

# Supplemental Material:

## Learning Over Large Alphabets

### Appendix A: A Proof for Theorem 1

We first introduce an important property that is used in our analysis.

**Proposition 1** *Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$ ,  $r \geq 1$  be a positive integer and  $n \in \mathbb{N}_+$ . Then the following holds,*

$$\sum_{u \in \mathcal{X}} p^r(u) e^{-np(u)} \leq \frac{(r-1)!}{n^{r-1}}. \quad (1)$$

**Proof.** Let  $X \sim p$  and define a random variable  $T(x) = \frac{(np(x))^{r-1} e^{-np(x)}}{(r-1)!}$ . Notice that  $T(x)$  is a Poisson distribution,  $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ , with a parameter  $\lambda = np(x)$  and  $k = r-1$ . Therefore,  $T(x) \in [0, 1]$ . The expected value of  $T(x)$  satisfies

$$\mathbb{E}T(X) = \sum_{x \in \mathcal{X}} p(x) \frac{(np(x))^{r-1} e^{-np(x)}}{(r-1)!} = \frac{n^{r-1}}{(r-1)!} \sum_{x \in \mathcal{X}} p^r(x) e^{-np(x)} \leq 1 \quad (2)$$

where the inequality follows from  $T(x) \in [0, 1]$ . ■

Let us now derive the bound for the risk.

$$\begin{aligned} \mathbb{E}_{X^n \sim p} (\hat{M}_\beta(X^n) - M_0(X^n))^2 &= \\ \mathbb{E}_{X^n \sim p} \left( \sum_{u \in \mathcal{X}} \mathbb{1}\{N_u = 0\} p(u) - \sum_{u \in \mathcal{X}} \mathbb{1}\{N_u = 0\} \beta \right)^2 &= \\ \sum_{u \in \mathcal{X}} \sum_{v \in \mathcal{X}} P_n(0, 0) (p(u) - \beta)(p(v) - \beta) \end{aligned} \quad (3)$$

where  $P_n(i, j) = \mathbb{E}_{X^n \sim p} (\mathbb{1}(N_u = i) \mathbb{1}(N_v = j))$ , and

$$P_n(i, j) = \begin{cases} \binom{n}{i, j} p^i(u) p^j(v) (1 - p(u) - p(v))^{n-i-j} & u \neq v \\ \binom{n}{i} p^i(u) (1 - p(u))^{n-i} & u = v, i = j \\ 0 & u = v, i \neq j \end{cases} \quad (4)$$

Plugging the definition of  $P_n(i, j)$  (4) to the estimation risk (3), we obtain

$$\begin{aligned} \mathbb{E}_{X^n \sim p} \left( \hat{M}_\beta(X^n) - M(X^n) \right)^2 = & \quad (5) \\ & \sum_{u \neq v} (1 - p(u) - p(v))^n (p(u) - \beta)(p(v) - \beta) + \sum_{u \in \mathcal{X}} (1 - p(u))^n (p(u) - \beta)^2 = \\ & \beta^2 \sum_{u \neq v} (1 - p(u) - p(v))^n \\ & - \beta \sum_{u \neq v} (1 - p(u) - p(v))^n (p(u) + p(v)) \\ & + \sum_{u \neq v} (1 - p(u) - p(v))^n p(u) p(v) \\ & + \sum_{u \in \mathcal{X}} (1 - p(u))^n (p(u) - \beta)^2 \end{aligned}$$

Let us separately study each of the terms in (5).

Let us examine the first term in (5):  $\beta^2 \sum_{u \neq v} (1 - p(u) - p(v))^n$ . We have

$$\beta^2 \sum_{u \neq v} (1 - p(u) - p(v))^n \leq \beta^2 \sum_{u \neq v} e^{-n(p(u) + p(v))} \quad (6)$$

where the inequality follows from

$$(1 - t)^n \leq e^{-nt} \quad (7)$$

for any  $t \in \mathbb{R}$  and  $n \in \mathbb{R}_+ [1]$ .

Let us examine the second term of (5):  $-\beta \sum_{u \neq v} (1 - p(u) - p(v))^n (p(u) + p(v))$ . Following from the inequality

$$(1 - t)^n \geq e^{-nt} (1 - nt^2) \quad (8)$$

for any  $0 \leq t \leq 1$  and  $n \in \mathbb{N}_+ [1, 2]$ , we get that

$$\begin{aligned} -\beta \sum_{u \neq v} (1 - p(u) - p(v))^n (p(u) + p(v)) &\leq -\beta \sum_{u \neq v} e^{-n(p(u)+p(v))} (1 - n(p(u) + p(v))^2) (p(u) + p(v)) \quad (9) \\ &= -\beta \sum_{u \neq v} e^{-n(p(u)+p(v))} (p(u) + p(v)) + \beta n \sum_{u \neq v} e^{-n(p(u)+p(v))} (p(u) + p(v))^3 \end{aligned}$$

Using Proposition 1, we can find an bound on  $\beta n \sum_{u \neq v} e^{-n(p(u)+p(v))} (p(u) + p(v))^3$ :

$$\begin{aligned} \beta n \sum_{u \neq v} e^{-n(p(u)+p(v))} (p(u) + p(v))^3 &= \quad (10) \\ \beta n \sum_{u \neq v} e^{-n(p(u)+p(v))} (p^3(u) + 3p^2(u)p(v) + 3p(u)p^2(v) + p^3(v)) &\leq \\ \beta n \sum_{u \in \mathcal{X}} e^{-np(u)} \sum_{v \in \mathcal{X}} e^{-np(v)} (p^3(u) + 3p^2(u)p(v) + 3p(u)p^2(v) + p^3(v)) &\leq \\ \beta n \left( \frac{4}{n^2} \sum_{u \in \mathcal{X}} e^{-np(u)} + \frac{6}{n} \right) &= \\ \frac{4\beta}{n} \sum_{u \in \mathcal{X}} e^{-np(u)} + 6\beta &\leq \\ 2\beta \left( \frac{2m}{n} + 3 \right) \end{aligned}$$

Let us examine the third term of (5):  $\sum_{u \neq v} (1 - p(u) - p(v))^n p(u)p(v)$ . We have

$$\sum_{u \neq v} (1 - p(u) - p(v))^n p(u)p(v) \leq \sum_{u \neq v} e^{-n(p(u)+p(v))} p(u)p(v) \quad (11)$$

where the inequality follows from (7).

So far we have

$$\begin{aligned} \mathbb{E}_{X^n \sim p} \left( \hat{M}_\beta(X^n) - M(X^n) \right)^2 &\leq \quad (12) \\ 2\beta \left( \frac{2m}{n} + 3 \right) + \\ \sum_{u \neq v} e^{-n(p(u)+p(v))} (\beta^2 - \beta(p(u) + p(v)) + p(u)p(v)) & \\ + \sum_{u \in \mathcal{X}} (1 - p(u))^n (p(u) - \beta)^2 & \end{aligned}$$

The term  $\sum_{u \neq v} e^{-n(p(u)+p(v))}(\beta^2 - \beta(p(u) + p(v)) + p(u)p(v))$  can be bounded by a simpler bound:

$$\begin{aligned}
& \sum_{u \neq v} e^{-n(p(u)+p(v))}(\beta^2 - \beta(p(u) + p(v)) + p(u)p(v)) \leq \\
& \sum_{u \in \mathcal{X}} e^{-np(u)} p(u) \sum_{v \in \mathcal{X}} e^{-np(v)} p(v) \\
& - \beta \sum_{u \in \mathcal{X}} e^{-np(u)} p(u) \sum_{v \in \mathcal{X}} e^{-np(v)} \\
& - \beta \sum_{u \in \mathcal{X}} e^{-np(u)} \sum_{v \in \mathcal{X}} e^{-np(v)} p(v) \\
& + 2\beta \sum_{u \in \mathcal{X}} e^{-2np(u)} p(u) \\
& + \beta^2 \sum_{u \in \mathcal{X}} e^{-np(u)} \sum_{v \in \mathcal{X}} e^{-np(v)} = \\
& \left( \sum_{u \in \mathcal{X}} p(u) e^{-np(u)} - \beta \sum_{u \in \mathcal{X}} e^{-np(u)} \right)^2 + 2\beta \sum_{u \in \mathcal{X}} e^{-2np(u)} p(u) = \\
& \left( \sum_{u \in \mathcal{X}} (p(u) - \beta) e^{-np(u)} \right)^2 + 2\beta \sum_{u \in \mathcal{X}} e^{-2np(u)} p(u) \leq \\
& \left( \sum_{u \in \mathcal{X}} (p(u) - \beta) e^{-np(u)} \right)^2 + 2\beta
\end{aligned} \tag{13}$$

Where the last inequality stems from Proposition 1. Next, we study the last term in (5). Following (7) we get:

$$\sum_{u \in \mathcal{X}} (1 - p(u))^n (p(u) - \beta)^2 \leq \sum_{u \in \mathcal{X}} e^{-np(u)} (p(u) - \beta)^2 \tag{14}$$

From (12), (13) and (14) we get Theorem 1:

$$\mathbb{E}_{X^n \sim p} \left( \hat{M}_\beta(X^n) - M(X^n) \right)^2 \leq 2\beta \left( \frac{2m}{n} + 4 \right) + \left( \sum_{u \in \mathcal{X}} (p(u) - \beta) e^{-np(u)} \right)^2 + \sum_{u \in \mathcal{X}} e^{-np(u)} (p(u) - \beta)^2 \tag{15}$$

## Appendix B

**Theorem 2** Let  $p$  be a probability distribution over a countable alphabet  $\mathcal{X}$  of size  $m < \infty$ . Let  $f_{n,\beta} = \sum_{u \in \mathcal{X}} e^{-np(u)} (p(u) - \beta)^2$  and  $g_{n,\beta} = \sum_{u \in \mathcal{X}} e^{-np(u)} (p(u) - \beta)$ . Let  $f_{n,\beta}^{max} = \max_{p \in \Delta_m} f_{n,\beta}$ ,  $g_{n,\beta}^{max} = \max_{p \in \Delta_m} g_{n,\beta}$  and  $p_f^* = \arg \max_{p \in \Delta_m} f_{n,\beta}$ ,  $p_g^* = \arg \max_{p \in \Delta_m} g_{n,\beta}$ . Then, the following holds:

- For  $g_{n,\beta}^{max}$ :

1. There exists no more than a single probability value  $p_g^*(u)$  such that  $p_g^*(u) \in \left(\frac{2}{n} + \beta, 1\right]$ .
2. If  $p_g^*(u), p_g^*(v) \in \left[0, \frac{2}{n} + \beta\right]$ , then  $p_g^*(u) = p_g^*(v)$ .

- For  $f_{n,\beta}^{max}$ :

1. There exist  $m_0 < m$  probability values such that  $p_f^*(u) = 0$ .
2. There exists at most a single probability value  $p_f^*(u)$  such that  $p_f^*(u) \in \left(0, \frac{2-\sqrt{2}}{n} + \beta\right)$ .
3. There exist  $m_1 < m$  probability values such that  $p_f^*(u) \in \left[\frac{2-\sqrt{2}}{n} + \beta, \frac{2+\sqrt{2}}{n} + \beta\right]$ . Furthermore,  $p_f^*(v) = p_f^*(u)$  for all  $p_f^*(v), p_f^*(u) \in \left[\frac{2-\sqrt{2}}{n} + \beta, \frac{2+\sqrt{2}}{n} + \beta\right]$ .
4. There exists at most a single probability value  $p_f^*(u)$  such that  $p_f^*(u) \in \left(\frac{2+\sqrt{2}}{n} + \beta, 1\right]$ .

**Proof.** Let us first study the function  $f = e^{-np}(p - \beta)^2$ :

$$\frac{\partial f}{\partial p} = -ne^{-np}(p - \beta)^2 + 2e^{-np}(p - \beta) = 0$$

$$e^{-np}(p - \beta)(2 - np + n\beta) = 0$$

$f$  is non-negative, therefore  $p_{min} = \beta$  which causes  $f$  to be 0 is a minimum. The other extremum is  $p_{max} = \frac{2}{n} + \beta$  - it is a maxima, because  $p_{min} < p_{max}$  and  $f(p_{min}) < f(p_{max})$ , meaning that the function is increasing in the range  $[p_{min}, p_{max}]$ , and since there are no other extremum and since the limit of the function approaches 0 as  $p$  approaches infinity, that means that the function is decreasing for  $p > p_{max}$ .

$$\frac{\partial^2 f}{\partial p^2} = n^2 e^{-np}(p - \beta)^2 - 4ne^{-np}(p - \beta) + 2e^{-np} = 0$$

$$e^{-np}(n^2(p - \beta)^2 - 4n(p - \beta) + 2) = 0$$

define  $t = (p - \beta)$ . Find the solutions to this equation:

$$t_{1,2} = \frac{4n \pm \sqrt{16n^2 - 8n^2}}{2n^2} = \frac{2 \pm \sqrt{2}}{n}$$

We get:  $p_1 = \frac{2+\sqrt{2}}{n} + \beta, p_2 = \frac{2-\sqrt{2}}{n} + \beta$

Therefore,  $f$  is convex in the range  $[0, p_2]$ , concave in the range  $(p_2, p_1)$ , and convex in the range  $(p_1, 1]$ .

**Proposition 2** *An optimal solution includes no more than one value in the range  $(0, p_2)$ .*

**Proof.** To prove it by contradiction, let us assume that an optimal solution with distribution  $p$  has a subset of values  $U$  such that for each  $u$  in  $U$ ,  $p(u)$  is in the range  $(0, p_2)$  with at least 2 values in the subset. Define the sum of these values  $\sum_{u \in U} p(u) = C$ . Let  $u_1$  and  $u_2$  be two values in  $U$ .

If  $0 \leq p(u_1) \leq \beta \leq p(u_2) \leq p_2$  (without loss of generality), then due to the convex nature of  $f$  in the range  $(0, p_2)$ , a probability distribution  $p'$  where  $p'(v) = p(v)$  for  $v \notin \{u_1, u_2\}$ ,  $p'(u_1) = p(u_1) - \min\{p(u_1), \beta - p(u_2)\}$ ,  $p'(u_2) = p(u_2) + \min\{p(u_1), \beta - p(u_2)\}$  yields a solution where the constraint

$\sum_{u \in U} p'(u) = C$  is satisfied, with a higher value than the optimal solution, in contradiction to it being optimal (maximal).

If  $0 \leq p(u_1) \leq p(u_2) \leq \beta$  (without loss of generality), then let  $p'$  be a probability distribution where  $p'(v) = p(v)$  for  $v \notin \{u_1, u_2\}$ , and if  $p(u_1) + p(u_2) \geq p_2$  then  $p'(u_1) = p(u_1) - p_2 + p(u_2)$ ,  $p'(u_2) = p_2$ , otherwise  $p'(u_1) = 0$ ,  $p'(u_2) = p(u_2) + p(u_1)$ . In both cases, the constraint  $\sum_{u \in U} p'(u) = C$  is satisfied.

By the definition of a convex function, for all  $x, y$  in the convex domain, and all  $\lambda \in [0, 1]$ , we have  $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ . Therefore, we also have  $f(\lambda x + (1 - \lambda)y) + f(\lambda y + (1 - \lambda)x) \leq f(x) + f(y)$ . In the first case, if we choose  $x = p'(u_1) = p(u_1) + p(u_2) - p_2$ ,  $y = p'(u_2) = p_2$  and  $\lambda = \frac{p(u_1) - p_2}{p(u_1) + p(u_2) - 2p_2}$ , we get  $f(p(u_1)) + f(p(u_2)) \leq f(p'(u_1)) + f(p'(u_2))$ , meaning that for  $p'$  we get a higher value than the optimal solution, in contradiction to it being optimal. Similarly, in the second case, if we choose  $x = p'(u_1) = 0$ ,  $y = p'(u_2) = p(u_1) + p(u_2)$ ,  $\lambda = \frac{p(u_1)}{p(u_1) + p(u_2)}$ , we get the same result. ■

**Proposition 3** *All values in the range  $[p_2, p_1]$  are equal in an optimal solution.*

**Proof.** To prove it by contradiction, let us assume that an optimal solution with distribution  $p$  has a subset of values  $U$  such that for each  $u$  in  $U$ ,  $p(u)$  is in the range  $[p_2, p_1]$  with at least 2 values in the subset. Define the sum of these values  $\sum_{u \in U} p(u) = C$ . Let  $u_1$  and  $u_2$  be two values in  $U$  so that  $p(u_1) \neq p(u_2)$ .

By the definition of a concave function, for all  $x, y$  in the concave domain, and all  $\lambda \in [0, 1]$ , we have  $f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$ . Therefore, we also have  $f(\lambda x + (1 - \lambda)y) + f(\lambda y + (1 - \lambda)x) \geq f(x) + f(y)$ . If we choose  $x = p(u_1)$ ,  $y = p(u_2)$  and  $\lambda = 0.5$ , we get  $2f(\frac{p(u_1) + p(u_2)}{2}) \geq f(p(u_1)) + f(p(u_2))$ , meaning that for  $p'$  where  $p'(v) = p(v)$  for  $v \notin \{u_1, u_2\}$ ,  $p'(u_1) = p'(u_2) = \frac{p(u_1) + p(u_2)}{2}$ , we get that  $p'$  provides larger value than the optimal solution, in contradiction to it being optimal. ■

**Proposition 4** *An optimal solution includes no more than one value in the range  $(p_1, 1]$ .*

**Proof.** To prove it by contradiction, let us assume that an optimal solution with distribution  $p$  has a subset of values  $U$  such that for each  $u$  in  $U$ ,  $p(u)$  is in the range  $(p_1, 1]$  with at least 2 values in the subset. Define the sum of these values  $\sum_{u \in U} p(u) = C$ . Let  $u_1$  and  $u_2$  be two values in  $U$ .

Without loss of generality, assume  $p_1 \leq p(u_1) \leq p(u_2) \leq 1$ , then due to the convex nature of  $f$  in the range  $(p_1, 1]$ , a probability distribution  $p'$  where  $p'(v) = p(v)$  for  $v \notin \{u_1, u_2\}$ ,  $p'(u_1) = p_1$ ,  $p'(u_2) = p(u_2) + p(u_1) - p_1$  we would get a solution where the constraint  $\sum_{u \in U} p'(u) = C$  is satisfied, with a higher value than the optimal solution, in contradiction to it being optimal (maximal). Because  $p(u_1) + p(u_2) \leq 1$ , so must  $p'(u_2) \leq 1$ , so the distribution  $p'$  is valid. ■

From propositions 2, 3, 4 we get that the possible optimal probability has the four following degrees of freedom:

1. Number of elements  $u$  for which  $p(u)=0$ .
2. The (single) value of all elements in the range  $p(u) \in [p_2, p_1]$ .

3. The value of the (single) element in the range  $(0, p_2)$ , if it exists.
4. The value of the (single) element in the range  $(p_1, 1]$ , if it exists.

Let us now analyze  $g = e^{-np}(p - \beta)$ :

$$\frac{\partial g}{\partial p} = e^{-np}(1 - np + n\beta) = 0$$

We get an extremum at  $p = \frac{1}{n} + \beta$ .

$$\frac{\partial^2 g}{\partial p^2} = e^{-np}n(np - 2 - n\beta) = 0$$

We get a critical point at  $p = \frac{2}{n} + \beta$ . At the extremum  $p = \frac{1}{n} + \beta$  we get that the second derivative is negative, so it is concave for  $p < \frac{2}{n} + \beta$  and therefore the extremum is a maximum. The function is convex for  $p > \frac{2}{n} + \beta$ .

This means that for the probability  $p$  which maximizes  $g_{n,\beta}$ :

1. There is no more than one value in the range  $(\frac{2}{n} + \beta, 1]$ . (see proposition 4).
2. All values in the range  $[0, \frac{2}{n} + \beta]$  must be equal. (see proposition 3).

■

It follows from Theorem 1 that the estimation risk is bounded from above by

$$2\beta\left(\frac{2m}{n} + 3\right) + (g_{n,\beta}^{max})^2 + f_{n,\beta}^{max} \quad (16)$$

For every  $\beta$  that satisfies the conditions of Theorem 1 we attain a missing mass estimation bound, that holds for every probability distribution over an alphabet of size  $m$ . Thus, we examine different values, and seek the lowest risk bound. Algorithm 1 provides a pseudo-code for our suggested scheme.

---

**Algorithm 1** Missing mass estimation bound for a known alphabet size

---

**Require:**  $m, n$

- 1: **for**  $\beta$  that satisfies the conditions of Theorem 1 (specifically,  $0 \leq \beta \leq \frac{1}{m}$ ) **do**
  - 2:   Numerically evaluate  $f_{n,\beta}^{max}$  and  $g_{n,\beta}^{max}$  according to Theorem 2.
  - 3:   Compute the upper bound according to (16).
  - 4: **end for**
  - 5: **return** the lowest upper bound and the  $\beta_0$  that attains it.
- 

## References

- [1] D. S. Mitrinovic and P. M. Vasic, *Analytic inequalities*. Springer, 1970, vol. 61.

- [2] C. P. Niculescu and A. Vernescu, “A two-sided estimate of  $e^{x-(1+x/n)^n}$ ,” *Journal of Inequalities in Pure and Applied Mathematics*, vol. 5, no. 3, 2004.