# Local KL Geometry

The local KL geometry of statistical manifolds deviates from euclidean geometry in two ways: its quadratic part is curved and it contains data beyond its quadratic part. We examine how the new geometry distorts our euclidean picture of a large-$N$ sample as a tight Gaussian on an inner product space. We'll focus on the open simplex $\mathcal{M}$ with $D$ vertices.

The divergence $\mathrm{KL} : \mathcal{M} \times \mathcal{M} \to [0, \infty]$ given by

$$\mathrm{KL}(q : p) = \mathbb{E}_{x \sim q}[\log(q_x/p_x)]$$

is smooth and vanishes on and only on the diagonal.

Let's first examine the chance $\mathbb{P}[\hat{p}; p]$ of $N$ i.i.d. samples from $p$ yielding the empirical distribution $\hat{p}$. Fixing $D$ and allowing proportionality constants to depend on $N$, we consider $N$ large with respect to $\hat{p}$; that is, we take the sample granularity $\alpha = \max_x 1/(\hat{p}_x N)$ toward $0$. Plugging Stirling's formula (here $u$ is uniform)

$$\binom{N}{\hat{p}N} \propto \exp(NH(\hat{p}) - (D/2)\mathrm{KL}(u : \hat{p}) + O(\alpha))$$

into the lovely and routine formula (think of relative entropy)

$$\mathbb{P}[\hat{p}; p] \propto \binom{N}{\hat{p}N} \exp(-N\mathrm{KL}(\hat{p} : p) - NH(\hat{p}))$$

we see

$$\mathbb{P}[\hat{p}; p] \propto \exp(-N\mathrm{KL}(\hat{p} : p) - (D/2)\mathrm{KL}(u : \hat{p}) + O(\alpha))$$

Now let's rescale per the central limit theorem. With $\mathrm{KL}(p + v : p) = H(v, v)/2 + J(v, v, v)/6 + o(v^3)$ and $\mathrm{KL}(u : p + v) = \mathrm{KL}(u : p) + \tilde{G}(v) + o(v^1)$ we have for $v = u/\sqrt{N}$ and $u$ bounded (so that $\alpha \in O(1/N)$) that

$$\begin{aligned}\mathbb{P}[p + u/\sqrt{N}; p] \propto \exp(&-NH(u, u)/2 - NJ(u, u, u)/6\sqrt{N} \\ &-D\tilde{G}(u)/2\sqrt{N} + O(1/N))\end{aligned}$$

The $\tilde{G}$ term translates the Gaussian toward $u$; the $J$ term skews it (to have heavier tails) toward close vertices. For a sense of scale, note that $\hat{p}$'s true distribution has mean $p$, but this approximation suggests a shift on the order of $u \rightsquigarrow u + 1/\sqrt{N}$ or $v \rightsquigarrow v + 1/N$. This is comparable to the resolution of $\hat{p}$'s support and thus not a contradiction.

Let's visualize this for $D = 3, N = 30, p = (2/3, 3/12, 1/12)$.

## Concentration in Graphs

**Markov.** Suppose we have a width-$W$ depth-$D$ 'dense network' where each of $D$ layers has $W$ bit-valued nodes and where adjacent nodes have $\leq \delta$ mutual information. How does the total number of 'on' bits tend to deviate from the mean? (We might assume $\mu = WD/2$ and call the deviation $\Delta$).

First, if $\delta = 0$, then we have $D$ independent subgaussians with parameter at most $W$ so we expect concentration like $-\Delta^2/2W^2D$. A typical value for $\Delta$ is at most $W\sqrt{D}$. At another extreme, if $\delta = 1$, then each layer can communicate its majority to the next, so we get no concentration.

Intuitively, if $\delta$ mutual information corresponds to a probability $p$ of corruption, then concentration should be as good as if we had $pD$ many layers. Hmm... what's a lower bound for $p$ in terms of $\delta \leq p\log(1/p)+p$? Well, $\log(1/p) \leq \sqrt{1/p}$, so $\delta \leq 2\sqrt{p}$.

$$p\log(1/p) \approx \delta \qquad \delta/\log(1/\delta) \approx p \leq \delta \qquad p \approx (\delta/\log(1/\delta))/\log(\log(1/\delta)/\delta)$$