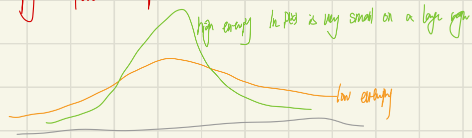
· Entropy of Exponential Family and Graphical Model

$$H_{P_\theta}(X) = -E_{x\sim P_\theta}[\ln P_\theta(x)]$$

$$= -E_{x\sim P_\theta}\left[\ln \frac{1}{Z(\theta)} \cdot \exp\langle t(\theta), T(x)\rangle\right]$$

$$= \ln Z(\theta) - \langle t(\theta), E_{x\sim P_\theta}(T(x))\rangle$$

low entropy: few low-prob instances



e.g. entropy of Gaussian $X \sim N(\mu, \sigma^2)$

$$t(\theta)\cdot T(x) = -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} = -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}$$

$$t(\theta) = \begin{bmatrix} -1/2\sigma^2 \\ \mu/\sigma^2 \\ -\mu^2/2\sigma^2 \end{bmatrix} \quad T(x) = \begin{bmatrix} x^2 \\ x \\ 1 \end{bmatrix}$$

$$H_{P_\theta}(X) = \ln\sqrt{2\pi}\,\sigma - \left(-\frac{1}{2\sigma^2}E[x^2] + \frac{\mu}{\sigma^2}E[x] - \frac{\mu^2}{2\sigma^2}\right) \qquad E[X^2] = \mu^2 + \sigma^2$$

$$= \frac{1}{2}\ln 2\pi\sigma^2 + \frac{(\mu^2+\sigma^2)}{2\sigma^2} - \frac{2\mu^2}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}$$

$$= \frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2} = \ln 2\pi e\,\sigma^2$$

if $p(x) = \frac{1}{Z}\prod_k \phi_k(D_k)$ is a markov network, then $H_p(X) = \ln Z + \sum_k E_p[\ln \phi_k(D_k)]$

$$T(x) = \ln \prod_k \phi_k(D_k)$$

if $P(X) = \prod_i P(X_i \mid Pa(X_i))$ is a distribution over a Bayesian network G

then $H_p(X) = \sum_i H_p(X_i \mid Pa(X_i))$

$$H_p(X) = E_p[-\ln p(x)] = E_p\left[-\sum_i \ln p(x_i \mid Pa(x_i))\right] = \sum_i E_p[-\ln p(x_i \mid Pa(x_i))] = \sum_i H_p(X_i \mid Pa(X_i))$$

· Relative Entropy

$$D(P\|Q) = E_{X\sim P}\left[\ln \frac{P(X)}{Q(X)}\right]$$

if Q is a distribution and P_θ is an exponential family defined by $T(\cdot)$ and $t(\cdot)$

$$D(Q\|P_\theta) = E_{x\sim Q}\left[\ln \frac{Q(X)}{P_\theta(X)}\right] = E_{x\sim Q}[\ln Q(x)] + E_{x\sim Q}\left[-\ln \frac{1}{Z(\theta)}\cdot \exp\langle t(\theta)\cdot T(x)\rangle\right]$$

$$= -H_Q(X) + \ln Z(\theta) - \langle t(\theta), E_{x\sim Q}[T(x)]\rangle$$

if P_{θ_1} and P_{θ_2} are in the same exponential family

$$D(P_{\theta_1}\|P_{\theta_2}) = E_{x\sim P_{\theta_1}}\left[\ln \frac{P_{\theta_1}(X)}{P_{\theta_2}(X)}\right]$$

$$= E_{x\sim P_{\theta_1}}\left[\ln \frac{1}{Z(\theta_1)}\exp\langle t(\theta_1), T(x)\rangle\right] - E_{x\sim P_{\theta_1}}\left[\ln \frac{1}{Z(\theta_2)}\cdot \exp\langle t(\theta_2), T(x)\rangle\right]$$

$$= -\ln \frac{Z(\theta_1)}{Z(\theta_2)} + E_{x\sim P_{\theta_1}}\left[\langle t(\theta_1), T(x)\rangle - \langle t(\theta_2), T(x)\rangle\right]$$

$$= -\ln \frac{Z(\theta_1)}{Z(\theta_2)} + \langle E_{x\sim P_{\theta_1}}[T(x)], t(\theta_1) - t(\theta_2)\rangle$$

if P is a distribution over a Bayesian Network G,

$$D(Q\|P) = E_{x\sim Q}[\ln Q] - E_{x\sim Q}[\ln P]$$
$$= -H_Q(X) - \sum_i E_{(x_i, pa(x_i))\sim Q}[\ln P(x_i \mid pa(x_i))]$$

$$E_{(x_i, pa(x_i))\sim Q}[\ln p(x_i \mid pa(x_i))] = \int_{x_i}\int_{pa(x_i)} Q(x_i, pa(x_i)) \cdot \ln p(x_i \mid pa(x_i))\, dx_i\, dpa(x_i)$$
$$= \int_{x_i}\int_{pa(x_i)} Q(pa(x_i)) \cdot Q(x_i \mid pa(x_i)) \cdot \ln p(x_i \mid pa(x_i))\, dx_i\, dpa(x_i)$$
$$= \int_{pa(x_i)} Q(pa(x_i))\, E_{x_i \sim Q(x_i \mid pa(x_i))}[\ln p(x_i \mid pa(x_i))]$$
$$= -H_Q(X) - \sum_i \sum_{pa(x_i)\in val(pa(x_i))} Q(pa(x_i))\, E_{x_i\sim Q(x_i \mid pa(x_i))}[\ln P(X_i \mid pa(x_i))]$$

$$\int\int_{x_1\, x_2} p(x_1,x_2)\ln f_1(x_1) f_2(x_1,x_2)\, dx_1 dx_2$$
$$= \int\int_{x_1\, x_2} p(x_1,x_2)\ln f_1(x_1)\, dx_1 dx_2 + \int\int p(x_1,x_2)\ln f_2(x_1,x_2)\, dx_1 dx_2$$
$$= \int_{x_1} p(x_1)\ln f_1(x_1)\, dx_1 + \cdots$$
$$= E_{x_1\sim p}[\ln f_1(x_1)] + E_{(x_1,x_2)\sim p}[\ln f_2(x_1,x_2)]$$

if P Q are 2 distributions over G.

$$D(Q\|P) = E_{x\sim Q}[\ln Q] - E_{x\sim Q}[\ln P]$$
$$= \sum_i \sum_{pa(x_i)} Q(pa(x_i))\, E_{x_i\sim Q(x_i \mid pa(x_i))}[\ln Q(X_i \mid pa(x_i))] - \sum_i \sum_{pa(x_i)} Q(pa(x_i))\, E_{x_i\sim Q(x_i \mid pa(x_i))}[\ln P(X_i \mid pa(x_i))]$$
$$= \sum_i \sum_{pa(x_i)} Q(pa(x_i))\, D\big(Q(X_i \mid pa(x_i)) \,\|\, P(X_i \mid pa(x_i))\big)$$

· Projection

View relative entropy as a notion of distance between two distributions

let P be a distribution and let Q be a convex set of distribution

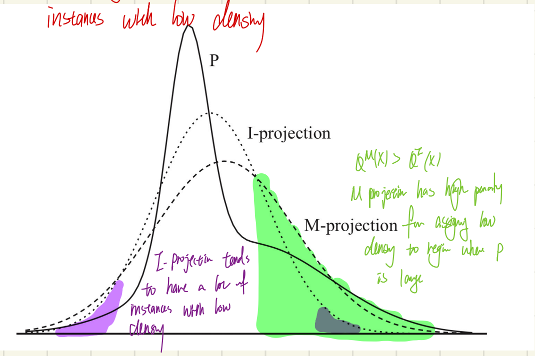
I-projection (information projection) of P onto Q: $Q^I = \arg\min_Q D(Q\|P)$

M-projection (moment projection) of P onto Q: $Q^M = \arg\min_Q D(P\|Q)$

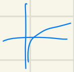
(: if $P \in Q$, then $Q^I = Q^M = P$)

because the relative entropy is not symmetric, $Q^M \neq Q^I$ in general

M-projection tends to give all assignments high density; I-projection tend to have some instances with low density



eg. project P onto Gaussians



M: $D(P\|Q) = \int_x p(x)\ln\frac{p(x)}{Q(x)}\,dx$

$$= -H_p(X) - E_{x\sim p}[\ln Q(x)]$$

$Q^M = \arg\min_Q -H_p(X) - E_{x\sim p}[\ln Q(x)] = \arg\max_Q E_{x\sim p}[\ln Q(x)]$

large penalty for assigning low density to regions where P is large.

I: $D(Q\|P) = \int_x Q(x)\ln\frac{Q(x)}{p(x)}\,dx$

$$= -H_Q(X) - E_{x\sim Q}[\ln p(x)]$$

$Q^I = \arg\min_Q -H_Q(X) + E_{x\sim Q}[\ln p(x)] = \arg\max_Q \underbrace{H_Q(X)}_{\text{small variance}} + \underbrace{E_{x\sim Q}[\ln p(x)]}_{\text{assign high/low prob to regions where } P \text{ is large/small}}$

· M-projection

let $P(X_1 \dots X_n)$ be a distribution and $\mathcal{Q}$ be the family of distributions over empty graph $G_\emptyset$

then $Q^M = \arg\min_{Q \in G_\emptyset} D(P\|Q) = P(X_1) \cdots P(X_n)$

$$\begin{aligned} D(P\|Q) &= E_{X\sim P}[\ln P(X) - \ln Q(X)] \\ &= E_{X\sim P}[\ln P(X)] - \sum_i E_{X_i \sim P}[\ln Q(X_i)] \\ &= E_{X\sim P}\left[\ln \frac{P(X)}{P(X_1)\cdots P(X_n)}\right] + \sum_i E_{X_i\sim P}\left[\ln \frac{P(X_i)}{Q(X_i)}\right] \\ &= D(P\|Q^M) + \sum_i D(P(X_i)\|Q(X_i)) \\ &\geq D(P\|Q^M) \end{aligned}$$

$$\int_{R^n} P(X_1 \dots X_n) \ln Q_1(X_1) \cdots Q_n(X_n)\, dx_1 \cdots dx_n$$

$$= \sum_i \int_{R^n} P(X_1 \cdots X_n) \ln Q_i(X_i)\, dx_1 \cdots dx_n$$

$$= \sum_i \int_R P(X_i) \ln Q(X_i)\, dx_i$$

equality holds only when $P(X_i) = Q(X_i)$

the M-Projection onto factored distribution is simply the product of marginals

Let P be a distribution over $\mathcal{X}$. and $\mathcal{Q}$ be an exponential family defined by $t(\theta)$ and $\tau(X)$

if there is a parameter θ s.t. $E_{X\sim Q_\theta}[\tau(X)] = E_{X\sim P}[\tau(X)]$, then M-Projection of P is Q_θ

Suppose $E_{X\sim P}[\tau(X)] = E_{X\sim Q_\theta}[\tau(X)]$

$$\begin{aligned} D(P\|Q_{\theta'}) - D(P\|Q_\theta) &= E_{X\sim P}\left[\ln \frac{P(X)}{Q_{\theta'}(X)}\right] - E_{X\sim P}\left[\ln \frac{P(X)}{Q_\theta(X)}\right] \\ &= E_{X\sim P}[\ln Q_\theta(X)] - E_{X\sim P}[\ln Q_{\theta'}(X)] \\ &= E_{X\sim P}\left[\ln \frac{1}{Z(\theta)} \cdot \exp\langle t(\theta), \tau(X)\rangle\right] - E_{X\sim P}\left[\ln \frac{1}{Z(\theta')} \exp\langle t(\theta'), \tau(X)\rangle\right] \\ &= \ln \frac{Z(\theta')}{Z(\theta)} + \langle t(\theta), E_{X\sim P}[\tau(X)]\rangle - \langle t(\theta'), E_{X\sim P}[\tau(X)]\rangle \\ &= \ln\left(\frac{Z(\theta')}{Z(\theta)}\right) + \langle t(\theta) - t(\theta'), E_{X\sim P}[\tau(X)]\rangle \\ &= \ln\left(\frac{Z(\theta')}{Z(\theta)}\right) + \langle t(\theta) - t(\theta'), E_{X\sim Q_\theta}[\tau(X)]\rangle \\ &= D(Q_\theta \| Q_{\theta'}) \geq 0 \end{aligned}$$

$$= \int_x Q_\theta(x) \cdot \ln \frac{\frac{1}{Z(\theta)} \cdot \exp\langle t(\theta), \tau(X)\rangle}{\frac{1}{Z(\theta')} \exp\langle t(\theta'), \tau(X)\rangle}\, dx$$

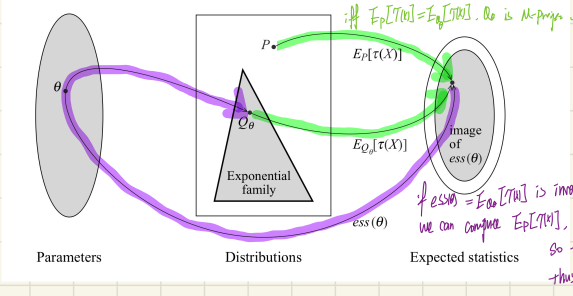
instead of describing a distribution in the family by parameter θ,
we can describe it in terms of the expected sufficient statistics.

define a mapping from parameter to expected sufficient statistics: $ess(\theta) = E_{Q_\theta}[\tau(X)]$

if $E_P[\tau(X)]$ is in the image of $ess(\cdot)$, then the M-projection of P is Q_θ that $E_{Q_\theta}[\tau(X)] = E_P[\tau(X)]$

if ess is invertible, parameters of M-projection of P are simply inverse of $ess(\cdot)$

iff $E_P[\tau(X)] = E_{Q_\theta}[\tau(X)]$. Q_θ is M-projection of P



if $ess(\theta) = E_{Q_\theta}[\tau(X)]$ is invertible.
we can compute $E_P[\tau(X)]$, then compute $\theta = ess^{-1}(E_P[\tau(X)])$
so that $E_{Q_\theta}[\tau(X)] = E_P[\tau(X)]$.
thus we found M-projection of P in $\{Q_\theta\}$

In many exponential families, the sufficient statistics $T(x)$ are moments. (eg. when $Q \sim N(\mu, \sigma^2)$ $t(\theta) = \begin{bmatrix} -1/2\sigma^2 \\ \mu/\sigma^2 \\ -\mu^2/2\sigma^2 \end{bmatrix}$ $T(x) = \begin{bmatrix} x^2 \\ x \\ 1 \end{bmatrix}$)
when $E_P[T(x)] = E_Q[T(x)]$. Q preserves the moments $E[X^n]$ of P