1 The Hölder exponent

Given a measure μ over a (pseudo-)metric space X, the *Hölder exponent* or *local dimension* $\alpha(x)$ of μ at a point $x \in X$ is given by:

$$\alpha(x) = \lim_{\epsilon \to 0} \frac{\ln(\mu(B_\epsilon(x)))}{\ln(\epsilon)}$$

where $B_\epsilon(x)$ is a ball of radius ϵ under the (pseudo-)metric.

Equivalently, we can say that the Hölder exponent $\alpha(x)$ is the unique real number such that

$$\mu(B_\epsilon(x)) \approx c\epsilon^{\alpha(x)} \tag{1}$$

for some constant c asymptotically as $\epsilon \to 0$.

2 Singular learning theory setup

We review some of the setup of singular learning theory. Let $W \subset R^n$ be our parameter space and X our input space. We have a parameterized statistical model defined by a probability distribution $p(x|w) : X \times W \to \mathbb{R}$. The space of probability distributions has a natural notion of distance given by the KL divergence:

$$\mathrm{KL}(p(x) \mid\mid q(x)) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

The *behavioral local learning coefficient (behavioral LLC)* at a parameter w^* is defined as the unique rational number $\lambda(w^*)$ such that

$$V(\epsilon) \approx c\epsilon^{\lambda(w^*)}(-\log(\epsilon))^{m(w^*)-1}$$

asymptotically as $\epsilon \to 0$ for some integer $m(w^*)$ and constant c, where

$$V(\epsilon) = \int_{\mathrm{KL}(p(x|w^*) \mid\mid p(x|w)) < \epsilon} dw.$$

When $m(w^*) = 1$, this becomes

$$V(\epsilon) \approx c\epsilon^{\lambda(w^*)} \tag{2}$$

Comparing Eq 1 and Eq 2 one may notice some similarity between $\lambda(w^*)$ and the Hölder exponent. This connection may be made rigorous.

We must first solve some technical issues. The KL divergence is not even a pseudometric - it is a divergence, and is not symmetric. However, it is closely related to the *Jensen-Shannon metric* d_{JS}, defined as:

$$d_{JS}(p, q) = \sqrt{\frac{1}{2}\mathrm{KL}(p \mid\mid m) + \frac{1}{2}\mathrm{KL}(q \mid\mid m)}$$

where $m = \frac{1}{2}(p+q)$ is a mixture distribution of q and m. The Jensen-Shannon metric symmetrizes the KL divergence by introducing the distribution m "halfway" between p and q, and the square root makes it a metric instead of a divergence.

Note that while the Jensen-Shannon metric is a metric on the space of probability distributions, if the parameterization map $w \mapsto p(x|w)$ is not injective (as typically the case in singular models), it will only be a *pseudometric* on the parameter space W, because distinct points may have distance zero if they map to the same probability distribution. This is why it was necessary to allow pseudometrics in the definition of the Hölder exponent above.

Over large distances, the KL divergence and Jensen-Shannon metric may not have a simple relationship. However, we only need them to be related for small distances. For two sufficiently close distributions p and q, it holds that

$$d_{JS}(p,q)^2 = \frac{1}{8}\,\mathrm{KL}(p \mid\mid q) + O(d_{JS}(p,q)^3). \tag{3}$$

Now we are ready to introduce our main claim.

Proposition 1. Assume the setup of the prior section, including the definition of the behavioral LLC $\lambda(w^*)$ and behavioral local multiplicity $m(w^*)$ for a point $w^* \in W$. Let $\alpha(w^*)$ be the Hölder exponent for the point $w^* \in W$ under the uniform measure on W and the Jensen-Shannon metric d_{JS}. Then, if $m(w^*) = 1$:

$$\alpha(w^*) = 2\lambda(w^*)$$

Proof. We leverage Eq 3 to relate the volume close to $p(x|w^*)$ under the KL divergence with the volume close to $p(x|w^*)$ under the Jensen-Shannon metric:

$$\begin{aligned}
\mu(B_\epsilon) &= \int_{d_{JS}(p(x|w^*),\,p(x|w)) < \epsilon} dw \\
&= \int_{d_{JS}(p(x|w^*),\,p(x|w))^2 < \epsilon^2} dw \\
&= \int_{\mathrm{KL}(p(x|w^*) \,||\, p(x|w)) < 8\epsilon^2 + O(\epsilon^3)} dw \\
&= V(8\epsilon^2 + O(\epsilon^3))
\end{aligned}$$

Then

$$\begin{aligned}
\mu(B_\epsilon) &= V(8\epsilon^2 + O(\epsilon^3)) \\
&\approx c(8\epsilon^2 + O(\epsilon^3))^{\lambda(w^*)} \\
&\approx c\epsilon^{2\lambda(w^*)} + O(\epsilon^{2\lambda(w^*)+1})
\end{aligned}$$

where the last step leverages the Taylor approximation of binomial expansion. Asymptotically for small ϵ, the $O(\epsilon^{2\lambda(w^*)+1})$ term is dominated, and we have:

$$\mu(B_\epsilon) \approx c\epsilon^{2\lambda(w^*)}$$

By the definition of the Hölder exponent $\alpha(w^*)$ we can conclude that $\alpha(w^*) = 2\lambda(w^*)$. □