Note on Random Projection Estimators

acm

The problem (as I understand it): we have some density on a high dimensional random variable, $\pi(x)$ where $x \in \mathbb{R}^D$ (and D is big). Our goal is to take expectations of the form

$$\mathbb{E}_{x\sim\pi}\left[f(x)\right] = \int \pi(x) f(x) dx \tag{1}$$

for some class of functions $f(x)$.

We would like to develop an estimator that is a function of a lower dimensional random variable that is the result of a projection

$$z \triangleq Ax \tag{2}$$

$$\int p(z) f(z) d\mu(z) \approx \int \pi(x) f(x) dx \tag{3}$$

where $A \in \mathbb{R}^{d\times D}$ is some projection that projects x onto a d-dimensional subspace of $\mathbb{R}^D$.

> Question: what is the class of f? In the RHS above, does f map $\mathbb{R}^D \mapsto \mathbb{R}$? If so, then $f(z)$ maps $\mathbb{R}^d \mapsto \mathbb{R}$. What is an example of a function for which we have both $f(x)$ and $f(z)$?
> One approach could be define $f(z) = f([Ax; u])$ where u comes from some ambient noise distribution (or is fixed to some value).

Orthonormal projections

One way we can define a (potentially) tractable transformation is to consider random orthonormal projections. Sample an orthonormal A that sends x into some subspace $S = \{z : Ax = z, x \in \mathbb{R}^D\}$. Then choose A^c to be any orthonormal projection into the complement of S, $\neg S$. Now we have a one-to-one mapping, $y = \tilde{A}x$ where

$$y = [\underbrace{Ax}_{\text{dim } d \text{ proj.}}, \underbrace{A^c x}_{\text{dim } D-d \text{ proj}}] \tag{4}$$

$$\tilde{A} \triangleq [A; A^c] \tag{5}$$

so we can divide the vector y into two parts

$$y = [y_{1:d}; y_{d+1:D}] \quad \text{full one-to-one transformation} \tag{6}$$

$$y_{1:d} = Ax \quad \text{random projection component} \tag{7}$$

$$y_{d+1:D} = A^c x \quad \text{random complement component} \tag{8}$$

Together, the distribution $\pi(x)$ and the transformation A induce a distribution over $y_{1:d} = Ax$. Ignoring the function f for now, we can keep track of the probability measure on z by integrating over the last $D - d$ dimensions of y. If $\tilde{A}$ is orthonormal, then it is essentially just a rotation that will preserve measure (i.e. the determinant is 1).

$$
\begin{aligned}
p(y)dy &= \pi(x)dx && (9)\\
\implies p(y) &= \pi(x)\left|\frac{dy}{dx}\right|^{-1} && (10)\\
p(y_{1:d}) &= \int p(y_{1:d}, y_{d+1:D})dy_{d+1:D} && (11)\\
&= \int \pi(x)\left|\frac{dy}{dx}\right|^{-1} dy_{d+1:D} && (12)\\
&= \int \pi(x)dy_{d+1:D} && (13)\\
&= \int_{\neg S} \pi(x)dx && (14)
\end{aligned}
$$

The intuition behind the above is that the marginalizing out $p(y_{d+1:D})$ corresponds to integrating over $\mathbb{R}^D$ restricting to the complement of the original subspace defined by A, $\neg S$.

Now if we consider estimators of the form

$$
\begin{aligned}
\mathbb{E}_{y_{1:d}\sim p(y_{1:d})}\left[f(y_{1:d})\right] &= \int p(y_{1:d})f(y_{1:d})dy_{1:d} && (15)\\
&= \int_S \left(\int_{\neg S} \pi(x)dx\right) f(y_{1:d})dy_{1:d} && (16)\\
&= \int \pi(x)f(y_{1:d})dx && (17)
\end{aligned}
$$

which is the correct value as long as $f(y_{1:d})$ is a reasonable surrogate for $f(x)$.

I suspect all of the above can simply be restated as "marginal distributions of a well-defined joint must be coherent" — I think the interesting part will be finding a class of functions f where the above trick could work.

Other thoughts + q's

- Efficiency: can we learn a distribution over projections A that are more efficient (e.g. lower variance) than other distributions?
- Variational Inference: can we do something useful with only point-wise access to $\tilde{\pi}(x)$, an unnormalized version of π? If our goal is to learn an approximation for some posterior, we won't necessarily have samples from $x \sim \pi$ to manipulate.

- Another conceptual hangup I had when I was thinking about this problem before is keeping straight the difference between restriction onto a subset (i.e. conditioning) and projection of probability mass onto a subset (i.e. marginalization). When we have a sample $x \sim \pi$ and we project it into a subspace, we're doing a sort of "Monte Carlo Marginalization". However, if we have an unnormalized posterior $\tilde{\pi}(x)$ and we think about this function on the subsets $\{x : Ax + b = y, x \in mathbbR^D\}$ (e.g. rays, planes, or linear subspaces in $\mathbb{R}^D$) then we're conditioning, and it's harder to bridge to JL. I got the impression that one of the Ermon papers was doing this sort of restricting to a subspace.