Some questions

For some distribution like $p(x, y) = \frac{1}{Z}\sin(x + y)e^{-x^2-y^2}$, we may want to find good approximations $q(x, y)$ that are more convenient to work with.

$$\min_q \mathrm{KL}(q\|p) = ?$$

$$\min_q \mathrm{KL}(p\|q) = ?$$

For where do you pick the $q(x, y)$?
Does the choice of $KL(p\|q)$ or $KL(q\|p)$ matter ?

Let $\mathcal{G}$ be family of distributions and let $\mathcal{F} \subset \mathcal{G}$. For a given $p \in \mathcal{G}$, there might not be a $q \in \mathcal{F}$ with $KL(q\|p) = 0$.

$$\arg\min_{q \in \mathcal{G}} KL(q\|p) = \arg\min_{q \in \mathcal{G}} KL(p\|q) = p$$

$$\arg\min_{q \in \mathcal{F}} KL(q\|p) = ?$$

If we have a parametrized q, how do we find the right values for the parameters ?

Variational methods and exponential family

The exclusive divergence, $\mathrm{KL}(q_1(x)q_2(y)\|p(x,y))$, is equal to

$$
\begin{aligned}
&= \int_X \int_Y q_1(x)q_2(y) \log \left[\frac{q_1(x)q_2(y)}{p(x,y)} \right] dxdy \\
&= \int_X q_1(x) \log q_1(x) dx - \int_X q_1(x) \left(\int_Y q_2(y) \log p(x,y) dy \right) dx \\
&\qquad + \text{some constant not involving } q_1(x) \\
&= KL\left(q_1(x) \| \mathbb{E}_{q_2(y)} \left[\log p(x,y) \right] \right) + \text{const}
\end{aligned}
$$

For the inclusive divergence,

$$\begin{aligned}\mathrm{KL}(p(x,y)\|q_1(x)q_2(y)) &= \int_X \int_Y p(x,y)\log\left[\frac{p(x,y)}{q_1(x)q_2(y)}\right] dxdy \\ &= \mathrm{KL}(\int_Y p(x,y)dy\|q_1(x)) + \mathrm{const}\end{aligned}$$

even nicer since it doesn't even involve $q_2(y)$.
This mean that if we look for a fully-factorized distribution q in every variable involved in p, we simply model the marginals $\int_{\mathbf{x}\setminus\mathbf{x}_i} p(\mathbf{x}d\mathbf{x})$.
We saw last time that we could play with conjugacy by taking our approximating family $\mathcal{F}$ to be a subset of the exponential family when minimizing $KL(q\|p)$.

Why bother with the exponential family ?

Given a density function $p(x)$, let $\mathcal{F}$ be the set of all density functions that can be written as

$$q(x) = exp(\sum_j \nu_j g_j(x))$$

where the ν_j are the parameters and the $g_j(x)$ are the fixed features (i.e. $g_j(x) = x^j$ in the gaussian case, $j = 0, 1, 2$). Then

$$q = \arg\min_{\mathcal{F}} \mathrm{KL}(p\|q) \iff \forall j, \int g_j(x)q(x)dx = \int g_j(x)p(x)dx.$$

Basically, finding the right density function in $\mathcal{F}$ amounts to matching "moments".

In some case, those integrals can be solved analytically. If we restrict ourselves to the gaussian case $q(x) = exp(\nu_0 + \nu_1 x + \nu_2 x^2)$ and we have a way to compute the moments $\int x^j p(x) dx$ numerically, we can obtain the right values for the ν_j with a bit of algebra.

In some case, those integrals can be solved analytically. If we restrict ourselves to the gaussian case $q(x) = exp(\nu_0 + \nu_1 x + \nu_2 x^2)$ and we have a way to compute the moments $\int x^j p(x) dx$ numerically, we can obtain the right values for the ν_j with a bit of algebra.
This would not be true if we took $q(x)$ to be of the form

$$\begin{aligned} & q(x) = exp(g_0 + g_1 \log(x^2) + g_2 \cos(x)) \\ \text{or} \quad & q(x) = exp(g_0 + g_1 |x|^3 + g_2 x^4) \end{aligned}$$

or if the domain of x was restricted to $[-1, 1]$, for example.

Theorem (fixed point)

Let $\mathcal{F}$ be indexed by a continuous parameter θ, possibly with constraints. If $\alpha \neq 0$, the following are equivalent.

- *q is a stationary point of $D_\alpha(p\|q)$*
- *q is a stationary point of $proj_{\mathcal{F}}\left(p(x)^\alpha q(x)^{1-\alpha}\right)$*

Why bother with the exponential family ?

Proposition (Maximal entropy principle)

Of all the possible probability density functions $p(x)$ satisfying a given finite set of constraints

$$\int f_k(x)p(x)dx = F_k \in \mathbb{R},$$

the one having the maximal entropy is of the form

$$P(x) = \frac{1}{Z} exp\left(\sum_k w_k f_k(x)\right).$$

$$KL(p\|q) = \int p(x) log \left(\frac{p(x)}{q(x)} \right) \tag{1}$$