Probability Estimation in Statistical Natural Language Processing

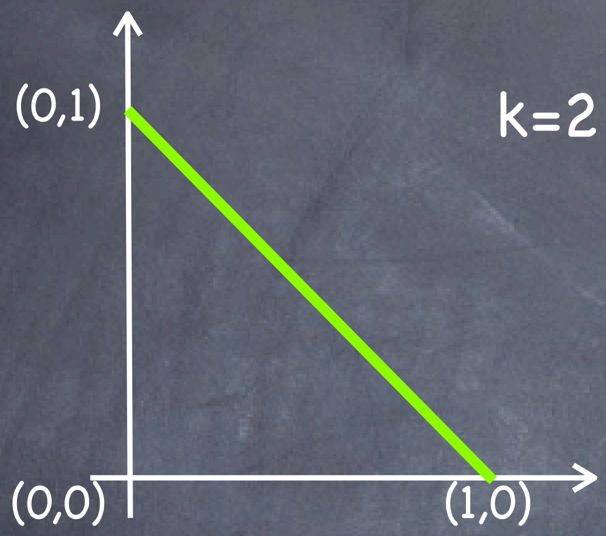
Sanjeev Khudanpur
(co-workers: B. Jedynak, D. Karakos, A. Yazgan)
Department of Electrical & Computer Engineering
Center for Language & Speech Processing
Johns Hopkins University, Baltimore, MD, U.S.A.

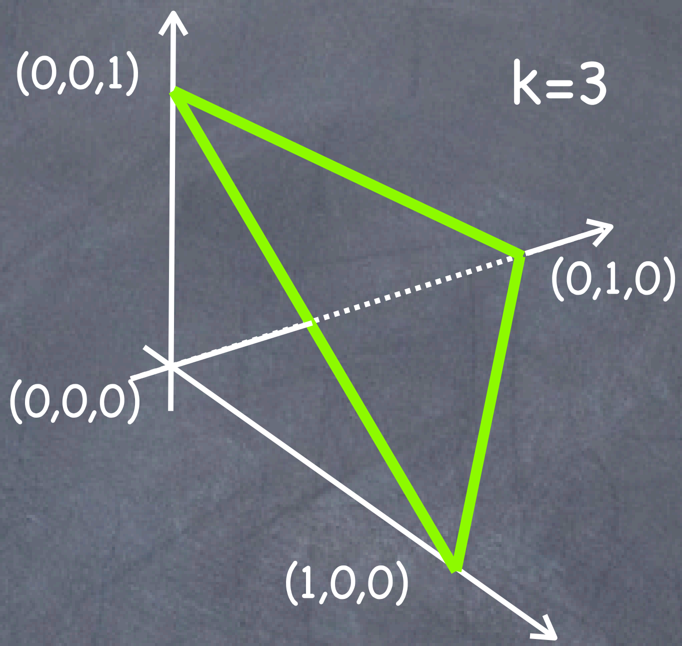
July 13, 2005

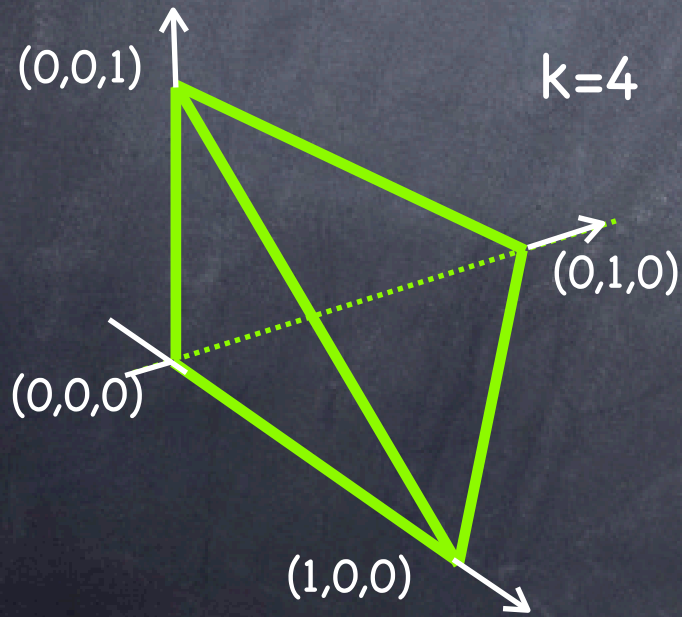
The Density Estimation Problem

- Consider a random variable taking values in $\mathcal{X}$ = {1,2,...,k}
- Let p(.) denote a probability mass function on $\mathcal{X}$
  - $0 \leq p(x) \leq 1$ for all x in $\mathcal{X}$, and $\Sigma_{x \in \mathcal{X}}\, p(x) = 1$
  - p(.) is usually unknown, and needs to be estimated from some sample data
- Let the samples $x_1, x_2, \ldots, x_n$ be drawn independently, each according to p(.)
- An estimator of p(.) is a function $\hat{p} : \mathcal{X}^n \rightarrow \mathcal{P}^k$

The k-Dimensional Unit Simplex







In general, it is

- a (k-1)—dim hyperplane
- restricted to the +ve orthant
- closed, bounded and convex

Likelihood of the Observed Data

- The likelihood of observing $x_1,\ldots,x_n$ under a probability mass function or pmf p is given by
- $$p(x_1,\ldots,x_n) = \prod_{t=1}^{n} p(x_t) = \prod_{x\in\mathcal{X}} [p(x)]^{n(x)}$$
- where n(x) is the number of times the value x is seen in the sample $x_1,\ldots,x_n$
- $$n(x) = \sum_{t=1}^{n} \mathbf{1}(x_t = x)$$
- Note: permuting $x_1,\ldots,x_n$ does not change its likelihood

Types and Typical Sequences

$$p(x_1,\ldots,x_n) = \prod_{x\in\mathcal{X}} [p(x)]^{n(x)} = \exp\left\{\log \prod_{x\in\mathcal{X}} [p(x)]^{n(x)}\right\}$$

$$= \exp\left\{\sum_{x\in\mathcal{X}} n(x) \log p(x)\right\}$$

$$= \exp\left\{n \sum_{x\in\mathcal{X}} \frac{n(x)}{n} \log p(x)\right\}$$

A sufficient statistic

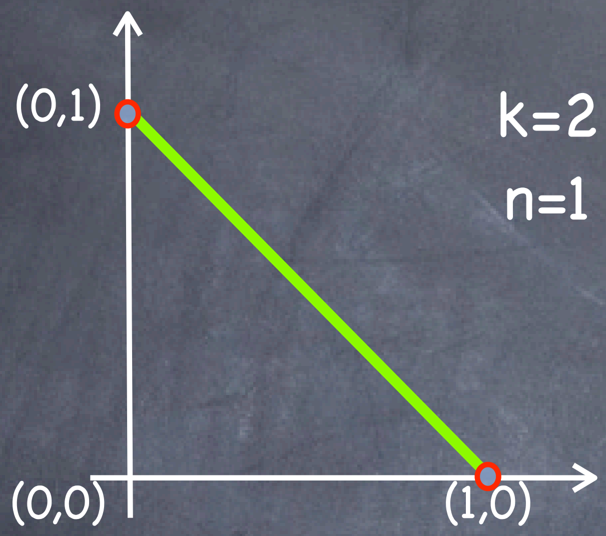
$$= \exp\left\{n \sum_{x\in\mathcal{X}} \hat{p}(x) \log p(x)\right\}$$

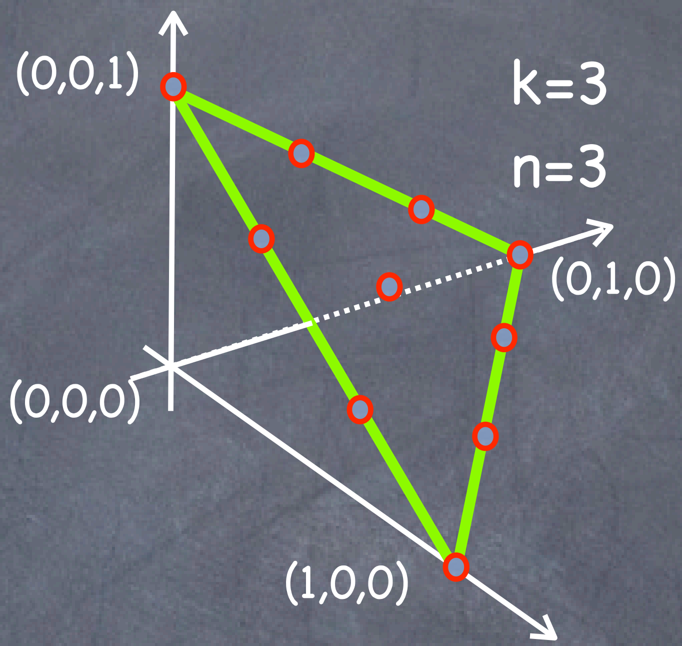
The type of a sequence is $\hat{p} \equiv \left(\frac{n(1)}{n},\ldots,\frac{n(k)}{n}\right)$

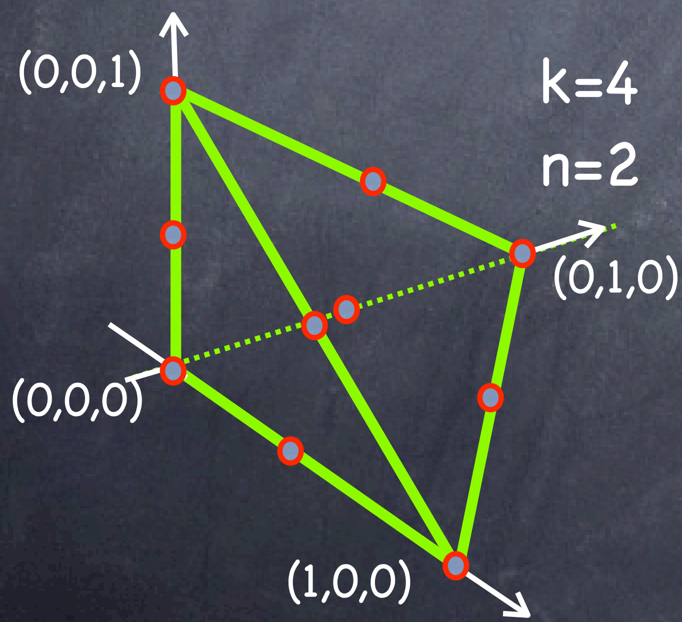
Number of sequences whose type is $\hat{p} = \dfrac{n!}{n(1)!\,n(2)!\,\cdots\,n(k)!}$

The number of distinct possible types is $\left|\mathcal{P}_n^k\right| = \binom{n+k-1}{k-1} \approx (n+1)^k$

Possible Types on the Simplex







In general, the possible types are

- reminiscent of the integer lattice in (k-1)–dimensional space
- "evenly spaced" on the simplex
- grow close together as $n \to \infty$

Likelihood, Entropy and Divergence

$$
\begin{aligned}
p(x_1, \ldots, x_n) &= \exp\left\{-n\left[\sum_{x\in\mathcal{X}} \hat{p}(x)\log\frac{1}{p(x)}\right]\right\} \\
&= \exp\left\{-n\left[-\sum_{x\in\mathcal{X}} \hat{p}(x)\log\hat{p}(x) + \sum_{x\in\mathcal{X}} \hat{p}(x)\log\frac{\hat{p}(x)}{p(x)}\right]\right\} \\
&= \exp\left\{-n\left[H(\hat{p}) + D(\hat{p}\|p)\right]\right\}
\end{aligned}
$$

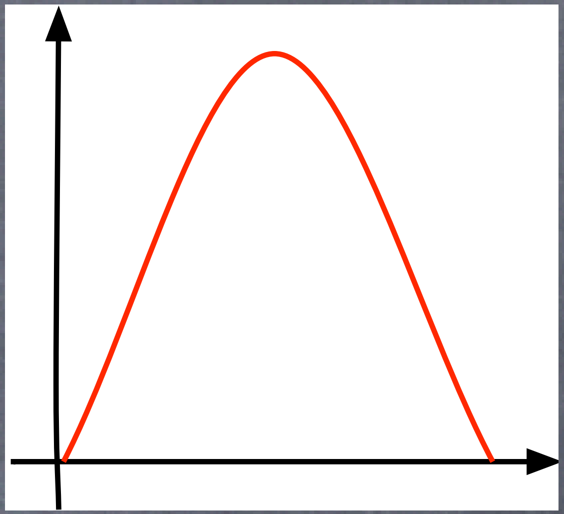
Entropy

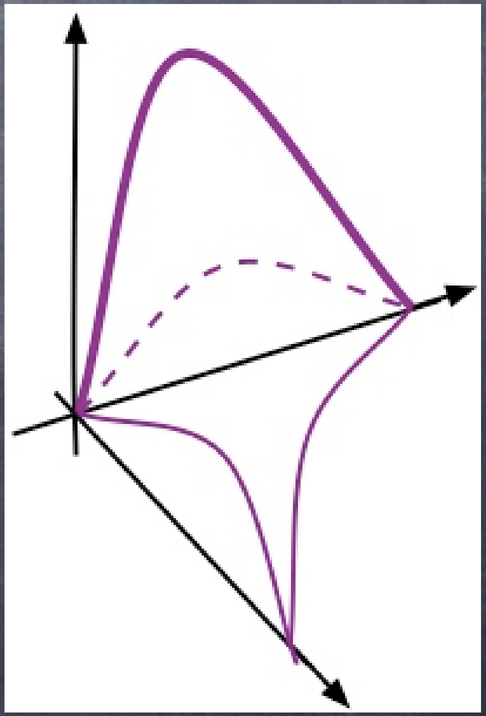
Kullback-Leibler Divergence

- Note that the sample (not the choice of p) fixes the entropy
- Therefore, if we wish to choose a p that assigns high likelihood to the observed sample, we much choose a p that is “close” in K-L divergence to $\hat{p}$

Properties of (Shannon) Entropy

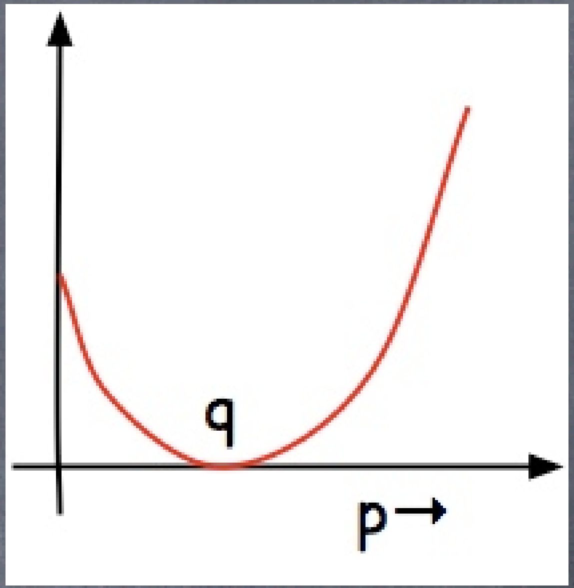
- $0 \leq H(p) \leq \log(k)$
- $H(p) = 0$ iff p is a degenrate pmf
- $H(p) = \log(k)$ iff p is the uniform pmf
- $H(p)$ is a continouous function of p
- $H(p)$ is a concave function of p
- $H(p)$ is the nonparametric analog of smoothness for continuous densities

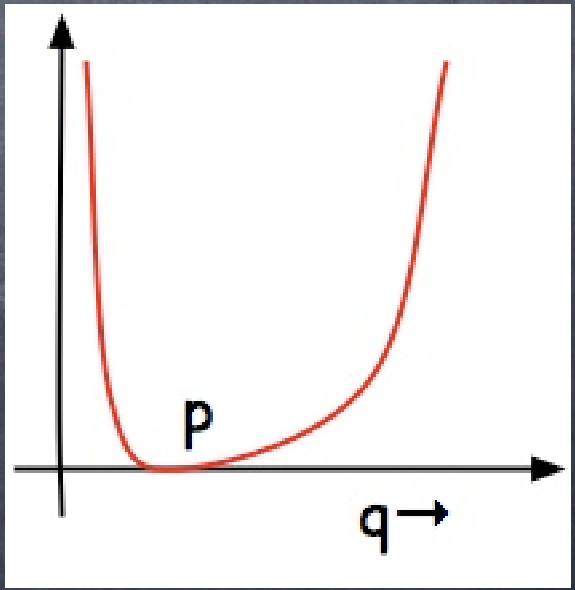




Properties of K-L Divergence

- $D(p||q) \geq 0$ with equality iff p==q
- $D(p||q)$ is continuous in (p,q)
- $D(p||q)$ is convex in (p,q)
  - If p is fixed, it is convex in q
- For the uniform pmf
  - $D(p||u) = \log(k) - H(p)$
- Maximize $H(p) \Leftrightarrow$ Minimize $D(p||u)$





Popular Density Estimates

- Maximum likelihood estimation:
  - choose the type itself as the estimate of p
- Bayesian estimation:
  - assume a (prior) probability density π on the on the simplex of pmfs, e.g. the Dirichlet density
  - assume a cost function $L(p,q)$ for estimating p as q, e.g. $||p-q||^2$
  - find the estimate q that minimizes expected cost $E_{\pi}[L(p,q)|x_1,\ldots x_n]$
  - Often leads to an "add-β" estimate $q^*(x) = \{n(x)+\beta\}/\{n+k\beta\}$
- Maximum entropy estimation:
  - Find a few marginals that may be reliably estimated from the type
  - Consider all pmfs that agree with these marginals as admissible
  - Choose the admissible pmf with the highest entropy
  - The type is always admissible, but a "smoother" pmf near it is chosen

Variations on Maximum Entropy

$$\mathcal{M} = \{p \in \mathcal{P}^k \, : \, p(A_j) = \hat{p}(A_j), \; j = 1, \ldots, J\}$$

$$p^*(x) = \frac{1}{Z(\Lambda)} \exp\left\{ \sum_{j=1}^{J} \lambda_j \mathbf{1}(x \in A_j) \right\}$$

- Enlarge the class of admissible pmfs
- $\mathcal{M} = \{p \in \mathcal{P}^k \, : \, \hat{p}(A_j) - \epsilon \leq p(A_j) \leq \hat{p}(A_j) + \epsilon, \; j = 1, \ldots, J\}$
- p* is also the ML estimate from an exponential family $\mathcal{Q}$
- Seek something other than the maximum entropy pmf in $\mathcal{M}$

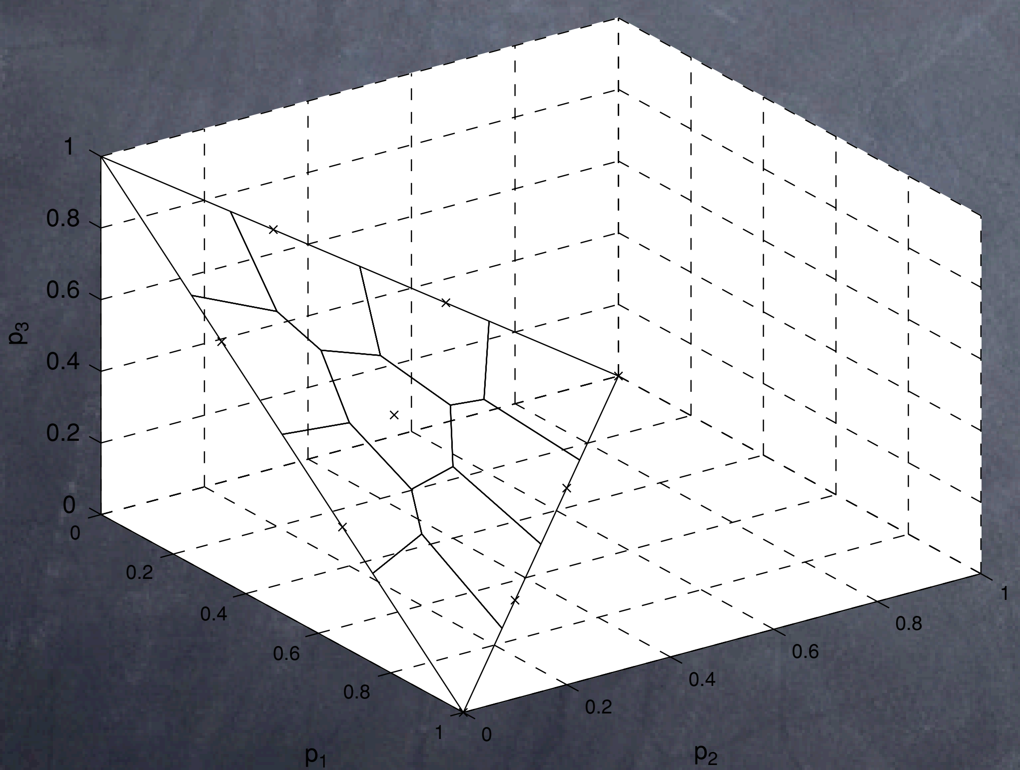
$$q^* = \arg\max_{q \in \mathcal{Q}} \left[q(x_1, \ldots, x_n) - \rho \|\Lambda\|^2 \right]$$

The Maximum Likelihood Set (new!)

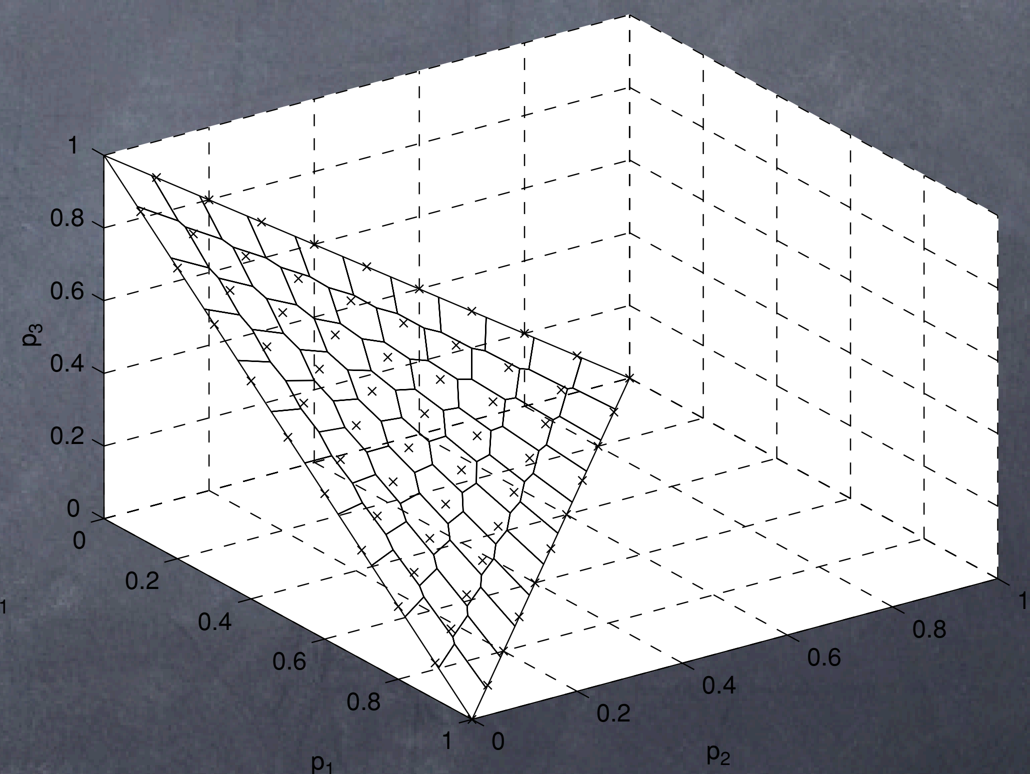
- Recall that the observed type is a sufficient statistic for estimating p from the sample data
- Recall that the type can take only a finite number of values for a finite sample size
- Define a pmf p to be admissible if it assigns a higher likelihood to the observed type than to any other type!
- $$p(\hat{p}) = \frac{n!}{n(1)! \cdots n(k)!} \prod_{x \in \mathcal{X}} [p(x)]^{n(x)}$$
-
- $$\mathcal{M} = \left\{ p \in \mathcal{P}^k \ : \ p(\hat{p}) \geq p(\hat{q}), \ \forall \ \hat{q} \in \mathcal{P}_n^k \right\}$$
- Key idea: the type we observed should be at least as likely as one we didn't

Visualizing the Max Likelihood Set

k=3 and n=3



k=3 and n=10



Characterizing the MLS

- The maximum likelihood set is equivalently given by

$$\mathcal{M} = \left\{ p \in \mathcal{P}^k \;:\; \frac{\hat{p}(x)}{\hat{p}(x') + \frac{1}{n}} \leq \frac{p(x)}{p(x')} \leq \frac{\hat{p}(x) + \frac{1}{n}}{\hat{p}(x')} \quad \forall\, x, x' \in \mathcal{X} \right\}$$

- Every MLS is a closed, bounded and convex set
  - bounded by linear hyper-planes
  - very useful when searching numerically for p*
- Every MLS contains the observed type, but no other type
  - the collection of MLS's tessellates the unit simplex
- The diameter of every MLS is O(1/n)

$$\text{For every pmf in the MLS } \|p - \hat{p}\|_1 \leq \frac{2(k-1)}{n}$$

More Properties of the MLS

- Every pmf in the MLS is a strongly consistent estimate of p
- $$\lim_{n\to\infty} \sup_{p\in\mathcal{M}} \|p - \tilde{p}\| = 0 \quad \tilde{p} - \text{almost surely}$$
- If n(x)>0, then p(x)>0 for every p in the MLS
- In every MLS, there is a pmf with p(x)>0 for all x in $\mathcal{X}$.
  - i.e. each MLS is guaranteed to contain "smooth" candidates
- Faithfulness to the observed evidence:
  - if n(x)>n(x') then, for every pmf p in the MLS, p(x)≥p(x')
  - this property isn't guaranteed for the Bayesian estimates, Good-Turing, etc.

Choosing an Estimate from the MLS

- If some reference pmf q is available (e.g. an estimate you would use for n=0), then it may be used to choose one of the admissible members of the MLS
- $$p^* = \arg\max_{p \in \mathcal{M}} D(p \| q)$$
- If no q is available, q could be assumed to be uniform
- Using this criterion to choose from the MLS has another desirable property, faithfulness to prior beliefs when the evidence is equivocal:
- $$n(x) = n(x') \quad \text{and} \quad q(x) \geq q(x') \quad \Longrightarrow \quad p^*(x) \geq p^*(x')$$
- This leads to considerable computational savings

Examples for Discussion

- The special case when n=1
- Estimating a unigram distribution for words using Zipf's law as a reference distribution
- Estimating a bigram (conditional) distribution using the unigram distribution as a reference distribution
- Implications for growing decision trees and random forests
- Implications for estimation of entropy & mutual information
- The case when $k \to \infty$ (i.e. unbounded alphabet sizes)

Concluding Remarks

- Density estimation is at the core of a lot of statistical methods in language and speech processing
- Spare data is always an issue (Google notwithstanding)
- Learning from small samples is vital; methods for incorporating structural constraints in these estimates need to be investigated further
- The MLS based estimate is a parameter-free technique that characterizes the uncertainty of the estimate, and provides a means for incorporating prior domain knowledge