Statistics G 8243
Ioannis Kontoyiannis
Tuesday, March 24, 2009
Handout #12

Summary of last lecture

Recall that, according to Sanov's theorem, if $\hat{P}_n = \hat{P}_{X_1^n}$ is the type of an i.i.d. random string $X_1^n = (X_1, X_2, \ldots, X_n)$ generated according to some distribution Q on A, then the probability that $\hat{P}_n$ will belong to a set of probability distributions E behaves exponentially,

$$Q^n(\hat{P}_n \in E) \doteq 2^{-nD(P^*\|Q)},$$

and the rate $D(P^*\|Q)$ in the exponent is determined by the relative entropy distance $D(P^*\|Q) = \inf_{P\in E} D(P\|Q)$ between Q and the set E; see Figure 1.

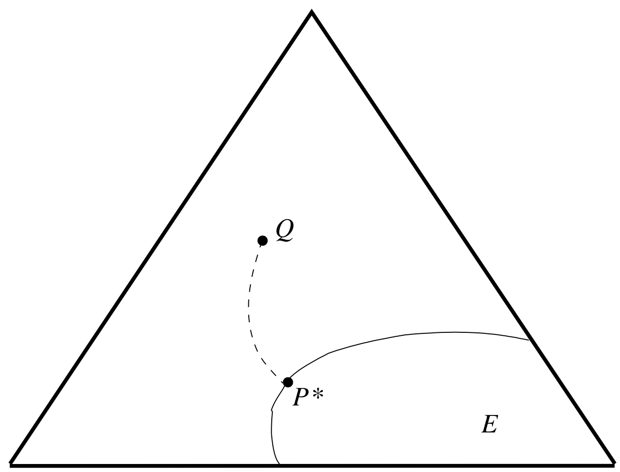

Figure 1: Representation of the projection of the distribution Q onto the set E; the point P^* achieves the infimum $\inf_{P\in E} D(P\|Q)$ in Sanov's theorem.

The following result enriches this geometric intuition, by showing that projections in terms of $D(P\|Q)$ as in Sanov's theorem satisfy an orthogonality property similar to Euclidean projections. That is, if we write $\tilde{P}$ for the usual orthogonal projection of Q onto the set E,

$$|\tilde{P}Q| = \inf_{P\in E} |PQ|,$$

where $|PQ| = [\sum_x (P(x) - Q(x))^2]^{1/2}$ denotes the Euclidean norm, then the Pythagorean theorem says that, for any other $P \in E$,

$$|PQ|^2 \geq |P\tilde{P}|^2 + |\tilde{P}Q|^2,$$

see Figure 2.

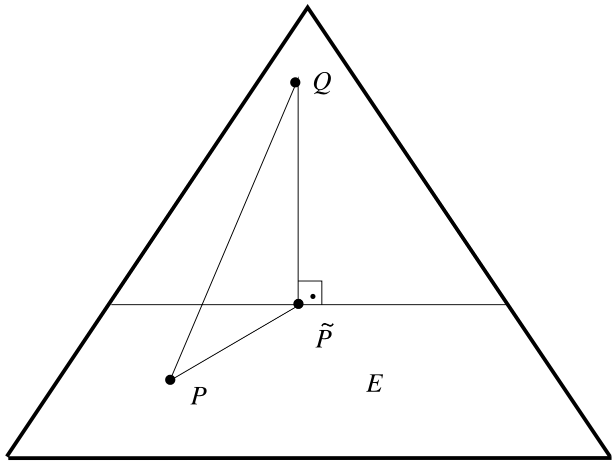

Figure 2: Graphical representation of the Pythagorean theorem for Euclidean projections on the simplex $\mathcal{P}$.

Theorem. [Pythagorean Identity] Let Q be a probability distribution on a finite alphabet $A = \{a_1, a_2, \ldots, a_m\}$, and let E be a (nonempty) set of probability distributions on A, not containing Q. If E is convex and closed, then there is a unique $P^* \in E$ achieving the infimum $\inf_{P \in E} D(P\|Q)$, and it satisfies,

$$D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q), \quad \text{for all } P \in E.$$

The proof is based on a local argument, similar in spirit to arguments commonly used in differential geometry. Indeed, the connection can be made deeper and more precise. Using the local structure of the relative entropy it is possible to define a Riemannian metric on the simplex of probability distributions (and more generally, on the space of probability measures on general spaces), leading to a very rich and interesting "information geometry;" see, e.g., the text [1].

Sanov's theorem and the Pythagorean identity clearly identify the distributions P^* achieving the projections $\inf_{P \in E} D(P\|Q)$ as important in this context. It is then natural to ask if they can be expressed in a particularly easy form. Indeed, for the special case when E is the intersection of $\mathcal{P}$ with a half-space in $\mathbb{R}^m$, their form can be evaluated explicitly.

For example, suppose that for some function $V : A \to \mathbb{R}$ and a constant c we let, $E = \{P \in \mathcal{P} \ : \ E_P[V(X)] \geq c\}$. We can try to evaluate the infimum $\inf_{P \in E} D(P\|Q)$ using the classical method of Lagrange multipliers. We define the functional,

$$J(P) = D(P\|Q) + \beta \sum_x P(x)V(x) + \alpha \sum_x P(x),$$

and differentiating $J(P)$ with respect to any one of the components of P and setting equal to zero,

$$\frac{\partial J(P)}{\partial P(x)} = \log \frac{P(x)}{Q(x)} + P(x)\frac{Q(x)}{P(x)}\frac{1}{Q(x)}(\log e) + \beta V(x) + \alpha = 0,$$

which gives,

$$P(x) = (\text{const})2^{\beta V(x)}.$$

The next proposition states that this is precisely the form of the projections P^*. Although it is not hard to make the above computation rigorous, we present an alternative proof below.

Proposition. [MAXIMUM ENTROPY AND GIBBS' DISTRIBUTIONS] Let $V : A \to \mathbb{R}$ be an arbitrary function on a finite alphabet $A = \{a_1, a_2, \ldots, a_m\}$.

(a) If $\frac{1}{m}\sum_x V(x) < c < \max_x V(x)$, then the maximum,

$$\sup\{H(P) \;:\; E_P[V(X)] \geq c\},$$

is achieved by the distribution,

$$P^*(x) = \frac{2^{\beta V(x)}}{\sum_y 2^{\beta V(y)}},$$

where $\beta > 0$ is chosen such that the constraint is satisfied with equality, $E_{P^*}[V(X)] = c$.

(b) If $\sum_x Q(x)V(x) < c < \max_x V(x)$, for some distribution Q on A, then the minimum,

$$\inf\{D(P\|Q) \;:\; E_P[V(X)] \geq c\},$$

is achieved by the distribution,

$$P^*(x) = \frac{Q(x)2^{\beta V(x)}}{\sum_y Q(y)2^{\beta V(y)}},$$

where $\beta > 0$ is chosen such that the constraint is satisfied with equality, $E_{P^*}[V(X)] = c$.

We remark that outside the range of the values of c considered above the problems are trivial. For example, if $c \leq \frac{1}{m}\sum_x V(x)$ in (a), then we can trivially take P to be the uniform distribution, and $H(P)$ takes its (unconstrained) maximum value of $\log m$. The other cases are similar.

Next we state a refinement to Sanov's theorem, stating that, conditional on a "rare event" of the form $\{\hat{P}_{X_1^n} \in E\}$, the distribution of the individual random variables X_i is close to P^*.

Theorem. [CONDITIONAL LIMIT THEOREM] Let $\hat{P}_n = \hat{P}_{X_1^n}$ denote the type of an i.i.d. random string $X_1^n = (X_1, X_2, \ldots, X_n)$ with distribution Q on A, and let let E be a (nonempty) convex set of probability distributions on A, not containing Q. If E is equal to the closure of its interior, then for all $a \in A$,

$$\Pr\{X_1 = x \mid \hat{P}_n \in E\} \to P^*(x), \quad \text{as } n \to \infty,$$

where P^* is the unique distribution in E that achieves the infimum $\inf_{P \in E} D(P\|Q)$.

References

[1] S-i. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000.