1 Notation

X random variable
x instantiation/sample of the random variable X
$u(X_1, \ldots, X_N)$ sufficient statistic
θ parameterization of the distribution
η natural parameter
$Z(\eta)$ normalization factor
$g(\eta)$ inverse normalizer
$h(x)$ intrinsic/carrier measure

2 Pitman-Koopman-Dermois Theorem

Suppose we have the sufficient statistic is additive: $u(X_1, \ldots, X_N) = \sum_{n=1}^{N} u(X_1)$. Furthermore assume it is a minimal sufficient statistic, which means mathematically there exists a mapping from any other sufficient statistic to u.

$$f(u'(X_1, \ldots, X_N)) = u(X_1, \ldots, X_N)) \tag{1}$$

Then we can express parametrize the distribution by the natural parameters, η, such that

$$P(x|\eta) = g(\eta)h(x)e^{\langle u(x),\eta\rangle} \tag{2}$$

where

$$\begin{aligned} \tilde{P} &:= h(x)e^{\langle u(x),\eta\rangle} \\ Z(\eta) &:= \int h(x)e^{\langle u(x),\eta\rangle} \\ g(\eta) &:= \frac{1}{Z(\eta)} \end{aligned} \tag{3}$$

Problem 2.1. *Show that it satisfies the Fisher-Neyman factorization.*

Problem 2.2. *The German Tank Problem Find a probability distribution that has a sufficient statistic, but is not an exponential family distribution. Is it an additive sufficient statistic?*

Problem 2.3. *What is the difference between a normal distribution and an exponential family of distributions?*

Problem 2.4. *Bishop shows that the family of Bernoulli distributions is an exponential family (and Bernoulli is a special case of the binomial distribution). However, there is a mistake/problem with Bishop's definition, what is it??*

Problem 2.5. *Give an exponnential family that contains the binomial distribution (discrete distribution of n coin flips with the probability of heads being p).*

Problem 2.6. *Given a pdf, we $p(x|\theta)$, we can write $p(x) = e^{\log p(x|\theta)}$. This seems to suggest any pdf can be turned into an exponential family. But that is not quite correct. Give conditions when this trick will and will not work.*

Problem 2.7. *The Weibull distribution was used to model the distribution for the onset of covid cases.*
The Weibull distribution has the cumulative distribution function

$$P(X \leq x|\lambda, k) = \begin{cases} 1 - e^{-\frac{x}{\lambda}^k} & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{4}$$

Express the Weibull distribution as an exponential family, with a sufficient statistic and natural parameters.

Problem 2.8. *The Gumbel distribution, which is used for the Gumbel-softmax trick for backpropagating through categorical distributions, is defined as*

$$\begin{aligned} P(x|\mu, \beta) = \frac{1}{\beta} e^{-(z+e^{-z})}, \quad z := \frac{x-\mu}{\beta} \\ P(X \leq x|\mu, \beta) = e^{-e^{-\frac{x-\mu}{\beta}}} \end{aligned} \tag{5}$$

Is it an exponential family?

3 Sufficiency

Fisher-Neyman factorization

One explanation of a sufficient statistic is that it contains all relevant information about observations for inferring the parametrization. We wish to understand what this means.

Problem 3.1. *Suppose we know nothing about our data distribution, $P_{data}(X)$.*
What would be an appropriate sufficient statistic?
What do we need to specify beforehand before we can define a sufficient statistic?

Problem 3.2. *Suppose we have two sets of samples of size N, $\{x_1, \ldots, x_N\}, \{x'_1, \ldots, x'_N\}$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known but μ is unknown.*
Further suppose that we have a normal prior, $P(\mu) = \mathcal{N}(0, 1)$.
The sufficient statistic is the sum of the samples, so

$$\begin{aligned} u(x_1, \ldots, x_N) = \sum_n x_n \\ = u(x'_1, \ldots, x'_N) = \sum_n x'_n \end{aligned} \tag{6}$$

Is it true that we assume equal posterior probabilities?

$$P(\mu = 1|x_1, \ldots, x_N) = P(\mu = 2|x'_1, \ldots, x'_N) \tag{7}$$

Suppose the sufficient statistic is the same. How do the posteriors compare?

$$P(\mu = 1|x_1, \ldots, x_N, u(x_1, \ldots, x_N) = C) = P(\mu = 2|x'_1, \ldots, x'_N, u(x'_1, \ldots, x'_N) = C)? \tag{8}$$

Problem 3.3. *Suppose two sets of samples of size N, $\{x_1, \ldots, x_N\}, \{x'_1, \ldots, x'_N\}$, from a family of distributions parametrized by θ with sufficient statistic u.*

Suppose $u(x_1, \ldots, x_N) = u(x'_1, \ldots, x'_N)$. Can you find an example where $\prod_n P(x_n|\theta) \neq \prod_n P(x_n|\theta')$?

Problem 3.4. *Suppose now that we now have two gaussian distributions with the same variance, $\mathcal{N}(-1,1)$ and $\mathcal{N}(1,1)$. We pick with equal probability one of these distributions but without knowing which one, and then we can sample from it repeatedly.*
Now we repeatedly to sample this distribution N times, but we reject/discard and resample any samples with $x < 0$ (but we always end up with a total of N accepted samples).

The sufficient statistic is still the mean (you can try to justify this, though it might be hard).

Can you come up with an example of $\{x_1, \ldots, x_N\}, \{x'_1, \ldots, x'_N\}$ with the same sufficient statistic but different likelihoods?

We can conclude

$$P(x_1, \ldots, x_N|\theta, u(x_1, \ldots, x_N) = C) \neq P(x'_1, \ldots, x'_N|\theta, u(x'_1, \ldots, x'_N) = C) \quad (9)$$

So sufficiency does not mean that given all samples with the same sufficient statistic are equally likely.

It does not mean that a sample has the same likelihood regardless of the parametrization (see ancillary statistic).

It does mean that the likelihood ratio of two samples with the same sufficient statistic is the same no matter what the parametrization:

$$\frac{\prod_n P(x_n|\theta)}{\prod_n P(x'_n|\theta)} = \frac{\prod_n P(x_n|\theta')}{\prod_n P(x'_n|\theta')} \quad (10)$$

That is, $P(x_1, \ldots, x_N|u(x_1, \ldots, x_N) \perp \theta$.

It also means that

$$\frac{\prod_n P(x_n|\theta)}{\prod_n P(x_n|\theta')} = \frac{\prod_n P(x'_n|\theta)}{\prod_n P(x'_n|\theta')} \quad (11)$$

Problem 3.5. *Explain what this equation means in terms of inferring the parameters given the sufficient statistic.*

Problem 3.6. *Assuming the above, prove the Fisher-Neyman factorization theorem.*

$$p(x|\theta) = h(x) f_\theta(u(x)) \quad (12)$$

where $h(x)$,f_θ are non-negative functions.

Problem 3.7. *Draw a (graphical) relationship between a sample, $\mathcal{D} = X_1, \ldots, X_N)$, the sufficient statistic $u(x_1, \ldots, x_N)$, and the parametrization, θ. Use a double forward arrow to indicate deterministic relationships.*

4 Gradient properties of exponential families

Problem 4.1. *Derive an expression for $\nabla_\eta \log Z(\eta)$ in terms of the sufficient statistics directly.*

Problem 4.2. *Derive an expression for $\nabla_\eta \log Z(\eta)$ from the fact that $\int P(x|\eta)dx = 1$.*

Problem 4.3. *Derive an expression for $\nabla_\eta \nabla_\eta \log Z(\eta)$ in terms of the sufficient statistics.*

Problem 4.4. *Derive an expression for $\nabla_\eta \log P(x|\eta)$ in terms of the expected sufficient statistic.*

Problem 4.5. *Suppose we have online samples of a bernoulli distribution with an unknown mean. Derive a first order and second order gradient-based update of the distribution parameters.*

Problem 4.6. *Suppose we have a online samples $\{(X_n, Y_n)\}_n$ for a logistic regression model*

$$P(Y|X, W) = \sigma(\langle W, X \rangle) \tag{13}$$

Derive a first order and second order gradient-based update for the parameters W.

Problem 4.7. *Prove that $\log Z(\eta)$ is a convex function. (Hint: use results from the above problem)*

Problem 4.8. *Prove that the domain of natural parameters is convex.*

Problem 4.9. *Assuming that the covariance of the sufficient statistic is positive definite (equivalently, that the variance constrained along any vector is positive), prove that there is a one-to-one relationship between η and the expected sufficient statistic.*

Problem 4.10. *Assuming the above, prove the canonical link function, $f(E_{x\sim P(x|\eta)}[u(x)]) = \eta$, exists and is one-to-one.*

Problem 4.11. *(Moment matching) Using gradients / derivatives, derive a condition for the M-projection:*

$$argmin_\eta KL\left(P_{data}(X) \| P(X|\eta)\right) = argmin_\eta \int P_{data}(X) \log \frac{P_{data}(X)}{P(X|\eta)} \tag{14}$$

Problem 4.12. *Derive the M-projection of a probability distribution $P(X)$ on to the space of Gaussian distributions.*
What if you only had samples of $P(X)$? How would you estimate the M-projection?

5 Probabilistic Graphical Models

Undirected Markov networks are defined by $\mathcal{M} = (\mathcal{H}, \Phi)$ are defined by their undirected graph $\mathcal{H}$ and their factors, $\Phi = \{\phi_i(D_i)\}_i$. The unnormalized probability distribution is called the Gibbs distribution:

$$\tilde{P}_\Phi(\mathcal{X}) = \prod_i \phi_i(D_i) \tag{15}$$

Problem 5.1. *If each of the factors is an exponential family distribution, write the joint distribution as an exponential family.*

6 Exponential families and convex optimization duality

The convex conjugate / Fenchel transform / Legendre-Fenchel transform is the function to a (potentially non-convex) function.

A notational note, we will use lowercase x here for random variables, instead of instantiations, since we are not considering samples in this section.

$$f^*(\lambda) = \sup_{x \in X} \left(\langle \lambda, x \rangle - f(x) \right) \tag{16}$$

Let $G(\eta)$ be the log normalizer:

$$G(\eta) = \log Z(\eta) = \int h(x) e^{\langle u(x), \eta \rangle} dx \tag{17}$$

Then we can re-expresss the exponential family formula as

$$P(x|\eta) = h(x) e^{\langle u(x), \eta \rangle - G(\eta)} \tag{18}$$

Let's consider the convex conjugate of the log normalizer:

$$F(\lambda) = \sup_{\eta} \left(\langle \lambda, \eta \rangle - G(\eta) \right) \tag{19}$$

Problem: Solve for $\nabla_\eta \left(\langle \lambda, \eta^* \rangle - G(\eta) \right) = 0$, specifically show that $\lambda = E_{x \sim P(x|\eta^*)} \left[u(x) \right]$. Is this a global maximum?

Using the above formula for λ, what is $F(\lambda)$?

What is $\nabla_\lambda F(\lambda)$? This gives us a formula for the canonical link function.

Fenchel's inequality is a direct result of the definition of the convex conjugate

$$f^*(\lambda) + f(x) \leq \langle \lambda, x \rangle \tag{20}$$

We observe that the exponential form of an exponential family resembles the term inside the convex conjugate:

$$P(x|\eta) = h(x) e^{\langle u(x), \eta \rangle - G(\eta)} \leq h(x) e^{F(\bar{u})} \tag{21}$$

And that it is maximized when $u(x)$ is the expected sufficient statistic. This means the mean sufficient statistic is related to its mode (disregarding $h(x)$ for now).

Suppose we were to define an exponential family for the natural parameter:

$$P(\eta|\lambda) = h^*(\eta) e^{\langle \lambda, \eta \rangle - F(\lambda)} \tag{22}$$

We can use Fenchel's inequality for $F(\lambda)$

$$P(\eta|\lambda) = h^*(\eta) e^{\langle \lambda, \eta - \eta^* \rangle + G(\eta)} \tag{23}$$

¡++¿

7 EM and variational inference

Free energy?