Supplementary Material for: "On the Power and Limits of Distance-Based Learning"

Periklis A. Papakonstantinou Jia Xu Guang Yang

May 24, 2016

This supplementary material document contains:

1. Full proofs for every statement made in the main body.
2. Additional statements that introduce more details and more techniques for analyzing bi-Lipschitz learning settings.
3. Empirical results from the Machine Translation domain that indicate an underlying geometric structure in human and machine translation maps. The empirical results also show these properties for two "experts": (1) Google translate and (1) Bing Translate.

Contents

1 Roadmap to the Proofs

The cylindric example is listed on Section 3 of the main paper. This example is artificial but more accessible than the natural settings in Section 2 and Section 3. In this example, we assert that a selector C constructed with merely metric information is always be worse than an optimal one.

Section 2 contains the bounds for the hamming space $(\mathsf{GF}[2]^n, \ell_1)$, which is a previously very well-studied metric structure. In this setting we also achieve a lower bound for the generaliztion error of the selector C. Section 2.1 considers the extension when the input is not exactly uniformly distributed, and Section 2.2 extends to the case when the target concept is arbitrarily chosen and the learner uses random queries to construct the selector C. Lower bounds for C are proved in both extensions.

Then, we investigate more involved metric structures equipped with edit distance. A similar lower bound as in Section 2 is given as well as the extension to arbitrary target concept and the leaner with random queries. Section 3.3 contains the proof of Lemma 7 which is used in Section 3.1 and Section 3.2.

Finally, we support the naturalness of our setting with experimental evidence in Section 4. We put the experimental results only in the appendix because we want to stay focus on our in-depth results on lower bounds.

2 Choosing between two in the Hamming Cube

The previous cylindric example is mainly for expository reasons. In this section, we investigate the hamming cube example defined over $(\mathsf{GF}[2]^n, \ell_1)$, where the metric structure is more natural and well-studied.

Theorem 1. *For every $\rho \in \mathbb{Z}^{\geq 0}$ and $q(n)$, there is a concept class $\mathfrak{C}$, two experts $\mathsf{A_0}, \mathsf{A_1} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \ell_1)$, such that for $\mathsf{A_0}, \mathsf{A_1}$ and for every $\mathsf{r} \in \mathfrak{C}$ the bi-Lipschitz condition holds with constant 2, and*

- *for every $\mathsf{r} \in \mathfrak{C}$, $\mathcal{R}(\mathsf{A_0}) + \mathcal{R}(\mathsf{A_1}) = \rho$,*
- *if the input follows uniform distribution $\mathsf{D} = U_{2n}$, then for every generic selector $\mathsf{C} = \mathsf{C}_{\mathsf{A_0},\mathsf{A_1},[\mathsf{r}]}$ constructed from q queries, there is a target concept $\mathsf{r} \in \mathfrak{C}$ satisfying $\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{n^{\rho+1}q(n)}{2^{n+1}}\right)\rho$.*

Corollary 2. *Given the conditions of Theorem 1 and $\rho = \left\lfloor \frac{n - \log q(n)}{\log n} - 3 \right\rfloor$, there is $\frac{n^{\rho+1}q(n)}{2^n} \leq \frac{1}{n^2}$, and hence $\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{1}{2n^2}\right)\rho \approx \rho/2$. For example, $q(n)$ can be $\mathsf{poly}(n)$ or even $n^{\log n}$.*

Proof of Theorem 1. We first introduce the example and then verify the bi-Lipschitz conditions. Finally, we analyze the generalization error of the experts and the generic selector.

Construction (hamming space example) For $x \in \{0,1\}^n, y \in \{0,1\}^n$, let $\mathsf{A_0}$ and $\mathsf{A_1}$ be defined as

$$\begin{aligned} \mathsf{A_0}(x,y) &= \quad (x,y) \\ \mathsf{A_1}(x,y) &= \big(x, y + m(\rho)\big) \end{aligned} \tag{1}$$

where $m : \{0,1,\cdots,n\} \to \{0,1\}^n$ is a mask defined as $m(t) = (\underbrace{0,\cdots,0}_{n-t}, \underbrace{1,1,\cdots,1}_{t})$, and the computation of $y + m(\rho)$ is bitwise over $\{0,1\}^n$.

Note that $q(n)$ upper bounds the number of queries used to construct the selector C. Let $Q = \big\{q_i \big| 1 \le i \le q(n)\big\}$ denote the set of all queries, where $q_i = (q_i', q_i'') \in \{0,1\}^{2n}$ for every $q_i \in Q$ and $q_i', q_i'' \in \{0,1\}^n$. For completeness, let $Q' = \big\{q_i' \big| \exists q_i \in Q, q_i = (q_i', q_i'')\big\}$ and $Q'' = \big\{q_i'' \big| \exists q_i \in Q, q_i = (q_i', q_i'')\big\}$.

We also construct the concept class $\mathfrak{C} = \{\mathsf{r}_0\} \bigcup \Big\{\mathsf{r}_Q \big| Q \subseteq \{0,1\}^{2n}, |Q| \le q(n)\Big\}$, where r_0 denotes the identity mapping, and for every Q, r_Q is defined as

$$\mathsf{r}_Q(x,y) = (x, y + m(w_Q(x)))$$

where $m(t) = (\underbrace{0,\cdots,0}_{n-t}, \underbrace{1,1,\cdots,1}_{t})$ as in (1), and $w_Q(x) = \min\Big\{\min_{q_i' \in Q'}\big\{d\big(x, q_i'\big)\big\}, \rho\Big\}$ for q_i' being the first half of q_i as before. By definition, $\mathsf{r}_Q(x,y) = \mathsf{r}_0(x,y)$ immediately holds if $x \in Q'$, and in particular for every $(x,y) \in Q$.

Metric properties of $\mathsf{A_0}, \mathsf{A_1}$ and $\mathfrak{C}$ Bi-Lipschitz condition is trivial for $\mathsf{A_0}, \mathsf{A_1}$ and r_0, since both their contraction and expansion are exactly 1. For $\mathsf{r}_Q \in \mathfrak{C}$, the expansion of r_Q is bounded as follows,

$$\begin{aligned} & d\big(\mathsf{r}_Q(x,y), \mathsf{r}_Q(x',y')\big) \\ =& |x - x'| + |y + m(w_Q(x)) - y' - m(w_Q(x'))| \\ \le& |x - x'| + |y - y'| + |w_Q(x) - w_Q(x')| \\ \le& |x - x'| + |y - y'| + |x - x'| \\ \le& 2d\big((x,y),(x',y')\big) \end{aligned}$$

On the other hand, noticing $\mathsf{r}_Q\big(\mathsf{r}_Q(x,y)\big) = (x,y)$, the contraction is also bounded by 2 since $d\big((x,y),(x',y')\big) \le 2d\big(\mathsf{r}_Q(x,y), \mathsf{r}_Q(x',y')\big)$. Therefore, r_Q has bi-Lipschitz constant at most 2.

Generalization error of $\mathsf{A}_0, \mathsf{A}_1$ For $\mathsf{r} = \mathsf{r}_0$, trivially $\mathcal{R}(\mathsf{A}_0) = 0$, $\mathcal{R}(\mathsf{A}_1) = \rho$, and the conclusion follows. We prove that for arbitrary input distribution and every $\mathsf{r}_Q \in \mathfrak{C}$, $\mathcal{R}(\mathsf{A}_0) + \mathcal{R}(\mathsf{A}_1) = \rho$.

$$\begin{aligned}
\mathcal{R}(\mathsf{A}_0) + \mathcal{R}(\mathsf{A}_1) &= \mathbb{E}\left[d\big(\mathsf{A}_0(x,y), \mathsf{r}_Q(x,y)\big) + d\big(\mathsf{A}_1(x,y), \mathsf{r}_Q(x,y)\big)\right] \\
&= \mathbb{E}\left[d\Big((x,y), \big(x, y + m\big(w_Q(x)\big)\big)\Big) + d\Big(\big(x, y + m(\rho)\big), \big(x, y + m(w_Q(x))\big)\Big)\right] \\
&= \mathbb{E}\left[\left\|m(w_Q(x))\right\|_1 + \left\|m(\rho) - m(w_Q(x))\right\|_1\right] \\
&= \mathbb{E}\left[w_Q(x) + \big(\rho - w_Q(x)\big)\right] = \mathbb{E}[\rho] = \rho
\end{aligned}$$

Lower bound for $\mathcal{R}(\mathsf{C})$ We lower bound the generalization error $\mathcal{R}(\mathsf{C})$. The intuition is that as long as Q contains all the queries used to construct C, C cannot distinguish two concepts r_0 and r_Q, neither can it be close to both of them. Thus, $\mathcal{R}(\mathsf{C})$ is lower bounded.

The generalization distance of r_0 and r_Q, under uniform distribution, is bounded as

$$\begin{aligned}
d(\mathsf{r}_0, \mathsf{r}_Q) = &\mathop{\mathbb{E}}_{(x,y)\sim U_{2n}}\left[d\big(\mathsf{r}_0(x,y), \mathsf{r}_Q(x,y)\big)\right] \\
=&\frac{1}{2^{2n}} \sum_{(x,y)\in\{0,1\}^{2n}} \left\|m(w_Q(x))\right\|_1 \\
=&\frac{1}{2^{2n}} \sum_{(x,y)\in\{0,1\}^{2n}} w_Q(x) \\
=&\frac{1}{2^{n}} \sum_{x\in\{0,1\}^{n}} w_Q(x) = \mathop{\mathbb{E}}_{x\sim U_n}\left[w_Q(x)\right]
\end{aligned}$$

Recalling that $\mathcal{B}_z(\rho) \stackrel{\text{def}}{=} \{x \mid d(x,z) \le \rho\}$ denotes the closed *ball of radius ρ centered at z*, if $x \notin \bigcup_{q'_i \in Q'} \mathcal{B}_{q'_i}(\rho)$, then by definition of w_Q, it follows

$$w_Q(x) = \min\left\{\min_{q'_i \in Q'}\left\{d\big(x, q'_i\big)\right\}, \rho\right\} = \rho$$

and hence $\mathsf{r}_Q(x,y) = \big(x, y + m(w_Q(x))\big) = (x, y + m(\rho)) = \mathsf{A}_1(x,y)$ when $x \notin \bigcup_{q'_i \in Q'} \mathcal{B}_{q'_i}(\rho)$. Since $|Q'| \le |Q| \le q(n)$ and $\left|\mathcal{B}_{q'_i}(\rho)\right| \le \sum_{j=0}^{\lfloor\rho\rfloor} \binom{n}{j} < n^{\rho+1}$,

the volume of $\bigcup_{q_i' \in Q'} \mathcal{B}_{q_i'}(\rho)$ can be upper bounded by

$$\left| \bigcup_{q_i' \in Q'} \mathcal{B}_{q_i'}(\rho) \right| < n^{\rho+1} q(n), \tag{2}$$

Consequentially, we have

$$\begin{aligned} \left|\left\{x \in \{0,1\}^n \middle| w_Q(x) = \rho\right\}\right| =& 2^n - \left|\left\{x \in \{0,1\}^n \middle| \min_{q_i' \in Q'} \left\{d(x, q_i')\right\} < \rho\right\}\right| \\ \geq& 2^n - \left| \bigcup_{q_i' \in Q'} \mathcal{B}_{q_i'}(\rho) \right| > 2^n - n^{\rho+1} q(n) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathop{\mathbb{E}}_{x \sim U_n} [w_Q(x)] \geq& \frac{\left|\left\{x \in \{0,1\}^n \middle| w_Q(x) = \rho\right\}\right|}{2^n} \cdot \rho \\ >& \frac{(2^n - n^{\rho+1} q(n))\rho}{2^n} = \left(1 - \frac{n^{\rho+1} q(n)}{2^n}\right) \rho \end{aligned}$$

Plugging the above inequality into the bound for $d(\mathsf{r}_0, \mathsf{r}_Q)$, there is

$$d(\mathsf{r}_0, \mathsf{r}_Q) = \mathop{\mathbb{E}}_{x \sim U_n} [w_Q(x)] > \left(1 - \frac{n^{\rho+1} q(n)}{2^n}\right) \rho \tag{3}$$

For $\mathsf{A}_0, \mathsf{A}_1$ as in (1), recalling that Q contains all queries made by C, then the selector C cannot distinguish the concept r_0 from r_Q, since these two concepts coincide on every query in Q. Therefore, there is $\mathsf{r} \in \mathfrak{C}$ (in particular, $\mathsf{r} \in \{\mathsf{r}_0, \mathsf{r}_Q\}$) such that

$$\mathcal{R}(\mathsf{C}) \geq d(\mathsf{r}_0, \mathsf{r}_Q)/2 > \left(\frac{1}{2} - \frac{n^{\rho+1} q(n)}{2^{n+1}}\right) \rho$$

□

2.1 Extending to sparse almost uniform distribution.

In Theorem 1 and Theorem 1, the generalization errors are studied under the uniform distribution. Formally, to prove a PAC lower bound it is sufficient to exhibit a single example of a distribution. However, we would like to obtain even stronger bounds, aiming to understand better more realistic situations. We consider the following sparse almost uniform distribution as an extension.

Definition 3 (sparse almost uniform distribution)**.** *A distribution* D *over* $\{0,1\}^{2n}$ *is called a* k*-sparse almost uniform distribution if for every* $(x,y) \sim \mathsf{D}$ *where* $|x| = |y|$*,* y *is independent from* x *and distributes uniformly,* $\max_{s\in\{0,1\}^n} \Pr_{(x,y)\sim\mathsf{D}}[x = s] \leq 2^{-k}$*.*

Theorem 4. *For every* $\rho \in \mathbb{Z}^{\geq 0}$ *and* $q(n)$*, there is a concept class* $\mathfrak{C}$ *and two experts* $\mathsf{A_0}, \mathsf{A_1} : \mathcal{X} \to \mathcal{Y}$ *where* $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \ell_1)$ *for* $\mathsf{A_0}, \mathsf{A_1}$ *and for every* $\mathsf{r} \in \mathfrak{C}$ *the bi-Lipschitz condition holds with constant* 2*, such that for every* k*-sparse almost uniform distribution* D*,*

- *for every* $\mathsf{r} \in \mathfrak{C}$*,* $\mathcal{R}(\mathsf{A_0}) + \mathcal{R}(\mathsf{A_1}) = \rho$*,*
- *for every selector* $\mathsf{C} = \mathsf{C}_{\mathsf{A_0},\mathsf{A_1},[\mathsf{r}]}$ *constructed from* q *queries, there is a target concept* $\mathsf{r} \in \mathfrak{C}$ *satisfying* $\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{n^{\rho+1}q(n)}{2^{k+1}}\right)\rho$*.*

Proof. Let us consider exactly the same construction as in Theorem 1 (and the bi-Lipschitz condition automatically follows). For $x \in \{0,1\}^n, y \in \{0,1\}^n$,

$$\begin{array}{ll} \mathsf{A_0}(x,y) = & (x,y) \\ \mathsf{A_1}(x,y) = & \big(x, y + m(\rho)\big) \end{array}$$

The concept class is $\mathfrak{C} = \{\mathsf{r}_0\} \bigcup \Big\{\mathsf{r}_Q \big| Q \subseteq \{0,1\}^{2n}, |Q| \leq q(n)\Big\}$, where r_0 denotes the identity mapping, and for every Q, r_Q is defined as

$$\mathsf{r}_Q(x,y) = (x, y + m(w_Q(x)))$$

The masking function $m : \{0,1,\cdots,n\} \to \{0,1\}^n$ is defined as before, $m(t) = (\underbrace{0,\cdots,0}_{n-t},\underbrace{1,1,\cdots,1}_{t})$. $Q = \big\{q_i \big| 1 \leq i \leq q(n)\big\}$ and $Q' = \big\{q'_i \big| \exists q_i \in Q, q_i = (q'_i, q''_i)\big\}$, $Q'' = \big\{q''_i \big| \exists q_i \in Q, q_i = (q'_i, q''_i)\big\}$. And $w_Q(x) = \min\Big\{\min_{q'_i \in Q'} \big\{d\big(x, q'_i\big)\big\}, \rho\Big\}$ for q'_i being the first half of $q_i \in Q$. In particular, $w_0(x) = 0$ for every $x \in \{0,1\}^n$.

Generalization error of $\mathsf{A_0}, \mathsf{A_1}$**.** When $\mathsf{r} = \mathsf{r}_0$, it turns out $\mathcal{R}(\mathsf{A_0}) = 0$ and $\mathcal{R}(\mathsf{A_1}) = \rho$, and the conclusion follows. For every $\mathsf{r}_Q \in \mathfrak{C}$, $\mathcal{R}(\mathsf{A_0}) + \mathcal{R}(\mathsf{A_1}) = \rho$,

with exactly the same argument as in Theorem 1.

$$\begin{aligned}
\mathcal{R}(\mathsf{A}_0) + \mathcal{R}(\mathsf{A}_1) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} \left[d\big(\mathsf{A}_0(x,y), \mathsf{r}_Q(x,y)\big) + d\big(\mathsf{A}_1(x,y), \mathsf{r}_Q(x,y)\big) \right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} \left[d\Big((x,y), \big(x, y + m\big(w_Q(x)\big)\big)\Big) + d\Big(\big(x, y + m(\rho)\big), \big(x, y + m(w_Q(x))\big)\Big) \right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} \left[\left\| m(w_Q(x)) \right\|_1 + \left\| m(\rho) - m(w_Q(x)) \right\|_1 \right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} \left[w_Q(x) + \big(\rho - w_Q(x)\big) \right] = \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} [\rho] = \rho
\end{aligned}$$

Upper bound for the influence of "Beacons". It is clear that the set Q affects the value of $\mathsf{r}_Q(x,y)$ only when $x \in \bigcup_{q'\in Q'} \mathcal{B}_{q'}(\rho)$, since otherwise $w_Q(x) = \rho$ and $\mathsf{r}_Q = \mathsf{A}_1$. Let $\mathcal{B} = \mathcal{B}(Q,\rho) \stackrel{\text{def}}{=} \bigcup_{q'\in Q'} \mathcal{B}_{q'}(\rho)$. We bound the probability that $x \in \mathcal{B}$, when $(x,y) \sim \mathsf{D}$. Recalling that $|\mathcal{B}| < n^{\rho+1}q(n)$ as in (2) and D is k-sparse almost uniform distribution,

$$\Pr_{(x,y)\sim\mathsf{D}}[x \in \mathcal{B}] \le |\mathcal{B}| \cdot \max_{s\in\mathcal{B}} \Pr_{(x,y)\sim\mathsf{D}}[x = s] \le n^{\rho+1} q(n) \cdot 2^{-k} \tag{4}$$

Lower bound for $\mathcal{R}(\mathsf{C})$ Note that $w_Q(x) = \rho$ for $x \notin \mathcal{B}$, thus we lower bound the distance of r_0 and r_Q as follows:

$$\begin{aligned}
d(\mathsf{r}_0, \mathsf{r}_Q) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} \left[d\big(\mathsf{r}_0(x,y), \mathsf{r}_Q(x,y)\big) \right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathsf{D}} [w_Q(x)] \\
&\ge \rho \cdot \Pr_{(x,y)\sim\mathsf{D}} [x \notin \mathcal{B}] \ge (1 - n^{\rho+1} q(n) \cdot 2^{-k})\rho
\end{aligned}$$

Therefore, there exists $\mathsf{r} \in \mathfrak{C}$ (in particular, $\mathsf{r} \in \{\mathsf{r}_0, \mathsf{r}_Q\}$) such that

$$\mathcal{R}(\mathsf{C}) \ge d(\mathsf{r}_0, \mathsf{r}_Q)/2 > \left(\frac{1}{2} - \frac{n^{\rho+1} q(n)}{2^{k+1}} \right) \rho$$

□

2.2 Strong bounds for all target concepts and random queries

In Theorem 1 we fix the two experts and consider a target concept chosen in an adversarial way with respect to the q queries from which C is constructed.

Now, we consider the extension when C is constructed from random queries. The following theorem can be realized as a dual to Theorem 1.

Theorem 5. *For every $\rho \in \mathbb{Z}^{\geq 0}$, polynomial $q(n)$, and every target concept r with bi-Lipschitz constant c, there exists a set $\mathcal{A}$ of experts defined over $\mathcal{X} \to \mathcal{Y}$ where $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \ell_1)$ and all experts in $\mathcal{A}$ have bi-Lipschitz constant $\leq 2\mathsf{c}$, such that when the instances follow the uniform distribution $\mathsf{D} = U_{2n}$, then for every q-queries generic selector $\mathsf{C} = \mathsf{C}_{\mathsf{A}_0,\mathsf{A}_1,[\mathsf{r}]}$, there are two experts $\mathsf{A}_0, \mathsf{A}_1 \in \mathcal{A}$ such that $\min\{\mathcal{R}(\mathsf{A}_0), \mathcal{R}(\mathsf{A}_1)\} = 0$ but*

$$\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{n^{\rho+1}q(n)}{2^{n+1}}\right)\frac{\rho}{\mathsf{c}}$$

Proof. We begin with the construction of $\mathcal{A}$ with respect to r. Then, we verify its metric properies, and finally we will analyze the lower bound.

Construction Recalling that $q(n)$ upper bounds the number of queries used to construct the selector C, let $Q = \{q_i | 1 \leq i \leq q(n)\}$ denote the set of all queries, where $q_i = (q_i', q_i'') \in \{0,1\}^{2n}$ for every $q_i \in Q$ and $q_i', q_i'' \in \{0,1\}^n$. For completeness, let $Q' = \{q_i' | \exists q_i \in Q, q_i = (q_i', q_i'')\}$ and $Q'' = \{q_i'' | \exists q_i \in Q, q_i = (q_i', q_i'')\}$.

For $x \in \{0,1\}^n, y \in \{0,1\}^n$, define A_Q as

$$\mathsf{A}_Q(x,y) = \mathsf{r}\big(x, y + m(w_Q(x))\big) \tag{5}$$

where $m : \{0,1,\cdots,n\} \to \{0,1\}^n$ is a mask defined as $m(t) = (\underbrace{0,\cdots,0}_{n-t}, \underbrace{1,1,\cdots,1}_{t})$, and $w_Q(x) = \min\{d(x,Q'), \rho\} = \min\left\{\min_{q_i' \in Q'}\{d\big(x, q_i'\big)\}, \rho\right\}$, and the addition is bitwise over $\{0,1\}^n$.

By definition, $w_Q(x) = 0$ and immediately $\mathsf{A}_Q(x,y) = \mathsf{r}(x,y)$ holds if $x \in Q'$, and in particular when $(x,y) \in Q$. Intuitively, A_Q is close to r on input (x,y) only when x is close to some element in the set Q'. When x is far away from Q', i.e. $w_Q(x) = \rho$, the distance of A_Q and r is bounded by $d(\mathsf{A}_Q(x,y), \mathsf{r}(x,y)) \geq d(y + m(w_Q(x)), y)/\mathsf{c} = \rho/\mathsf{c}$ and similarly $d(\mathsf{A}_Q(x,y), \mathsf{r}(x,y)) \leq \rho\mathsf{c}$.

Then, we construct the class $\mathcal{A} = \{\mathsf{r}\} \bigcup \left\{\mathsf{A}_Q | Q \subseteq \{0,1\}^{2n}, |Q| \leq q(n)\right\}$

Metric properties of $\mathcal{A}$ Recalling that r has bi-Lipschitz constant c, we consider the bi-Lipschitz condition of $\mathsf{A}_Q \in \mathcal{A}$. For the expansion of A_Q, we

have

$$\begin{aligned}
&d\big(\mathsf{A}_Q(x,y),\mathsf{A}_Q(x',y')\big)\\
=&d\left(\mathsf{r}\big(x,y+m(w_Q(x))\big),\mathsf{r}\big(x',y'+m(w_Q(x'))\big)\right)\\
\le&\mathsf{c}\cdot d\left(\big(x,y+m(w_Q(x))\big),\big(x',y'+m(w_Q(x'))\big)\right)\\
=&\mathsf{c}\left(\|x-x'\|_1+\|y+m(w_Q(x))-y'-m(w_Q(x'))\|_1\right)\\
\le&\mathsf{c}\cdot\left(d\left((x,y),(x',y')\right)+\|w_Q(x)-w_Q(x')\|_1\right)\\
\le&\mathsf{c}\cdot d\left((x,y),(x',y')\right)+\mathsf{c}\cdot d(x,x')\\
\le&2\mathsf{c}\cdot d\big((x,y),(x',y')\big)
\end{aligned}$$

On the other hand, the contraction of A_Q is also bounded by $2\mathsf{c}$,

$$\begin{aligned}
&d\big(\mathsf{A}_Q(x,y),\mathsf{A}_Q(x',y')\big)\\
=&d\left(\mathsf{r}\big(x,y+m(w_Q(x))\big),\mathsf{r}\big(x',y'+m(w_Q(x'))\big)\right)\\
\ge&\mathsf{c}^{-1}\cdot d\left(\big(x,y+m(w_Q(x))\big),\big(x',y'+m(w_Q(x'))\big)\right)\\
=&\mathsf{c}^{-1}\left(\|x-x'\|_1+\|y+m(w_Q(x))-y'-m(w_Q(x'))\|_1\right)\\
\ge&\mathsf{c}^{-1}\cdot\max\left\{\|x-x'\|_1,\|x-x'\|_1+\|y-y'\|_1-\|w_Q(x)-w_Q(x')\|_1\right\}\\
\ge&\mathsf{c}^{-1}\cdot\max\left\{\|x-x'\|_1,\|y-y'\|_1\right\}\\
\ge&\mathsf{c}^{-1}\cdot\frac{\|x-x'\|_1+\|y-y'\|_1}{2}\\
=&\frac{d\big((x,y),(x',y')\big)}{2\mathsf{c}}
\end{aligned}$$

Thus, the bi-Lipschitz condition follows and A_Q has bi-Lipschitz constant $2\mathsf{c}$.

Lower bound for $\mathcal{R}(\mathsf{C})$ Let Q be the set of all queried points made in the construction of C. Randomly set $b\in\{0,1\}$, then let $\mathsf{A}_b=\mathsf{r}$ and $\mathsf{A}_{1-b}=\mathsf{A}_Q$. Since $\mathsf{A}_b(x,y)=\mathsf{A}_{1-b}(x,y)$ for every $(x,y)\in Q$, and moreover $\mathsf{A}_b(x,y)=\mathsf{A}_{1-b}\big(x,y+m(w_Q(x))\big)$, A_0 and A_1 are symmetric and hence indistinguishable to C. Therefore $\mathcal{R}(\mathsf{C})\ge(\mathcal{R}(\mathsf{A_0})+\mathcal{R}(\mathsf{A_1}))/2=\mathcal{R}(\mathsf{A}_Q)/2$. By the same calculation as in the proof of Theorem 1, we get the lower bound $\mathcal{R}(\mathsf{A}_Q)>\left(1-\frac{n^{\rho+1}q(n)}{2^n}\right)\rho/\mathsf{c}$ and as a result,

$$\mathcal{R}(\mathsf{C})>\left(\frac{1}{2}-\frac{n^{\rho+1}q(n)}{2^{n+1}}\right)\frac{\rho}{\mathsf{c}}$$

□

3 Choosing between two in Edit Spaces

Natural metric spaces lack the nice properties of the previous cylindric example. Our main result regards edit spaces. The technical obstacle is to give an explicit construction which deals with isoperimetric phenomena on spaces whose underlying graphs have high expansion. Roughly speaking this prevents constructing examples by simply gluing together independent copies (as we did for the cylinders).

In this section, we first introduce Theorem 6 and Theorem 9. We aslo exposit Corollary 8 for a better understanding of the theorem. Then, we provide the proof of Theorem 6 in Section 3.1. Section 3.2 contains the proof of Theorem 9 and Section 3.3 provides the proof to Lemma 7.

Theorem 6. *For every integer $\rho \leq n$ and q, there is a concept class $\mathfrak{C}$, two experts $\mathsf{A}_0, \mathsf{A}_1 : \mathcal{X} \to \mathcal{Y}$ (same experts for every concept), where $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \mathsf{ed})$, such that the following holds true for uniform $\mathsf{D} = U_{2n}$. For every learner with expert advice that makes q many queries there exists $\mathsf{r} \in \mathfrak{C}$ such that for the constructed selector C we have $\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{(4n+4\rho)^\rho q(n)}{2^{n+1}}\right) \Omega\left(\sqrt{\rho}\right)$, while the optimal selector has generalization error $\leq \frac{(4n+4\rho)^\rho q(n)}{2^n} \cdot \rho$. Furthermore, for $\mathsf{A}_0, \mathsf{A}_1$ and for every $\mathsf{r} \in \mathfrak{C}$ the bi-Lipschitz condition holds with constant less than 5.*

Proof sketch of Theorem 6. Consider two experts A_0, A_1 and the concept class $\mathfrak{C}$:

For $x \in \{0,1\}^n, y \in \{0,1\}^n$,

$$\begin{array}{ll} \mathsf{A}_0(x,y) = & (x,y) \\ \mathsf{A}_1(x,y) = & \big(x, y + m(\rho)\big) \end{array} \tag{6}$$

where the computation of $y+m(\rho)$ is bitwise over $\{0,1\}^n$, and $m : \{0,1,\cdots,n\} \to \{0,1\}^n$ is a mask defined as $m(t) = (\underbrace{0,\cdots,0}_{n-t}, \underbrace{1,1,\cdots,1}_{t})$. Let

$$\mathfrak{C} = \{\mathsf{r}_0\} \bigcup \left\{ \mathsf{r}_Q \middle| Q \subseteq \{0,1\}^{2n}, |Q| \leq q(n) \right\}$$

where $\mathsf{r}_0 = \mathsf{A}_0$ is identity and $\mathsf{r}_Q(x,y) = (x, y + m(w_Q(x)))$, where $w_Q(x) = \min\left\{ \min_{q_i' \in Q'} \left\{ \mathsf{ed}\big(x, q_i'\big) \right\}, \rho \right\}$, for $Q' = \left\{ q_i' \middle| \exists q_i \in Q, q_i = (q_i', q_i'') \right\}$.

Generalization error of $\mathsf{A}_0, \mathsf{A}_1$. For every $\mathsf{r} \in \mathfrak{C}$, $\mathcal{R}(\mathsf{A}_0) + \mathcal{R}(\mathsf{A}_1) \geq \mathbb{E}\left[\mathsf{ed}\big(y, y + m(\rho)\big)\right] = \Omega\left(\sqrt{\rho}\right)$ by the following lemma. This lemma is proved

by calculating the probability that ρ uniformly chosen bits have at least $(\rho + \sqrt{\rho})/2$ many 0's, and the full proof is defered to Section 3.3.

Lemma 7. *For every $\rho \le n$, and function m as in (6) we have*

$$\mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\right] \ge \Omega\left(\sqrt{\rho}\right)$$

Upper bound for the influence of "Beacons". Suppose Q consists of all queries instances, which are the "beacons". By definition, $\mathsf{r}_Q(x,y) = \mathsf{r}_0(x,y)$ only when $x \in Q'$, and the distance of r_0 and r_Q gradually grows to ρ as the input moves away from "beacons". Since each beacon influences a ball of radius ρ which has size at most $(4(n+\rho))^\rho$, the influence of all beacons can be upper bounded by the union of balls.

$$\left|\bigcup_{q_i'\in Q'} \mathcal{B}_{q_i'}(\rho)\right| < (4n+4\rho)^\rho\, q(n) \tag{7}$$

Lower bound for $\mathcal{R}(\mathsf{C})$ By inequality (7) and Lemma 7,

$$\begin{aligned}
&\mathsf{ed}(\mathsf{r}_0,\mathsf{r}_Q) = \mathop{\mathbb{E}}_{(x,y)\sim U_{2n}}\left[\mathsf{ed}\big(\mathsf{r}_0(x,y),\mathsf{r}_Q(x,y)\big)\right]\\
=&\frac{1}{2^{2n}}\sum_{(x,y)\in\{0,1\}^{2n}} \mathsf{ed}\Big((x,y),\big(x,y+m(w_Q(x))\big)\Big)\\
\ge&\frac{1}{2^n}\sum_{\substack{x\in\{0,1\}^n\\ w_Q(x)=\rho}}\left(\frac{1}{2^n}\sum_{y\in\{0,1\}^n}\mathsf{ed}\Big((x,y),\big(x,y+m(w_Q(x))\big)\Big)\right)\\
>&\frac{2^n-(4n+4\rho)^\rho\, q(n)}{2^n}\cdot\mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y,y+m(\rho)\big)\right] \ge \left(1-\frac{(4n+4\rho)^\rho\, q(n)}{2^n}\right)\Omega\left(\sqrt{\rho}\right)
\end{aligned}$$

Since C cannot distinguish r_0 from r_Q, there exists $\mathsf{r}\in\mathfrak{C}$ (in particular, $\mathsf{r}\in\{\mathsf{r}_0,\mathsf{r}_Q\}$) such that

$$\mathcal{R}(\mathsf{C}) \ge \mathsf{ed}(\mathsf{r}_0,\mathsf{r}_Q)/2 > \left(\frac{1}{2}-\frac{(4n+4\rho)^\rho\, q(n)}{2^{n+1}}\right)\Omega\left(\sqrt{\rho}\right)$$

The proof for the bi-Lipschitz condition (see the appendix) is technical but not difficult, which can be obtained by analyzing the optimal matching function. □

Corollary 8. *Given the conditions of Theorem 6 for* $q = q(n) \leq n^{\log n}$ *and* $\rho \leq \left\lfloor \frac{n - \log q(n)}{3 + \log n} - 3 \right\rfloor = \frac{n}{3+\log n} - \log n - \mathrm{O}(1)$*, then* $\frac{(4n+4\rho)^\rho q(n)}{2^n} < \frac{1}{n^3}$*. Thus, the optimal selector has generalization error* $\min\{\mathcal{R}(\mathsf{A_0}), \mathcal{R}(\mathsf{A_1})\} \leq \frac{(4n+4\rho)^\rho q(n)}{2^n} \cdot \rho < \frac{\rho}{n^3}$*, while the generalization error* $\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{1}{2n^3}\right) \Omega\left(\sqrt{\rho}\right) = \Omega\left(\sqrt{\rho}\right)$ *for every generic selector from* q *queries.*

Theorem 9 shares significant parts in its proof from that of Theorem 6. It also has a similar statement as the above one, where instead of a generic selector we have a *random samples generic selector* and the lower bound for $\mathcal{R}(\mathsf{C})$ holds *for every concept* $\mathsf{r} \in \mathfrak{C}$.

Theorem 9. *For every* $\rho \geq 0$*,* $q(n) \geq 0$*, and target concept* r *with distortion constant* c*, there exists a set* $\mathcal{A}$ *of bi-Lipschitz experts defined over* $\mathcal{X} \to \mathcal{Y}$ *for the edit space* $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \mathsf{ed})$*, such that for uniformly chosen input* $(x, y) \sim \mathsf{D} = U_{2n}$ *the following holds true. For every learner with black-box access that makes* q*-many queries to* r *and* q*-many queries to the two experts, there exist two experts* $\mathsf{A_0}, \mathsf{A_1} \in \mathcal{A}$ *with distortion* $\leq 5\mathsf{c}$ *and the optimal selector has* 0 *generalization error. Then, such a learner constructs* C *where* $\mathcal{R}(\mathsf{C}) \geq \left(\frac{1}{2} - \frac{(4n+4\rho)^\rho q(n)}{2^{n+1}}\right) \cdot \Omega\left(\frac{\sqrt{\rho}}{c}\right)$*. In particular,* $\mathcal{R}(\mathsf{C}) \geq \Omega\left(\frac{\sqrt{\rho}}{c}\right)$ *when* $\rho \leq \left\lfloor \frac{n - \log q(n)}{3 + \log n} - 1 \right\rfloor$*.*

3.1 Proof of Theorem 2

Proof of Theorem 6. Similar to the proof of Theorem 1, we first introduce the example, then analyze the generalization error of the experts and the generic selector, and finally verify the bi-Lipschitz conditions.

Construction (family of examples) For $x \in \{0,1\}^n, y \in \{0,1\}^n$, let $\mathsf{A_0}$ and $\mathsf{A_1}$ be defined as follows

$$\begin{aligned} \mathsf{A_0}(x, y) &= (x, y) \\ \mathsf{A_1}(x, y) &= \big(x, y + m(\rho)\big) \end{aligned} \tag{8}$$

where $m : \{0, 1, \cdots, n\} \to \{0,1\}^n$ is a mask $m(t) = (\underbrace{0, \cdots, 0}_{n-t}, \underbrace{1, 1, \cdots, 1}_{t})$, and the computation of $y + m(\rho)$ is bitwise over $\{0,1\}^n$.

Note that $q(n)$ upper bounds the number of queries made in the construction of the selector C. Let $Q = \left\{q_i \middle| 1 \leq i \leq q(n)\right\}$ denote the set of all queries, where $q_i = (q_i', q_i'') \in \{0,1\}^{2n}$ for every $q_i \in Q$ and $q_i', q_i'' \in$

$\{0,1\}^n$. For completeness, let $Q' = \left\{q'_i \middle| \exists q_i \in Q, q_i = (q'_i, q''_i)\right\}$ and $Q'' = \left\{q''_i \middle| \exists q_i \in Q, q_i = (q'_i, q''_i)\right\}$.

Then we construct the concept class $\mathfrak{C} = \{\mathsf{r}_0\} \bigcup \left\{\mathsf{r}_Q \middle| Q \subseteq \{0,1\}^{2n}, |Q| \leq q(n)\right\}$. r_0 denotes the identity mapping, and for every Q, r_Q is defined as

$$\mathsf{r}_Q(x,y) = (x, y + m(w_Q(x)))$$

where $m(t) = (\underbrace{0,\cdots,0}_{n-t},\underbrace{1,1,\cdots,1}_{t})$ and $w_Q(x) = \min\left\{\min_{q'_i \in Q'}\left\{\mathsf{ed}\big(x,q'_i\big)\right\}, \rho\right\}$ as before. By definition, $\mathsf{r}_Q(x,y) = \mathsf{r}_0(x,y)$ if and only if $x \in Q'$, which holds in particular for every $(x,y) \in Q$.

Upper bounding the volume of $\bigcup_{q'_i \in Q'} \mathcal{B}_{q'_i}(\rho)$. Let $\mathcal{B}_z(\rho) \stackrel{\text{def}}{=} \{x \mid \mathsf{ed}(x,z) \leq \rho\}$ denote the closed *ball of radius ρ centered at z*. For $x \notin \bigcup_{q'_i \in Q'} \mathcal{B}_{q'_i}(\rho)$, by definition of w_Q, it follows

$$w_Q(x) = \min\left\{\min_{q'_i \in Q'}\left\{\mathsf{ed}\big(x,q'_i\big)\right\}, \rho\right\} = \rho$$

To upper bound the volume of $\mathcal{B}_z(\rho)$, we calculate the number of distinct ρ-step operation sequences from $z \in \{0,1\}^n$. For each of the ρ steps, we can perform either insertion $0/1$ to one of at most $n+\rho$ positions, or deletion and substitution in one of at most $n+\rho-1$ positions, or no operation. Totally, we have less than $4(n+\rho)$ choices in each step. Thus, we bound $\mathcal{B}_{q'_i}(\rho)$ as

$$\left|\mathcal{B}_{q'_i}(\rho)\right| < (4(n+\rho))^\rho$$

By union bound, the volume of the union of $q(n)$ balls is upper bounded by

$$\left|\bigcup_{q'_i \in Q'} \mathcal{B}_{q'_i}(\rho)\right| < (4n+4\rho)^\rho\, q(n)$$

As a consequence,

$$\begin{aligned}\left|\left\{x \in \{0,1\}^n \middle| w_Q(x) = \rho\right\}\right| =& 2^n - \left|\left\{x \in \{0,1\}^n \middle| \min_{q'_i \in Q'}\left\{d(x,q'_i)\right\} < \rho\right\}\right| \\ \geq& 2^n - \left|\bigcup_{q'_i \in Q'} \mathcal{B}_{q'_i}(\rho)\right| > 2^n - (4n+4\rho)^\rho\, q(n)\end{aligned}$$

Generalization error of $\mathsf{A}_0, \mathsf{A}_1$ We consider the generalization error of an optimal selector for the concept class $\mathfrak{C}$. If $\mathsf{r} = \mathsf{r}_0$, then the generalization error is 0 since $\mathsf{A}_0 = \mathsf{r}_0$. Otherwise, if $\mathsf{r} = \mathsf{r}_Q$,

$$\begin{aligned}
\mathcal{R}(\mathsf{A}_1) &= \mathbb{E}\left[\mathsf{ed}\big(\mathsf{A}_1(x,y), \mathsf{r}(x,y)\big)\right] \\
&= \mathbb{E}\left[\mathsf{ed}\Big(\big(x, y+m(\rho)\big), \big(x, y+m(w_Q(x))\big)\Big)\right] \\
&\leq \mathbb{E}\left[\mathsf{ed}\Big(y+m(\rho), y+m\big(w_Q(x)\big)\Big)\right] \\
&\leq \Pr_{x\sim\mathsf{D}}\left[w_Q(x) < \rho\right] \mathbb{E}\left[\mathsf{ed}\Big(y+m(\rho), y+m\big(w_Q(x)\big)\Big)\Big| w_Q(x) < \rho\right] \\
&\leq \frac{(4n+4\rho)^\rho\, q(n)}{2^n} \cdot \rho
\end{aligned}$$

Thus, we conclude

$$\min\left\{\mathcal{R}(\mathsf{A}_0), \mathcal{R}(\mathsf{A}_1)\right\} \leq \frac{(4n+4\rho)^\rho\, q(n)}{2^n} \cdot \rho$$

Lower bound for $\mathcal{R}(\mathsf{C})$ The lower bound of $\mathcal{R}(\mathsf{C})$ follows the same intuition that C cannot distinguish two concepts which are far away from each other and on the other hand C can never be close to both of them.

To lower bounding the generalization distance of r_0 and r_Q, we first recall that

$$\left|\left\{x \in \{0,1\}^n \middle| w_Q(x) = \rho\right\}\right| > 2^n - (4n+4\rho)^\rho\, q(n)$$

Then using Lemma 7,

$$\begin{aligned}
\mathsf{ed}(\mathsf{r}_0, \mathsf{r}_Q) =& \mathop{\mathbb{E}}_{(x,y)\sim U_{2n}}\left[\mathsf{ed}\big(\mathsf{r}_0(x,y), \mathsf{r}_Q(x,y)\big)\right] \\
=& \frac{1}{2^{2n}} \sum_{(x,y)\in\{0,1\}^{2n}} \mathsf{ed}\left((x,y), \big(x, y+m(w_Q(x))\big)\right) \\
\geq& \frac{1}{2^n} \sum_{\substack{x\in\{0,1\}^n \\ w_Q(x)=\rho}} \left(\frac{1}{2^n} \sum_{y\in\{0,1\}^n} \mathsf{ed}\left((x,y), \big(x, y+m(w_Q(x))\big)\right)\right) \\
>& \frac{2^n - (4n+4\rho)^\rho\, q(n)}{2^n} \mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\right] \\
\geq& \left(1 - \frac{(4n+4\rho)^\rho\, q(n)}{2^n}\right) \Omega\left(\sqrt{\rho}\right)
\end{aligned}$$

For $\mathsf{A}_0, \mathsf{A}_1$ as in (6), reacalling that Q consists of all queries made by C, then C cannot distinguish the concept r_0 from r_Q, since these two concepts

coincide on every query in Q. Therefore, there exists $\mathsf{r} \in \mathfrak{C}$ (in particular, $\mathsf{r} \in \{\mathsf{r}_0, \mathsf{r}_Q\}$) such that

$$\mathcal{R}(\mathsf{C}) \geq \mathsf{ed}(\mathsf{r}_0, \mathsf{r}_Q)/2 > \left(\frac{1}{2} - \frac{(4n+4\rho)^\rho\, q(n)}{2^{n+1}}\right) \Omega\left(\sqrt{\rho}\right)$$

Metric properties of $\mathsf{A}_0, \mathsf{A}_1$ and $\mathfrak{C}$ The bi-Lipschitz condition is trivial for A_0 and r_0 since they are identity. Then, we prove the bi-Lipschitz constant is at most 3 for A_1 and no more than 5 for every $\mathsf{r}_Q \in \mathfrak{C}\backslash\{\mathsf{r}_0\}$.

Firstly, we bound $\mathsf{ed}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big) = \mathsf{ed}\left(\big(x, y+m(\rho)\big), \big(x', y'+m(\rho)\big)\right)$ with $\mathsf{ed}\big((x,y),(x',y')\big)$. Let us consider the optimal matching $\mathcal{M}\big((x,y),(x',y')\big)$ between (x,y) and (x',y'), which maps every pair of unchanged bits when editing[1] (x,y) to (x',y'). Denote by $\big|\mathcal{M}\big((x,y),(x',y')\big)\big|$ the number of matched pairs defined by the matching, then (note the input length is $2n$),

$$2n - \big|\mathcal{M}\big((x,y),(x',y')\big)\big| \leq \mathsf{ed}\big((x,y),(x',y')\big) \leq 4n - 2\big|\mathcal{M}\big((x,y),(x',y')\big)\big|$$

Similarly,

$$2n - \big|\mathcal{M}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big)\big| \leq \mathsf{ed}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big) \leq 4n - 2\big|\mathcal{M}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big)\big|$$

Therefore, it suffices to compare $\big|\mathcal{M}\big((x,y),(x',y')\big)\big|$ and $\big|\mathcal{M}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big)\big|$. Realizing the matchings as graphs with each edge corresponding to a pair of matched bits, we let Γ denote the cut in $\mathcal{M}\big((x,y),(x',y')\big)$ at the position $2n-\rho$, i.e. Γ consists of all pairs containing one bit from the first $2n-\rho$ positions and the other from the last ρ positions.

On one hand, we assert that Γ only contains edges in either of the following two cases: from the first $2n-\rho$ positions of (x,y) to the last ρ of (x',y'), or from the last ρ positions of (x,y) to the first $2n-\rho$ positions of (x',y'). Because in the process of editing, the relative order of remaining bits can never be changed. Then, it trivially follows that $2|\Gamma| \leq \mathsf{ed}\big((x,y),(x',y')\big)$, since each single edge in Γ will cause at least two insertion/deletion operations.

On the other hand, we notice that if a pair in $\mathcal{M}\big((x,y),(x',y')\big)$ is not in Γ, it can also appear in the matching between $\mathsf{A}_1(x,y)$ and $\mathsf{A}_1(x',y')$. Thus, $\big|\mathcal{M}\big((x,y),(x',y')\big)\big| - \big|\mathcal{M}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big)\big| \leq |\Gamma|$.

Combining above two parts,

$$\big|\mathcal{M}\big((x,y),(x',y')\big)\big| - \big|\mathcal{M}\big(\mathsf{A}_1(x,y), \mathsf{A}_1(x',y')\big)\big| \leq |\Gamma| \leq \mathsf{ed}\big((x,y),(x',y')\big)/2$$

[1] In case there are multiple ways to edit (x,y) to (x',y') with the same number of operations, fix $\mathcal{M}$ to any of them.

Then, recalling the lower and upper bound for $\mathsf{ed}\big((x,y),(x',y')\big)$ and $\mathsf{ed}\big(\mathsf{A}_1(x,y),\mathsf{A}_1(x',y')\big)$, the expansion of A_1 is at most 3.

$$\begin{aligned}
&\mathsf{ed}\big(\mathsf{A}_1(x,y),\mathsf{A}_1(x',y')\big)\\
\leq&4n-2\left|\mathcal{M}\big(\mathsf{A}_1(x,y),\mathsf{A}_1(x',y')\big)\right|\\
\leq&4n-2\left|\mathcal{M}\big((x,y),(x',y')\big)\right|+\mathsf{ed}\big((x,y),(x',y')\big)\\
\leq&3\mathsf{ed}\big((x,y),(x',y')\big)
\end{aligned}$$

Since $\mathsf{A}_1\big(\mathsf{A}_1(x,y)\big)=(x,y)$, the contraction of A_1 follows the same bound, i.e. $\mathsf{ed}\big((x,y),(x',y')\big)\leq 3\mathsf{ed}\big(\mathsf{A}_1(x,y),\mathsf{A}_1(x',y')\big)$. To conclude, A_1 has bi-Lipschitz constant 3.

For r_Q, we verify its bi-Lipschitz constnat is at most 5.

By definition of r_Q, we have

$$\mathsf{ed}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)=\mathsf{ed}\left(\big(x,y+m(w_Q(x))\big),\big(x',y'+m(w_Q(x'))\big)\right)$$

Again, let $\mathcal{M}\big((x,y),(x',y')\big)$ be the optimal matching between (x,y) and (x',y'). And similarly define $\mathcal{M}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)$, then,

$$\mathsf{ed}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)\leq 4n-2\left|\mathcal{M}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)\right|$$

Let Γ denote the cut in $\mathcal{M}\big((x,y),(x',y')\big)$ at the position $2n-w_Q(x)$ of (x,y) and the position $2n-w_Q(x')$ of (x',y'), i.e. Γ consists of all edges in the matching $\mathcal{M}\big((x,y),(x',y')\big)$ that

- either from the first $2n-w_Q(x)$ positions of (x,y) to the last $w_Q(x')$ positions of (x',y'),
- or from the last $w_Q(x)$ positions of (x,y) to the first $2n-w_Q(x')$ positions of (x',y').

Moreover, all edges in Γ must be in the same case. By counting the number of unmatched bits on both sides of the cut, we can get $\mathsf{ed}\big((x,y),(x',y')\big)\geq 2\big(|\Gamma|-|w_Q(x)-w_Q(x')|\big)$.

On the other hand, $\left|\mathcal{M}\big((x,y),(x',y')\big)\right|-\left|\mathcal{M}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)\right|\leq|\Gamma|$ since all edges in $\mathcal{M}\big((x,y),(x',y')\big)\backslash\Gamma$ can be simultaneously reserved in a matching from $\mathsf{r}_Q(x,y)$ to $\mathsf{r}_Q(x',y')$.

As a result,

$$\left|\mathcal{M}\big((x,y),(x',y')\big)\right|-\left|\mathcal{M}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)\right|\leq|\Gamma|\leq\mathsf{ed}\big((x,y),(x',y')\big)/2+\left|w_Q(x)-w_Q(x')\right|$$

Plugging it into previous lower and upper bound for $\mathsf{ed}\big((x,y),(x',y')\big)$ and

$\mathsf{ed}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)$, the expansion of r_Q is at most 5.

$$\begin{aligned}
&\mathsf{ed}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)\\
\leq&4n-2\left|\mathcal{M}\big(\mathsf{r}_Q(x,y),\mathsf{r}_Q(x',y')\big)\right|\\
\leq&4n-2\left|\mathcal{M}\big((x,y),(x',y')\big)\right|+\mathsf{ed}\big((x,y),(x',y')\big)+2\left|w_Q(x)-w_Q(x')\right|\\
\leq&3\mathsf{ed}\big((x,y),(x',y')\big)+2\left|w_Q(x)-w_Q(x')\right|\\
\leq&3\mathsf{ed}\big((x,y),(x',y')\big)+2\mathsf{ed}(x,x')\\
\leq&5\mathsf{ed}\big((x,y),(x',y')\big)
\end{aligned}$$

Since $\mathsf{r}_Q\big(\mathsf{r}_Q(x,y)\big)=(x,y)$, the contraction of r_Q is also bounded by 5. Thus, r_Q has bi-Lipschitz constant 5. □

3.2 Proof of Theorem 3: extending to all target concepts and random queries

Similarly as with the extension in Section 2.2, we consider the case where the target concept is a fixed mapping with low distortion in edit space. We show that for the fixed target concept there exists a pair of experts such that from which no selector with small generalization error can be constructed.

Theorem 9 (informally stated). For every $\rho\geq 0$, $q(n)\geq 0$, and target concept r with distortion constant c, there exists a set $\mathcal{A}$ of bi-Lipschitz experts defined over $\mathcal{X}\to\mathcal{Y}$ for the edit space $\mathcal{X}=\mathcal{Y}=(\{0,1\}^{2n},\mathsf{ed})$, such that for uniformly chosen input $(x,y)\sim\mathsf{D}=U_{2n}$ the following holds true. For every learner with black-box access that makes q-many queries to r and q-many queries to the two experts, there exist two experts $\mathsf{A_0},\mathsf{A_1}\in\mathcal{A}$ with distortion $\leq 5\mathsf{c}$ and the optimal selector has 0 generalization error. Then, such a learner constructs C where $\mathcal{R}(\mathsf{C})\geq\left(\frac{1}{2}-\frac{(4n+4\rho)^\rho q(n)}{2^{n+1}}\right)\cdot\Omega\left(\frac{\sqrt{\rho}}{c}\right)$. In particular, $\mathcal{R}(\mathsf{C})\geq\Omega\left(\frac{\sqrt{\rho}}{c}\right)$ when $\rho\leq\left\lfloor\frac{n-\log q(n)}{3+\log n}-1\right\rfloor$.

Proof. We begin with the construction of $\mathcal{A}$. Then, we disucss the metric property and the generalization error of experts $\mathsf{A}_Q\in\mathcal{A}\backslash\{\mathsf{r}\}$. Finally, we prove the lower bound for $\mathcal{R}(\mathsf{C})$.

Construction Recalling that $q(n)$ upper bounds the number of queries made by the selector C, let $Q=\big\{q_i\big|1\leq i\leq q(n)\big\}\subseteq\{0,1\}^{2n}$ denote the set of all queries, and define $Q'=\big\{q'_i\big|\exists q_i\in Q,q_i=(q'_i,q''_i)\big\}\subseteq\{0,1\}^n$ to be the set of first half of elements in Q.

For $x \in \{0,1\}^n, y \in \{0,1\}^n$, define A_Q as in (5)

$$\mathsf{A}_Q(x,y) = \mathsf{r}\big(x, y + m(w_Q(x))\big)$$

where $m : \{0,1,\cdots,n\} \to \{0,1\}^n$ is a mask defined as $m(t) = (\underbrace{0,\cdots,0}_{n-t},\underbrace{1,1,\cdots,1}_{t})$, and $w_Q(x) = \min\{d(x,Q'),\rho\} = \min\Big\{\min_{q_i' \in Q'}\big\{d\big(x,q_i'\big)\big\},\rho\Big\}$. When $x \in Q'$, in particular when $(x,y) \in Q$, $w_Q(x) = 0$ and immediately $\mathsf{A}_Q(x,y) = \mathsf{r}(x,y)$ follows by definition.

The expert class is constructed as $\mathcal{A} = \{\mathsf{r}\} \bigcup \Big\{\mathsf{A}_Q \big| Q \subseteq \{0,1\}^{2n}, |Q| \le q(n)\Big\}$

Metric properties of $\mathsf{A}_Q \in \mathcal{A}$ For the bi-Lipschitz condition of elements in $\mathcal{A}$, we need to bound $\mathsf{ed}\big(\mathsf{A}_Q(x,y), \mathsf{A}_Q(x',y')\big)$, with $\mathsf{ed}\big((x,y),(x',y')\big)$. Recalling that r has bi-Lipschitz constant c and $\mathsf{ed}\big(\mathsf{A}_Q(x,y), \mathsf{A}_Q(x',y')\big) = \mathsf{ed}\left(\mathsf{r}\big(x, y + m(w_Q(x))\big), \mathsf{r}\big(x', y' + m(w_Q(x'))\big)\right)$,

$$\frac{1}{\mathsf{c}} \le \frac{\mathsf{ed}\left(\mathsf{r}\big(x, y + m(w_Q(x))\big), \mathsf{r}\big(x', y' + m(w_Q(x'))\big)\right)}{\mathsf{ed}\left(\big(x, y + m(w_Q(x))\big), \big(x', y' + m(w_Q(x'))\big)\right)} \le \mathsf{c}$$

Following the bi-Lipschitz proof of r_Q in Theorem 6, we get the upper bound for expansion

$$\mathsf{ed}\left(\big(x, y + m(w_Q(x))\big), \big(x', y' + m(w_Q(x'))\big)\right) \le 5\mathsf{ed}\big((x,y),(x',y')\big)$$

and symmetrically for contraction

$$\mathsf{ed}\big((x,y),(x',y')\big) \le 5\mathsf{ed}\left(\big(x, y + m(w_Q(x))\big), \big(x', y' + m(w_Q(x'))\big)\right)$$

Therefore, the bi-Lipschitz constant of A_Q is at most $5\mathsf{c}$.

Generalization error of $\mathsf{A}_Q \in \mathcal{A}$ Before lower bounding the generalization error $\mathcal{R}(\mathsf{A}_Q)$, we upper bound the fraction of elements close to Q'. Let $B = \big\{x \in \{0,1\}^n \big| d(x,Q') < \rho\big\}$ be the set of all elements close to Q', then $B = \bigcup_{q_i' \in Q'} \mathcal{B}_{q_i'}(\rho)$ as in (7) and the volume is upper bounded by

$$|B| = \left|\bigcup_{q_i' \in Q'} \mathcal{B}_{q_i'}(\rho)\right| < (4n + 4\rho)^\rho \, q(n)$$

The generalization error of A_Q is $\Omega\left(\frac{\sqrt{\rho}}{\mathsf{c}}\right)$ as long as $(x,y) \sim U_{2n}$. Recalling Lemma 7 that $\mathbb{E}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\right] = \Omega\left(\sqrt{\rho}\right)$, we have

$$\begin{aligned}
\mathcal{R}(\mathsf{A}_Q) &= \mathbb{E}\left[\mathsf{ed}\big(\mathsf{A}_Q(x,y), \mathsf{r}(x,y)\big)\right] \\
&= \mathbb{E}\left[\mathsf{ed}\Big(\mathsf{r}\big(x, y+m(w_Q(x))\big), \mathsf{r}(x,y)\Big)\right] \\
&> \mathbb{E}\left[\mathsf{ed}\Big(\mathsf{r}\big(x, y+m(\rho)\big), \mathsf{r}(x,y)\Big)\Big| x \notin B\right] \cdot \Pr_{x\sim U_n}\left[x \notin B\right] \\
&\geq \mathbb{E}\left[\frac{1}{\mathsf{c}} \cdot \mathsf{ed}\Big(\big(x, y+m(\rho)\big), (x,y)\Big)\right] \cdot \left(1 - \frac{(4n+4\rho)^\rho\, q(n)}{2^n}\right) \\
&\geq \frac{1}{c} \cdot \mathbb{E}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\right] \cdot \left(1 - \frac{(4n+4\rho)^\rho\, q(n)}{2^n}\right) \\
&\geq \left(1 - \frac{(4n+4\rho)^\rho\, q(n)}{2^n}\right) \cdot \frac{\Omega\left(\sqrt{\rho}\right)}{c}
\end{aligned}$$

and when $\rho \leq \left\lfloor \frac{n - \log q(n)}{3+\log n} - 1 \right\rfloor$,

$$\mathcal{R}(\mathsf{A}_Q) = \Omega\left(\frac{\sqrt{\rho}}{c}\right)$$

Lower bound for $\mathcal{R}(\mathsf{C})$ Let Q be the set of all queried points made by C. Uniformly at random select $b \in \{0,1\}$, then let $\mathsf{A}_b = \mathsf{r}$ and $\mathsf{A}_{1-b} = \mathsf{A}_Q$. Trivially we have

$$\min\left\{\mathcal{R}(\mathsf{A}_0), \mathcal{R}(\mathsf{A}_1)\right\} = 0$$

Note that $\mathsf{A}_b(x,y) = \mathsf{A}_{1-b}(x,y)$ when $(x,y) \in Q$, and moreover $\mathsf{A}_b(x,y) = \mathsf{A}_{1-b}\big(x, y + +m(w_Q(x))\big)$. Therefore, A_0 and A_1 are indistinguishable to C, and hence

$$\mathcal{R}(\mathsf{C}) \geq \frac{\mathcal{R}(\mathsf{A}_0) + \mathcal{R}(\mathsf{A}_1)}{2} = \mathcal{R}(\mathsf{A}_Q)/2 = \left(1 - \frac{(4n+4\rho)^\rho\, q(n)}{2^n}\right) \cdot \Omega\left(\frac{\sqrt{\rho}}{c}\right)$$

□

3.3 Proof of Main Lemma

Lemma 10. *For every $\rho \leq n$, we have the following lower bound for the expected edit distance between a random string and its complement.*

$$\mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\right] \geq \Omega\left(\sqrt{\rho}\right)$$

Proof of Lemma 10. We lower bound the expected edit distance by calculating the probability the y has $\sqrt{\rho}$ more 0's than 1's in its last ρ bits.

Consider the case when y has at least $(\rho+\sqrt{\rho})/2$ bits equal to 0 in the last ρ bits, then the number of 0's in last ρ bits of $y + m(\rho)$ is at most $(\rho-\sqrt{\rho})/2$. As a result,

$$\mathsf{ed}\big(y, y+m(\rho)\big) \geq |\#\text{ of 0's in } y - \#\text{ of 0's in } y+m(\rho)| \geq \sqrt{\rho}$$

Then, we bound the probability of the above case. In particular, we lower bound it by the probability that in the last ρ bits of y there are at least $(\rho+\sqrt{\rho})/2$ but at most $\rho/2+\sqrt{\rho}$ bits equal to 0.

$$\begin{aligned}
&\Pr_{y\sim U_n}\left[\ (\rho+\sqrt{\rho})/2 \leq (\#\text{ of 0's in last } \rho \text{ bits of } y) \leq \rho/2+\sqrt{\rho}\ \right] \\
=&\sum_{i=\rho/2+\sqrt{\rho}/2}^{\rho/2+\sqrt{\rho}} \frac{\binom{\rho}{i}\cdot 2^{n-\rho}}{2^n} \\
\geq&\frac{1}{2^\rho}\cdot(\frac{\sqrt{\rho}}{2}+1)\cdot\binom{\rho}{\rho/2+\sqrt{\rho}} \\
\geq&\frac{\sqrt{\rho}}{2^{\rho+1}}\cdot\frac{\rho!}{(\rho/2+\sqrt{\rho})!(\rho/2-\sqrt{\rho})!}
\end{aligned} \tag{9}$$

Recalling that by Stirling's Formula (the version due to Robbins 1955),

$$\sqrt{2\pi}n^{n+1/2}\exp(-n+\frac{1}{12n+1}) \leq n! \leq \sqrt{2\pi}n^{n+1/2}\exp(-n+\frac{1}{12n})$$

we derive the lower bound for $\binom{\rho}{\rho/2+\sqrt{\rho}}$ as

$$
\begin{aligned}
&\binom{\rho}{\rho/2+\sqrt{\rho}} \\
=&\frac{\rho!}{(\rho/2+\sqrt{\rho})!(\rho/2-\sqrt{\rho})!} \\
\geq&\frac{\sqrt{2\pi}\rho^{\rho+1/2}\exp(-\rho+\frac{1}{12\rho+1})}{\left(\sqrt{2\pi}(\rho/2+\sqrt{\rho})^{\rho/2+\sqrt{\rho}+1/2}\exp(-(\rho/2+\sqrt{\rho})+1)\right)}\cdot \\
&\qquad\frac{1}{\left(\sqrt{2\pi}(\rho/2-\sqrt{\rho})^{\rho/2-\sqrt{\rho}+1/2}\exp(-(\rho/2-\sqrt{\rho})+1)\right)} \\
=&\frac{\rho^{\rho+1/2}\exp\left(-\rho+\frac{1}{12\rho+1}+(\rho/2+\sqrt{\rho})+(\rho/2-\sqrt{\rho})-2\right)}{\sqrt{2\pi}(\rho/2+\sqrt{\rho})^{\rho/2+\sqrt{\rho}+1/2}(\rho/2-\sqrt{\rho})^{\rho/2-\sqrt{\rho}+1/2}} \\
\geq&\frac{\rho^{\rho+1/2}\exp\left(\frac{1}{12\rho+1}-2\right)}{\sqrt{2\pi}(\frac{\rho}{2}+\frac{2\rho}{\rho+1})^{\rho+1}} \qquad (10)\\
\geq&\frac{\rho^{\rho+1/2}\exp\left(\frac{1}{12\rho+1}-2\right)}{\sqrt{2\pi}(\rho/2+2)^{\rho+1}} \\
=&\frac{1}{\sqrt{2\pi\rho}}\cdot\left(\frac{\rho}{\rho/2+2}\right)^{\rho+1}\exp\left(\frac{1}{12\rho+1}-2\right) \\
\geq&\frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^2}\cdot\left(\frac{\rho/2}{\rho/2+2}\right)^{\rho+1}=\frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^2}\cdot\left(1-\frac{1}{\rho/4+1}\right)^{\rho+1} \\
\geq&\frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^2}\cdot\left(1-\frac{1}{\rho/4+1}\right)^{4(\rho/4+1)} \\
>&\frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^2}\cdot\left(\frac{1}{5}\right)^5 \qquad (11)
\end{aligned}
$$

The inequality (10) holds since

$$
\begin{aligned}
&(\rho/2+\sqrt{\rho})^{\rho/2+\sqrt{\rho}+1/2}\cdot(\rho/2-\sqrt{\rho})^{\rho/2-\sqrt{\rho}+1/2} \\
\leq&\left(\frac{(\rho/2+\sqrt{\rho})(\rho/2+\sqrt{\rho}+1/2)+(\rho/2-\sqrt{\rho})(\rho/2-\sqrt{\rho}+1/2)}{\rho/2+\sqrt{\rho}+1/2+\rho/2-\sqrt{\rho}+1/2}\right)^{\rho/2+\sqrt{\rho}+1/2+\rho/2-\sqrt{\rho}+1/2} \\
=&\left(\frac{\rho^2/2+5\rho/2}{\rho+1}\right)^{\rho+1}=(\frac{\rho}{2}+\frac{2\rho}{\rho+1})^{\rho+1}
\end{aligned}
$$

And the inequality (11) holds for $\rho \geq 1$, since the function $\left(1-\frac{1}{x}\right)^x$ is monotonically increasing and converges to e^{-1} when $x > 1$, therefore

$$\left(1-\frac{1}{\rho/4+1}\right)^{\rho/4+1} \geq \left(\frac{1}{5}\right)^{5/4}$$

To conclude, we have the lower bound of $\binom{\rho}{\rho/2+\sqrt{\rho}}$,

$$\binom{\rho}{\rho/2+\sqrt{\rho}} > \frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^2\cdot 5^5} \tag{12}$$

In fact, this lower bound can be further improved to $\frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^6}$ when ρ is sufficiently large.

Then, plugging (12) into (9),

$$\begin{aligned}
&\Pr_{y\sim U_n}\left[\ \rho/2+\sqrt{\rho}/2 \leq (\#\text{ of 0's in last }\rho\text{ bits of }y) \leq \rho/2+\sqrt{\rho}\ \right] \\
\geq &\frac{\sqrt{\rho}}{2^{\rho+1}}\cdot\binom{\rho}{\rho/2+\sqrt{\rho}} \\
\geq &\frac{\sqrt{\rho}}{2^{\rho+1}}\cdot\frac{2^{\rho+1}}{\sqrt{2\pi\rho}\cdot e^2\cdot 5^5} = \frac{1}{\sqrt{2\pi}\cdot e^2\cdot 5^5} = \Omega(1)
\end{aligned} \tag{13}$$

Therefore,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\right] \\
\geq &\mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y, y+m(\rho)\big)\ \middle|\ \rho/2+\sqrt{\rho}/2 \leq (\#\text{ of 0's in last }\rho\text{ bits of }y) \leq \rho/2+\sqrt{\rho}\ \right] \\
&\cdot \Pr_{y\sim U_n}\left[\ \rho/2+\sqrt{\rho}/2 \leq (\#\text{ of 0's in last }\rho\text{ bits of }y) \leq \rho/2+\sqrt{\rho}\ \right] \\
\geq &\sqrt{\rho}\cdot\Omega(1) = \Omega(\sqrt{\rho})
\end{aligned}$$

The expected edit distance between y and $y+m(\rho)$ is $\Omega\left(\sqrt{\rho}\right)$. □

4 Experimental results

This work presents analytical results. This is the first theoretical study on metric embeddings between metric spaces of combinatorial nature. The hope is that it will provide machinery in areas such as in image processing, speech recognition, data mining, and natural language processing (NLP).

In this section, we take one of the mainstream applications nowadays in NLP, machine translation, and verify that the setting where our theoretical developments happen are consistent with this setting.

We consider all sentences of bounded length in one human language as a finite metric space. We impose metric structure by choosing edit (or Levenshtein) distance [Lev66] to measure the distance between sentences. Other scoring functions commonly used in machine translation won't be appropriate because they fail to satisfy the symmetry and triangle inequality. We find rather intriguing the fact that edit distance is able to reveal very nontrivial geometry in human and machine translations. This human language can be a natural (perfect) human language such as French, English, but can also be a machine generated language, such as translations of a natural language by a translation engine, e.g. Google translate. The translation of sentences between two languages now becomes a metric embedding.

In our setting, the source metric space is Chinese language and the target metric space is English. The mapping is translating from a Chinese sentence into an English sentence. The experts are Bing translate [Bin] (A_0) from Chinese to English and Google translate [Goo] (A_1). Therefore, we have three metric mappings: (i) from Chinese natural language to English natural language; (ii) from Chinese natural language to Google translation; (iii) from Chinese natural language to Bing translation.

4.1 Data sets

Experiment	Data	Original Words	Words	Vocabulary
Natural	Chinese	2752	2624	479
Language	English Reference	2755	2748	481
Google	Chinese	2752	2747	480
	Google Translation	2542	2536	444
Bing	Chinese	2752	2746	480
	Bing Translation	2554	2548	369

Table 1: IWSLT 2011 corpus statistics in three experimental settings: natural language, Google and Bing translations, on running words before and after filtering as well as vocabulary size in thousand units [K].

The data is provided from IWSLT [IWS] competition with 139K sentence pairs in Chinese and English. The details of this corpus can be found

in Table 1. We used Google and Bing online translation engine application to obtain the English machine translation results. To eliminate the effect of noise in data, such as incomplete sentences, and other sporadic translation anomalies (distortion is a worst-case quantity), we show the distortion value on selected subsets of the data disregarding sentences triggering high distortion. This happens either due to errors in the original corpus, or due to large difference in the sentence length between two languages. For filtering, we use simple greedy heuristics to compute these sets which in reality could be enough larger (finding these sets can be somewhat easily shown to be NP-hard by a reduction from Independent Set and it is not clear if any reasonable approximation exists – recall that independent set itself is vastly inapproximable). We remove the sentence pair with the highest value on contraction or expansion and re-calculate the distortion iteratively.

4.1.1 Bi-Lipschitz condition (low-distortion)

Our *metric embeddings in learning* developments in the main paper is based on the Bi-Lipschitz condition only. In this subsection, we show that translation within our setting satisfies this condition (at least within some relaxed notion of locality), thus our theory is very naturally related to language translation. The term bi-Lipschitz is used informally to refer to small distortion. The distortion is the product of *expansion* and *contraction*, where the contraction is the maximum value among each distance of two Chinese sentences divided by the distance their English translations; and the expansion is the maximum value among each distance of the two translated English sentences divided by the distance of the corresponding sentences in Chinese. Here, bi-Lipschitz means that for every two Chinese sentences, which are close in edit distance their English translations are also close, and vice-versa.

Figure 1 depicts the expansion, contraction, and distortion on the subsets of sentences selected based on their lengths, ranging from 1 to 100. Each subset is composed of 1000 sentence pairs (sampled uniformly random) with no more or less than two of the given length in average, when filtering out 10% of the sentences (which includes noise).

Shorter sentences tend to have higher distortion, because of the sensitivity of distortion on the number of words. For instance, two far related single-word Chinese idioms ("众说纷纭" and "接下来") are translated into an English sentence with eight and six words ("many things have been said about this ." and "now follow me , ok ? ") , respectively, their distance is two in Chinese and fourteen in English.

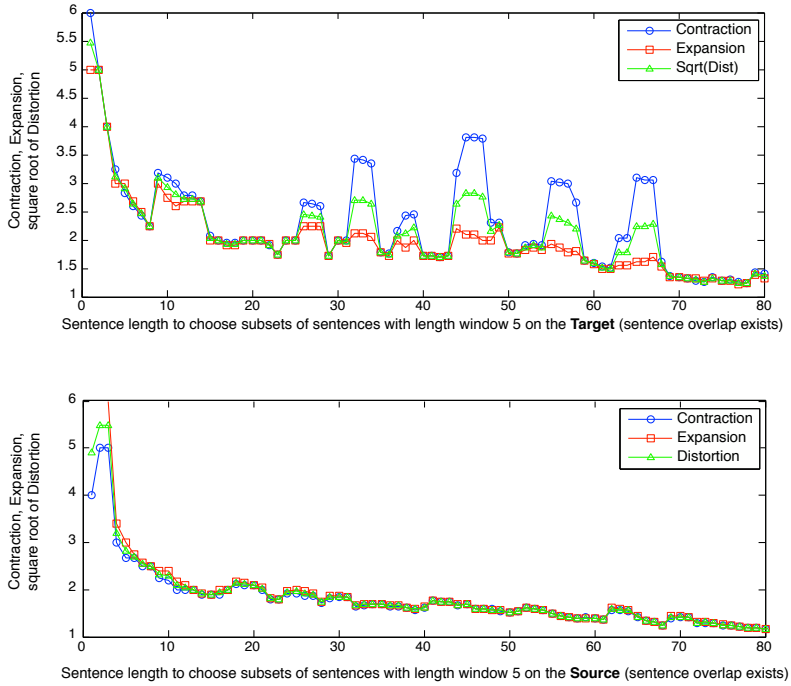

Figure 1: Distortion in sentence length range for the *natural language*. The x-axis depicts the window position x for sentence length in the window $[x-2, x+2]$. The y-axis is the distortion occuring in the corresponding window.

Our empirical study revealed two more conditions satisfied in translation setting: Metric Separability and Dense Neighboring. These conditions are not used in the statements of our theorems. However, we remark that *all of our constructions of the concept classes and the experts in our theorems do satisfy these conditions as well.* Further understanding in the geometry of human translation maps can lead to new classification techniques.

4.1.2 Metric separability condition

We partition the translation output in Table 1 into two sets: one that contains the translated sentences generated using Bing that has smaller edit distance than those generated using Google (this is the set: "Bing-better"); the other contains all translations generated using Google which has smaller

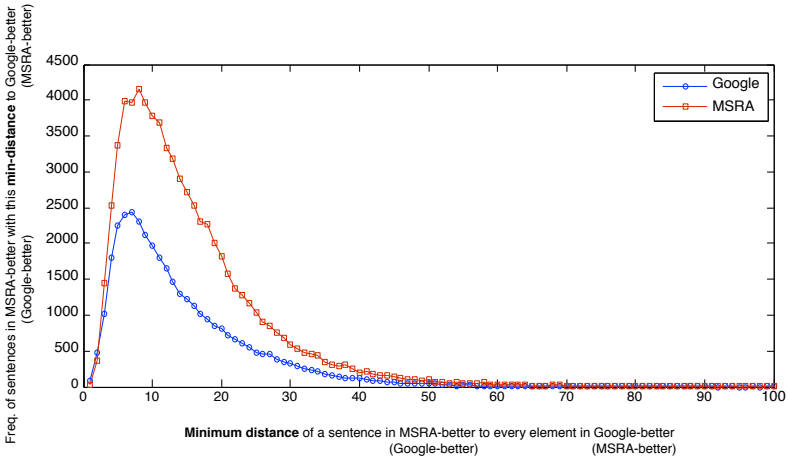

Figure 2: Frequency of sentences in Google or Bing better performed set given the mininum Levenshtein distance of this sentence from the other. Google-better: the set where Google does better than Bing; Bing-better: the set where Bing does better than Google. The x-axis is the minimum distance of Bing-better (Google-better) to Google-better (Bing-better). The y-axis is the number of sentences in the Bing-better (Google-better) set with the corresponding minimum distance from Google-better (Bing-better).

distance than the Bing one (Google-better).

There is a large subset of Bing-better that has significant distance from Google-better set (where distance between two sets is the min-distance of a point in one and a point in the other). This means that for every sentence in Bing-better, its minimum distance to all sentences in Google-better set is greater than a threshold, i.e. there is a certain distance from each sentence in Bing-better to the Google-better. Therefore Bing-better and Google-better set can be separated in a metric/geometric sense.

Figure 2 depicts the histogram of the number of sentences given its minimum distance to the target set. An important experimental finding is that the better performed translations are concentrated close to the area that has the minimum distance of 10 to the other set. The curve for Bing as in Figure 2 is greater than that for the Google, and is pushed slightly to the area with longer distances. This result indicates a strong geometric property: it is not only the case that there are large sets where Bing outperforms Google and vice-versa, but also when they do so they do it for sentences that are far away in an edit distance sense.

4.1.3 Dense neighboring condition

Consider one of the two systems performing better and the other performing worse for each sentence in a maximum subset of source sentences. For this subset if we translate a sentence with the worse system, then we can always find another sentence in the same subset whose translation is close to the first translation.

In our experiments, for each source sentence better translated with Google, we apply Bing to translate it as well. For this translation, we look for the closest translation generated by Google, whose source sentence is translated better by Google than by Bing with the respect to the edit distance. The results show that 10% of the worse translated sentences by Bing has a neighborhood with less than or equal to 5 in edit distance, where as a neighborhood we take a good translated sentence by Google. For edit distance 10 the ; and 50% of sentences with a edit distance of 10. Google versus Bing shows analogous result. We believe (and the tendency of the graph is such) that the percentage of sentences fulfilling this assumption will increase by increasing the corpus size.

References

[Bin] Bing. Bing online machine translation system.

[Goo] Google. Google translate online machine translation system.

[IWS] IWSLT. International workshop on spoken language translation 2011.

[Lev66] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.