

# 1 Notation

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

# 2 Sufficiency

Fisher-Neyman factorization

One explanation of a sufficient statistic is that it contains all relevant information about observations for inferring the parametrization.

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?

Suppose we have two sets of samples of size  $N$ , from a normal distribution  $\mathcal{N}(\mu, \sigma^2 I)$ , where  $\sigma^2$  is known but  $\mu$  is unknown. Further suppose that we have a normal prior,  $P(\mu) = \mathcal{N}(0, I)$ . The sufficient statistic is the either the mean or the sum of the samples, so

$$\begin{aligned} u(x_1, \dots, x_N) &= \frac{1}{N} \sum_n x_n \\ &= u(x'_1, \dots, x'_N) = \frac{1}{N} \sum_n x'_n \end{aligned} \tag{1}$$

Is it true that we assume equal posterior probabilities?

$$P(\theta = 1|x_1, \dots, x_N) = P(\theta = 2|x'_1, \dots, x'_N) \tag{2}$$

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 equally probability one of these distributions, and then we can sample from it repeatedly. Now we decide to sample  $N$  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).

For 2 samples, suppose we have a sufficient statistic of 2, so  $x_1 + x_2 = 2$ . Compare

$$\begin{aligned} P(x_1 = -1, x_2 = 3|\mu) \\ P(x_1 = 1, x_2 = 1|\mu) \end{aligned} \tag{3}$$

We can conclude

$$P(x_1, \dots, x_N|\theta, u(x_1, \dots, x_N) = C) \neq P(x'_1, \dots, x'_N|\theta, u(x'_1, \dots, x'_N) = C) \tag{4}$$

Suppose two sets of samples (assume with the same size) have the same sufficient statistic  $u$ , given a family of distributions parametrized by  $\theta$ .

$$u(x_1, \dots, x_N) = u(x'_1, \dots, x'_N) \quad (5)$$

Can we have

$$\prod_n P(x_n|\theta) \neq \prod_n P(x_n|\theta') \quad (6)$$

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:

$$\frac{\prod_n P(x_n|\theta)}{\prod_n P(x'_n|\theta)} = \frac{\prod_n P(x_n|\theta')}{\prod_n P(x'_n|\theta')} \quad (7)$$

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

It also means that

$$\frac{\prod_n P(x_n|\theta)}{\prod_n P(x_n|\theta')} = \frac{\prod_n P(x'_n|\theta)}{\prod_n P(x'_n|\theta')} \quad (8)$$

Problem: explain what this equation means in terms of inferring the parameters given the sufficient statistic.

Problem: assuming the above, prove the Fisher-Neyman factorization theorem.

$$p(x|\theta) = h(x)f_\theta(u(x)) \quad (9)$$

where  $h(x), f_\theta$  are non-negative functions.

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

### 3 Pitman-Koopman-Dermois Theorem

### 4 Gradient properties of exponential families

Problem:

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

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

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

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

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

Prove that the domain of natural parameters is convex.

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  $\eta$  and the expected sufficient statistic.

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

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

$$\operatorname{argmin}_{\eta} KL(P_{data}(X) \| P(X|\eta)) = \operatorname{argmin}_{\eta} \int P_{data}(X) \log \frac{P_{data}(X)}{P(x|\eta)} \quad (10)$$

Derive a way of estimating an M-projection given samples from an exponential family.