0.0.1 Natural exponential family

Likelihood:

$$\ell(\theta) = \pi(x) + \theta^{\mathsf{T}} x - G(\theta)$$

$G(\theta)$ is log partition function. Also generates moments:

$$\begin{aligned}
G(\theta) &= \log \int dx\, e^{\pi(x)+\theta^{\mathsf{T}} x} \\
G'(\theta) &= e^{-G(\theta)} \int dx\, e^{\pi(x)+\theta^{\mathsf{T}} x} x = E_\theta[x] \\
G''(\theta) &= e^{-G(\theta)} \int dx\, e^{\pi(x)+\theta^{\mathsf{T}} x} x^2 - G'(\theta)\, e^{-G(\theta)} \int dx\, e^{\pi(x)+\theta^{\mathsf{T}} x} x = E_\theta[x^2] - E_\theta[x]^2 = Var_\theta(x)
\end{aligned}$$

Since variances are positive semi-definite, G is a convex function.

0.0.2 Mean parameters and dual

The **conjugate** or **convex dual** to G is:

$$F(x) = \sup_\theta \left[\theta^{\mathsf{T}} x - G(\theta)\right]$$

That is, the greatest distance that a line of slope x starting from the origin rises above G. At that point, the derivative of the difference must be 0, so $G'(\theta^*) = x$. So equally, F gives intercept of the tangent to G with slope x.

Recall from above that $G'(\theta) = \mu$. So $F(\mu) = \theta^{\mathsf{T}}\mu - G(\theta)$.

$F(x)$ gives the maximum value of the likelihood (upto $\pi(x)$) for data with sufficient stat x.

Now G is generally strictly convex (otherwise variance of sufficient stat would be zero for some parameters). Thus, G' is strictly monotonic and there is a one-to-one map between θ and *feasible* values of μ. Thus, the exponential family can also be parametrised by μ.

Then $F(\mu)$ is the negative entropy of the distribution (upto $\pi(x)$):

$$-\mathbf{H}[x] = \langle \log p(x)\rangle = \left\langle \pi(x) + \theta^{\mathsf{T}} x - G(\theta)\right\rangle = \langle \pi(x)\rangle + \theta^{\mathsf{T}}\mu - G(\theta) = \langle \pi(x)\rangle_\mu + F(\mu)$$

We often write $g(\theta) = G'(\theta) = \mu$; also $f(\mu) = F'(\mu) = \theta$. So $f = g^{-1}$ and $f'(\mu) = 1/g'(\theta)$.

0.0.3 Bregman Divergences

The Bregman divergence under a differentiable, strictly convex function F is:

$$B_F(p|q) = F(p) - F(q) - f(q)(p-q)$$

that is, the difference between $F(p)$ and a first order approximation to $F(p)$ anchored at q. Strict convexity means that $B_F \geq 0$ with equality iff $p = q$.

ExpFam likelihood can be written:

$$\ell(\mu) = \pi(x) + F(x) - B_F(x|\mu)$$

Also:

$$B_G(\theta|\theta') = B_F(\mu\prime|\mu) = KL[p(x|\theta')|p(x|\theta)]$$

where last step follows from:

$$\begin{aligned}KL[p(x|\theta')|p(x|\theta)] &= \langle \log p(x|\theta') - \log p(x|\theta)\rangle_{\theta'} \\ &= \left\langle \pi(x) + (\theta')^\mathsf{T} x - G(\theta') - \pi(x) - \theta^\mathsf{T} x + G(\theta)\right\rangle_{\theta'} \\ &= G(\theta) - G(\theta') - (\theta - \theta')^\mathsf{T} \langle x\rangle_{\theta'} \\ &= G(\theta) - G(\theta') - (\theta - \theta')^\mathsf{T} \mu' \\ &= G(\theta) - G(\theta') - (\theta - \theta')^\mathsf{T} g(\theta')\end{aligned}$$

0.0.4 ML fitting

$$\begin{aligned}\ell(\theta) &= \sum_i \pi(x_i) + \theta^\mathsf{T} x_i - G(\theta) \\ \ell'(\theta) &= \sum_i x_i - G'(\theta) \\ \Rightarrow NG'(\theta^{ML}) &= \sum_i x_i \\ \Rightarrow \theta^{ML} &= f(\frac{1}{N}\sum_i x_i)\end{aligned}$$

0.0.5 GLMs

Consider scalar x_i and vector inputs $\mathbf{y}_i$.

$$\ell(\mathbf{w}) = \sum_i \pi(x_i) + x_i \mathbf{w}^\mathsf{T}\mathbf{y}_i - G(\mathbf{w}^\mathsf{T}\mathbf{y}_i)$$

so,

$$\ell'(\mathbf{w}) = \sum_i x_i\mathbf{y}_i - g(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\mathbf{y}_i = \sum_i (x_i - \mu_i)\mathbf{y}_i$$

and

$$\ell''(\mathbf{w}) = -\sum_i g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^\mathsf{T}$$

So a Newton update would be:

$$\begin{aligned}\Delta\mathbf{w} &= -(\ell''(\mathbf{w}))^{-1}\ell'(\mathbf{w}) \\ &= \left[\sum_i g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^\mathsf{T}\right]^{-1} \sum_i (x_i - \mu_i)\mathbf{y}_i \\ &= \left[\sum_i g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^\mathsf{T}\right]^{-1} \sum_i \underbrace{(x_i - \mu_i)f'(\mu_i)}_{\Delta z_i} g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\mathbf{y}_i \\ &= \left[\sum_i g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^\mathsf{T}\right]^{-1} \sum_i g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)\Delta z_i\mathbf{y}_i\end{aligned}$$

which looks like weighted linear regression with "inputs" $\mathbf{y}_i$, "outputs" $z_i = \mathbf{w}^\mathsf{T}\mathbf{y}_i + (x_i - \mu_i)f'(\mu_i)$ and "variances" $1/g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i) = f'(\mu_i)^2 g'(\mathbf{w}^\mathsf{T}\mathbf{y}_i)$ (i.e. variance in z_i estimated by linearisation around μ_i). This is IRLS.

TODO: Non-natural link functions and variance parameters.

0.0.6 Latent variables

Exponential Family PCA: $\theta_i = \mathbf{w}^\top \mathbf{y}_i$. Optimise jointly over $\mathbf{w}$ and $\mathbf{y}_i$.

NMF: $\mu_{ij} = \mathbf{w}_i^\top \mathbf{y}_j$. Optimise jointly over $\mathbf{w}_i$ and $\mathbf{y}_j$, both constrained to be non-negative.