1 Notation

X random variable
x instantiation of the random variable X
X_n sample of the random variable X
$u(X_1, \ldots, X_N)$ sufficient statistic
P the probability density function
θ parameterization of the distribution, generally denoted $P(x; \theta)$ in traditional statistics and $P(x|\theta)$ in Bayesian statistics
η natural parameter
$Z(\eta)$ normalization factor of the unnormalized distribution
$g(\eta)$ inverse normalizer
$h(x)$ intrinsic/carrier measure $\langle \cdot, \cdot \rangle$ inner product, assumed to be the dot product unless otherwise stated

2 Pitman-Koopman-Dermois Theorem

Suppose a sufficient statistic is additive: $u(X_1, \ldots, X_N) = \sum_{n=1}^{N} u(X_n)$. Furthermore, assume it is a minimal sufficient statistic, which means mathematically there exists a mapping f 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 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

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

Problem 2.1. *Find a probability distribution that has a sufficient statistic, but is not an exponential family distribution. Is it an additive sufficient statistic?*

Problem 2.2. *What is the difference between a normal distribution and an exponential family of distributions?*

The following are two more complicated distributions. Challenge yourself to see if you can convert them to exponential family form.

Problem 2.3. *The Weibull distribution was used to model the distribution for the onset of covid cases.*
The Weibull distribution has the cumulative distribution function

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

Express the Weibull distribution as an exponential family distribution (assume k is fixed), with a sufficient statistic and natural parameters.
Can we express it as an exponential family distribution it if k is considered to be a parameter?

Problem 2.4. *The Gumbel distribution, which is used for the Gumbel-softmax trick for backpropagating through categorical distributions, is defined as*

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

Is it an exponential family, and what assumptions do we have to make, if any? If so, find a minimal sufficient statistic and natural parameters.

The following are more conceptual questions.

Problem 2.5. *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 2.6. *Give an exponential family that contains the binomial distribution (discrete distribution of n coin flips with the probability of heads being p).*

Problem 2.7. *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 2.8. *Show that an exponential family distribution satisfies the Fisher-Neyman factorization theorem*

$$p(x|\theta) = h(x) f_\theta(u(x)) \tag{6}$$

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

3 Sufficiency

This is a more advanced, theoretical section (as in you may want to skip this for later), which is meant to solidify our understanding of the mysterious concept of sufficiency. An understanding of this is important for ideas about perfect and imperfect data reduction, such as Tishby's Information Bottleneck.

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

$$\begin{aligned} u(x_1, \ldots, x_N) &= \sum_{n=1}^{N} x_n \\ &= \sum_{n=1}^{N} x'_n = u(x'_1, \ldots, x'_N) \end{aligned} \tag{7}$$

Suppose the sufficient statistic is the same, $u(x_1, \ldots, x_N) = u(x'_1, \ldots, x'_N) = C$. How do the posteriors compare?

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

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 but different means, for example, $\mathcal{N}(-1, 1)$ 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 sample this distribution N times, but we procedurally reject/discard and resample any samples with $x < 0$, sot 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

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

So sufficiency does not mean that 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, that is, $P(x_1, \ldots, x_N|\theta) \neq P(x_1, \ldots, x_n|\theta')$ (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:

$$\frac{\prod_{n=1}^{N} P(x_n|\theta)}{\prod_{n=1}^{N} P(x'_n|\theta)} = \frac{\prod_{n=1}^{N} P(x_n|\theta')}{\prod_{n=1}^{N} P(x'_n|\theta')} \tag{10}$$

[1] A note on a change of notation, we will use lowercase x_n for samples instead of capital X_n. This is because we define our sufficient statistic through functions defined on the instantiations of the samples of the random variable.

That is, $X_1, \ldots, X_N \perp \theta | u(X_1, \ldots, X_N)$.
It also means that

$$\frac{\prod_{n=1}^{N} P(x_n|\theta)}{\prod_{n=1}^{N} P(x_n|\theta')} = \frac{\prod_{n=1}^{N} P(x'_n|\theta)}{\prod_{n=1}^{N} P(x'_n|\theta')} \tag{11}$$

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*

$$P(x|\theta) = h(x) f_\theta(u(x)) \tag{12}$$

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 Gradients of exponential families and statistical modeling

4.1 Gradients

The first three problems are very fundamental to exponential families and machine learning, and have some level of progression.

Problem 4.1. *Derive an expression for $\nabla_\eta \log Z(\eta)$ (Hint: in terms of the sufficient statistics or expectations of sufficient statistics).*
Derive an expression for $\nabla_\eta \log Z(\eta)$ from the fact that $\int P(x|\eta)dx = 1$.

Problem 4.2. *Derive an expression for $\nabla_\eta \nabla_\eta \log Z(\eta)$, the Hessian of the log normalizer, in terms of the expected sufficient statistics.*

Problem 4.3. *Derive an expression for $\nabla_\eta \log P(x|\eta)$ in terms of the expected sufficient statistic.*
Suppose we have samples $\{X_1, \ldots, X_N\}$ from a data distribution, $P_{data}(X)$. We can define an empirical distribution by the mean of N dirac delta functions, $P_{emp}(X) = \frac{1}{N} \sum_{n=1}^{N} \delta(X = X_n)$. [2].
Derive an expression for

$$\nabla_\eta KL\left(P_{emp}(X) \| P(X|\eta)\right) \tag{13}$$

, where

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

Hint: What is $\nabla_\eta \int P_{emp}(x) \log P_{emp}(x) dx$?
Hint: Assume gradients commute with integration and addition, since they are linear operators.

[2] The dirac delta function is a generalized function, which can be interpreted as a point mass: $\int f(x)\delta(x = x_n)dx = f(x_n)$

4.2 Discriminative Modeling

The next three problems are increasingly more complex versions of the same problem. Choose one.

Assume we are maximizing log likelihood, although we would like a more information theoretic justification.

A notational note, X may now be a random vector of K random variables, $X_1, \ldots, X_K$.
We can either denote samples of X, Y as $\{X_1^n, \ldots, X_K^n, Y^n\}_{n=1}^N$, which is often notationally simplified into a matrix form, $\{X_{n1}, \ldots, X_{nK}, Y_n\}_{n=1}^N$.

Problem 4.4. *Suppose we have online samples X^n of a Bernoulli distribution, $X \in \{0, 1\}$, with an unknown probability $P(X = 1) = p$. Derive an online first-order and Newton's method/second-order gradient-based update of the natural parameter of the Bernoulli distribution.*

Problem 4.5. *Suppose we have a online samples $X, Y \sim P_{data}(X, Y)$, with $X \in R^K$, $Y \in \{0, 1\}$. We model it with a logistic regression model, with weights $W = \{w_1, \ldots, w_k\}$.*
Show that our conditional Bernoulli distribution is equivalent to

$$P(Y = 1|X, W) = \sigma(\sum_k w_k X_k) \tag{15}$$

What is the relationship between this equation and the exponential family canonical link function relationship, $E_{Y \sim P(Y|\eta)}[u(Y)] = \nabla_\eta \log Z(\eta)$.
How would we model Y with a single Bernoulli distribution? What happens to X in that model?
What would we have to do in order to optimize each of the conditional distributions independently?
We tie these conditional distributions with a single weight vector, W. We could simply assume every data point has an equal gradient update contribution.
Derive a first order and second order gradient-based update for the parameters W.
We showed before that optimizing a log likelihood objective is equivalent to optimizing $KL(P_{emp}(X) \| P_{model}(X))$, for a non-conditional distribution. How can we interpret our conditional log likelihood objective as a KL-divergence?

Problem 4.6. *Generalized Linear Models (GLM)*
Derive a gradient update $\nabla_W \mathcal{L}(\mathcal{D})$, with a negative log likelihood loss function, $\mathcal{L}(\mathcal{D}) = \frac{\sum_{n=1}^N \log P_W(Y^n|X^n)}{N}$, with a discriminative model $Y|X, W \in Poisson(\lambda)$, where the natural parameter of the Poisson distribution is $\eta = \sum_k w_k X_k$.
How can we interpret our objective as a KL-divergence?

4.3 Convex Analysis

The next two problems are more theoretical and are optional if you're not overly concerned about convex analysis.

Problem 4.7. *Prove that $\log Z(\eta)$ is a convex function. (Hint: use results from the above gradients 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 η and the expected sufficient statistic.

Problem 4.8. *Assuming the above, prove the canonical link function, $f(E_{x\sim P(x|\eta)}[u(x)]) = \eta$, exists and is one-to-one.*

4.4 KL divergence, M-projections, and I-projections

$$KL\left(P(X)\|Q(X)\right) := \int P(x)\log\frac{P(x)}{Q(x)}dx \tag{16}$$

Problem 4.9. *What is $KL\left(\mathcal{N}\left(\mu_1,\Sigma_1\right)\|\mathcal{N}\left(\mu_2,\Sigma_2\right)\right)$?*
Derive a simplified expression for $KL\left(P(x|\eta')\|P(x|\eta)\right)$ in terms of the natural parameters and expected sufficient statistics.

Problem 4.10. *Moment matching / M-projections*
What is

$$\nabla_\eta KL\left(P(X|\eta')\|P(X|\eta)\right) \tag{17}$$

(Hint: do not use the final result of the above problem). Using gradients / derivatives or the previous problem, derive a condition for the M-projection:

$$argmin_\eta KL\left(P_{data}(X)\|P(X|\eta)\right) = argmin_\eta \int P_{data}(x)\log\frac{P_{data}(x)}{P(x|\eta)}dx \tag{18}$$

Problem 4.11. *Derive the M-projection of a probability distribution $P(X)$ on to the space of Gaussian distributions.*
What if you only had samples of a Bernoulli distribution $P(X)$? How would you estimate the M-projection?

The first part of this problem is about the "other" KL divergence, which is used in variational inference and EM, and when using Boltzmann distributions it ties together ideas in thermodynamics and maximum entropy.

Problem 4.12. *Information projections / I-projections and thermodynamics*
I-projections fix the right hand side of a KL divergence and vary the left-hand side.
Suppose are considering the I-projection of an exponential family distribution, over some class of distributions:

$$argmin_{Q(X)\in\mathcal{Q}} KL\left(Q(X)\|P(X|\eta)\right) \tag{19}$$

Assuming $P(x|\eta) = g(\eta)e^{\langle u(x),\eta\rangle}$, expand the expression for $KL\left(Q(X)\|P(X|\eta)\right)$ in terms of the entropy of $Q(X)$ and other terms.
For a Boltzmann distribution, the sufficient statistic $u(x) := -\epsilon(x)$ is the negative energy function, $\eta := \frac{1}{T}$ is the inverse temperature, and we use $Z(\eta) = \frac{1}{g(\eta)}$. Write out $KL\left(Q(X)\|P(X|\eta)\right)$, and use the fact that it is non-negative to write out an inequality for $-T\log Z(\eta)$.
This is called the Helmholtz free energy. When the expected energy is constrained

to be constant, how does the entropy of $Q(X)$ compare to the Helmholtz free energy?
For what distribution is the Helmholtz free energy maximized (called the Gibbs free energy)? How does this relate to the maximum entropy distribution?

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:

$$\tilde{P}_\Phi(\mathcal{X}) = \prod_i \phi_i(D_i) \tag{20}$$

Problem 5.1. *If each of the factors is an exponential family distribution (not necessary the same family), write the joint distribution as an exponential family.*

Problem 5.2. *Log-Linear PGMs For positive factors, we could could express themselves in terms of an energy function*

$$\phi_i(D_i) = e^{-\epsilon(D)} \tag{21}$$

For an energy-based model, we assume an energy function $\epsilon(D)$, but we have an additional temperature term, $e^{-\frac{\epsilon(D)}{\tau}}$.
If we take the product of different energy-based models, with different temperatures, what are the sufficient statistics and the natural parameters of the joint distribution?

Problem 5.3. *Express an Ising model and a restricted Boltzmann machine as a log-linear energy-based model. What are the energies?*

Expectation Propagation

Problem 5.4. *Whenever we perform a marginalization, we have to perform an integration, which is not necessarily analytic, or a conjugate prior for the next factor.*
Suppose $P(Z) = \mathcal{N}(\mu, \Sigma)$, and $P(X|Z)$ is Bernoulli (specifically it is a logistic model). Suppose we want to approximate the posterior, $P(Z|X)$. We can treat $P(X|Z)$ as an unnomralized distribution of Z, and make it conjugate to $P(Z)$. How would we do this, using M-projections?

6 Exponential families and convex optimization duality

This section is focuses on a convex optimization perspective of exponential families and can be skipped. It is still under construction.

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.

$$f^*(\lambda) = \sup_{x \in X} \left(\langle \lambda, x \rangle - f(x) \right) \tag{22}$$

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

$$G(\eta) = \log Z(\eta) = \int h(x) e^{\langle u(x), \eta \rangle} dx \tag{23}$$

Then we can re-expresss the exponential family formula as

$$P(x|\eta) = h(x) e^{\langle u(x), \eta \rangle - G(\eta)} \tag{24}$$

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

$$F(\lambda) = \sup_{\eta} \left(\langle \lambda, \eta \rangle - G(\eta) \right) \tag{25}$$

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

$$f^*(\lambda) + f(x) \leq \langle \lambda, x \rangle \tag{26}$$

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

$$P(x|\eta) = h(x) e^{\langle u(x), \eta \rangle - G(\eta)} \leq h(x) e^{F(\bar{u})} \tag{27}$$

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:

$$P(\eta|\lambda) = h^*(\eta) e^{\langle \lambda, \eta \rangle - F(\lambda)} \tag{28}$$

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

$$P(\eta|\lambda) = h^*(\eta) e^{\langle \lambda, \eta - \eta^* \rangle + G(\eta)} \tag{29}$$

¡++¿

7 Conjugate Priors, EM and variational inference

Bishop 2.4.2

For every exponential family of the form $P(x|\eta) = g(\eta) h(x) e^{\langle u(x), \eta \rangle}$, there exists a conjugate prior in the form

$$P(\eta|\chi, \nu) = f(\chi, \nu) g(\eta)^\nu e^{\nu \langle \eta, \chi \rangle} \tag{30}$$

Problem 7.1. *What is the normalized posterior distribution for η, in terms of f, g, X, χ, ν? Also, I believe Bishop may have made a mistake later on in chapter 10.4, it should be $\chi_N = \frac{\sum_n u(x_n)+\nu\chi}{\nu+N}$.*

Challenge problem

Problem 7.2. *Suppose $P(Y|X, W)$ is in an exponential family, and assume $\eta = W\phi(X)$. Come up with a prior for the weights that is conjugate to this exponential family, and derive a formula for the posterior of the weights.*

Expectation Maximization

Problem 7.3. *Suppose $P(X|Z)$ is a generalized linear model and $P(Z)$ is a conjugate prior.*

$$\begin{aligned} P(X|Z;W) &= g(WZ)h(X)e^{\langle u(X),WZ\rangle} \\ P(Z|\chi,\nu) &= f(\chi,\nu)g(Z)^\nu e^{\langle Z,\nu\chi\rangle} \end{aligned} \tag{31}$$

What is the posterior distribution, $P(Z|X, W, \chi, \nu)$?
Maximize W, χ, ν with respect to $W^{old}, \chi^{old}, \nu^{old}$. What expressions do you need to know expectations in order to calculate this maximization?

Variational Inference

Problem 7.4.

Dynamic Bayesian Networks

Problem 7.5. *Suppose we have a hidden markov model, where $P(X^{(n)|Z^{(n)}}$ is from an exponential family, and $P(Z^{(n+1)}|Z^{(n)})$ is based off the conjugate prior (maybe assume it's a linear model, to be similar to Kalman filters). Derive a forward message passing / belief propagation algorithm.*