1 Conjugate Function

The convex conjugate of the function \( f \) is the function
\[
f^*(y) = \sup_x [y^T x - f(x)]
\] (1)

The domain of \( f^* \) is the set of \( y \)'s for which the supremum above is finite. Note that \( f^* \) is always convex in \( y \), as a supremum of linear functions in \( y \).

Let \( g(x, y) = y^T x - f(x) \). If \( f \) is differentiable and convex, then \( \sup_x g(x, y) \) can be calculated by taking the derivative w.r.t \( x \).
\[
\nabla_x g(x, y) = y - \nabla f(x) = 0
\] (2)
\[
y = \nabla f(x) \Rightarrow \text{solution } x^*
\] (3)

If \( f \) is convex, then \( x^* \) is a maximum. If the solution above is unique, then we say the pair \( (x^*, y) = (x^*, \nabla f(x^*)) \) is a Legendre conjugate pair. If the solution is unique for every \( y \), then we can write
\[
f^*(y) + f(x^*) = y^T x^* = \nabla f(x^*)^T x^*
\] (4)

Because at \( x^* \) is the supremum of \( g(x, y) \), it follows that for every \( x \) in the domain of \( f \) the r.h.s is no larger than the l.h.s, that is
\[
f^*(y) + f(x) \geq y^T x \quad \text{for all } x, y
\] (5)
This is called the **Fenchel-Legendre inequality**.

**Proposition** If \( f \) is convex, and epi \( f \) is closed, then \( f^{**} = f \). In this case, we call \( f, f^* \) a **Legendre conjugate pair of functions**.

**Exercise** Calculate the conjugates of: \( e^x, -\ln x, Ax + b, x \in \mathbb{R}^d, 1 - x^2, x \in [0, 1], \ln(e^{x_1} + e^{x_2} + \ldots + e^{x_m}), \frac{1}{2}, ||x||^2 \). Find the comains of the respective conjugate functions, and find examples of vectors \( y \) which are not in those domains.

Remark: If \( f \) convex, and \( y \in \partial f(x) \) for some \( x \), then \( y \in \text{dom } f^* \) and \( y, x \) are a conjugate pair.

## 2 Bregman divergences

Let \( \phi \) be a strictly convex and differentiable function. The **Bregman divergence** between \( x, y \in \text{dom } \phi \) is

\[
d_\phi(y, x) = \phi(y) - \phi(x) - \nabla \phi(x)^T (y - x)
\]

(6)

The geometric significance of the Bregman divergence is illustrated by the following picture. The Bregman divergence is the vertical distance at \( y \) between the graph of \( f \) and the tangent to the graph of \( f \) in \( x \).
Examples

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>$d_\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td></td>
</tr>
<tr>
<td>$x \ln x$</td>
<td>$y \ln \frac{x}{y} - (y - x)$</td>
</tr>
<tr>
<td>$-H(p) = \sum_j p_j \ln p_j$</td>
<td>$KL(q</td>
</tr>
<tr>
<td>$\sum p_j = \sum q_j = 1$</td>
<td>btw. distributions $p, q$</td>
</tr>
</tbody>
</table>

Exercise Find the Bregman divergences corresponding $e^x$, $-\ln x$, $ax + b$, $\ln(1 + e^x)$, $\frac{1}{x}$.

Properties of the Bregman divergence

1. $d_\phi(y, x) \geq 0$ (because the tangent to the epigraph is always below the graph)
2. convex in $y$ (easy to verify)
3. linear in $\phi$ (easy to verify)
4. invariant to addition of affine function $d_\phi + bT = d_\phi$ (easy to verify)
5. **Linear separation** $\{ x \mid d_\phi(x, u) = d_\phi(x, v) \}$ is a hyperplane.

**Proof**

\[
d_\phi(x, u) = d_\phi(x, v)
\]

\[
\phi(x) - \phi(u) - \nabla \phi(u)^T (x - u) = \phi(x) - \phi(v) - \nabla \phi(v)^T (x - v)
\]

\[
[\nabla \phi(u) - \nabla \phi(v)]^T x - [\nabla \phi(u)^T u - \nabla \phi(v)^T v - \phi(u) + \phi(v)] = 0
\]

The last equation defines a hyperplane.

6. **Centering**

\[
\min_u E_p [d_\phi(X, u)] = E_p [d_\phi(X, \mu)] \text{ where } \mu_p \equiv E_p [X] \text{ for any probability distribution } p \text{ over } X
\]
Proof Denote $J(u) = E_p[d_\phi(X, u)]$. Then,

$$J(u) - J(\mu) = \sum_x p(x)d_\phi(x, u) - \sum_x p(x)d_\phi(x, \mu)$$

$$= \sum_x p(x)[\phi(x) - \phi(u) - \nabla\phi(u)^T(x - u) - \phi(x) + \phi(\mu) - \nabla\phi(\mu)^T(x - \mu)]$$

$$= \phi(\mu) - \phi(u) - \nabla\phi(u)^T[\sum_x p(x)x - u] - \nabla\phi(\mu)^T[\sum_x p(x)x - \mu]$$

$$= \phi(\mu) - \phi(u) - \nabla\phi(u)^T[\mu - u]$$

$$= d_\phi(u, \mu) \geq 0$$

7. Conjugate duality Let $\psi(\theta) = \phi^*(\theta)$ be the conjugate of $\phi(\mu)$. Then $d_\phi(\mu_1, \mu_2) = d_\psi(\theta_2, \theta_1)$

Proof

$$d_\phi(\mu_1, \mu_2) = \phi(\mu_1) - \phi(\mu_2) - (\mu_1 - \mu_2)^T \nabla\phi(\mu_2)$$

$$= \phi(\mu_1) - \phi(\mu_2) - (\mu_1 - \mu_2)^T \theta_2 + \mu_1^T \theta_1 - \mu_1^T \theta_1$$

$$= [-\mu_1^T \theta_1 + \phi(\mu_1)] + [\mu_2^T \theta_2 - \phi(\mu_2)] - \mu_1^T \theta_2 + \mu_1^T \theta_1$$

$$= d_\psi(\theta_2, \theta_1)$$

3 Exponential family models

A family of probability distributions that can be put in the form below is called and exponential family model.

$$p_\theta(x) = \frac{1}{Z(\theta)} e^{\theta^T x}$$

In the above, $x \in \mathbb{R}^n$ are the natural coordinates (or sufficient statistics), $\theta \in \mathbb{R}^n$ are the natural parameters of the exponential family, and $Z$
is the normalization constant.

\[ Z(\theta) = \sum_x e^{\theta^T x} \quad (21) \]

\( Z(\theta) \) is convex (in \( \theta \)) as sum of the convex functions \( e^{\theta^T x} \).

**Remarks**

1. The general form of an exponential family model is

\[ \log p(x) = \theta^T t(x) - \log Z(\theta) + \log a(\gamma) + \log c(x), \]

with \( \gamma \) another parameter called *nuisance parameter* and \( a > 0 \) a scaling function, and \( c(x) \) a given probability measure. Here we will ignore \( a \) and \( c \), as they do not have any influence on the estimation of the parameter \( \theta \), nor on any of the properties of the exponential family that we study here, but the transformation \( x \mapsto t(x) \) will be used frequently.

2. Exponential families are defined for \( x \in \Omega \subseteq \mathbb{R}^n \), which can be a discrete or a continuous sample space. Therefore we will alternate between \( \sum_x \) and \( \int dx \), where the sum and the integral are over \( \Omega \).

We will find it useful to work with \( \ln Z(\theta) \) which is the *partition function* or the *cumulant function*.

\[ \psi(\theta) = \ln Z(\theta) = \ln \sum_x e^{\theta^T x} \quad (22) \]

This function is always convex in \( \theta \) as the composition of the convex increasing function \( \log \sum_i e^{y_i} \) with the linear functions \( y_i = x_i^T \theta \).

In the following we will assume (implicitly) various regularity conditions, for instance that the normalization constant is finite in a domain that contains a convex, open set, and that the coordinates \( x \) are linearly independent functions.

Exponential family models comprise (multivariate) normal distributions, Markov random fields (with positive distributions), binomial and multinomial models, etc. They have many convenient properties, some of which are evident from the definition above. For example, exponential family models are essentially the only parametric models that have fixed dimensional sufficient statistics\(^1\); they have *conjugate priors*; from the differential geometry p.o.v, exponential families represent *flat manifolds*, i.e affine function spaces spanned by the vectors \( \theta_i \). We will show some of these properties here.

\(^1\)Distributions that are piecewise uniform may also have finite sufficient statistics. In their case, the sufficient statistics are intervals in which the data lie.
Using (22) we can express the distribution $p(x)$ as

$$p_\theta(x) = e^{\theta^T x - \psi(\theta)}$$  \hspace{1cm} (23)

3.1 Examples

Normal univariate distribution with unit variance

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} = \exp\left\{ -\frac{x^2}{2} + \mu x - \frac{\mu^2}{2} + \frac{1}{2} \ln(2\pi) \right\}$$ \hspace{1cm} (24)

The natural parameter is $\mu \in \mathbb{R}$, the vector of sufficient statistics is one dimensional, equal to $x$, and there is a non-trivial $c(x)$ component of the model, that influences what $Z(\theta)$ is, but not the natural parameter $\mu$.

Normal univariate distribution

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ \hspace{1cm} (25)

$$= e^{\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2} + \ln(2\pi\sigma^2)}$$ \hspace{1cm} (26)

This univariate distribution has 2 natural parameters $\theta_1 = \frac{1}{2\sigma^2}$, $\theta_2 = \frac{\mu}{\sigma^2}$ and a vector of sufficient statistics $t(x) = [x^2, x] \in \mathbb{R}^2$. Note that in this case the natural coordinates/sufficient statistics have a different dimension than the original variable $x$. The log-partition function $\psi$ is expressed in natural parameters is

$$\psi(\theta) = -\frac{\theta_2^2}{\theta_1} + \ln(-\pi/\theta_1)$$ \hspace{1cm} (27)

which is strictly convex (verify by taking the Hessian) when $\theta_1 < 0$. The domain $\Theta$ of the natural parameters is $(-\infty, 0) \times \mathbb{R}$.

Bernoulli distribution Here $\Omega = \{0, 1\}$ and let $p = Pr[X = 1]$ the probability of success in the Bernoulli trial.

$$P_p(x) = p^x(1-p)^{1-x} = e^{(\ln p)x + \ln(1-p)(1-x)}$$ \hspace{1cm} (28)

and the natural parameters are $\theta = [\ln p \ln(1-p)]$ with $\psi(\theta) = 0$, $Z(\theta) = 1$. We will return to this model in the next section.
3.2 Expectations, moments and convexity

1. $E_{\theta}[X] \equiv \mu(\theta) = \nabla \psi(\theta)$

Proof

\[
\nabla \psi(\theta) = \frac{\nabla_{\theta} \left( \sum_x e^{\theta^T x} \right)}{Z(\theta)} = \frac{\sum_x x e^{\theta^T x}}{Z(\theta)} = \sum_x \frac{x e^{\theta^T x}}{Z(\theta)} = \sum_x x p(x) = E_{\theta}[X]
\]

2. $\text{Var}_{\theta}[X] = \nabla^2 \psi(\theta)$

Proof

\[
\nabla^2 \psi(\theta) = \nabla^T_{\theta} \left[ \frac{\sum_x x e^{\theta^T x}}{Z(\theta)} \right] = \sum_x \left\{ x x^T e^{\theta^T x} \frac{Z(\theta)}{Z^2(\theta)} + x e^{\theta^T x} \left[ - \frac{\nabla^T Z(\theta)}{Z^2(\theta)} \right] \right\} = \left\{ \sum_x x x^T p(x) - x e^{\theta^T x} \left[ - \frac{\sum_{x'} x' e^{\theta^T x'}}{Z^2(\theta)} \right] \right\} = E_{\theta}[x x^T] - E_{\theta}[x] (E_{\theta}[x])^T = \text{Var}_{\theta}X
\]

3. From Property 2, because the variance is always positive definite, we obtain an alternative proof that $\psi(\theta)$ is convex.

4. $\ln p_{\theta}(x) = \theta^T x - \psi(\theta)$ is concave in $\theta$ and linear in $x$. Hence $p$ is log-concave in $\theta$, and is a log-linear model in $x$.

5. From 4 we also expect that, (under mild regularity conditions) the Maximum Likelihood estimate (when it exists) to be unique, and computationally easy to find, as the unique local maximum of the log-likelihood.
Remark: Let us assume that the sufficient statistics $x_1, \ldots, x_n$ are affinely independent random variables. Then, $\text{Var} X$ is non-singular, and consequently, $\nabla^2 \ln p_\theta(x) = -\nabla^2 \psi(\theta) \prec 0$, implying that the log-likelihood is strictly concave, and has at most one global maximum.

**Example: Univariate Normal** By taking the gradient of $\psi(\theta)$, we obtain

$$
\nabla \psi = \begin{bmatrix}
-\frac{1}{2n} + \frac{\theta_2^2}{2 \sigma^2} \\
-\frac{\theta_2}{2n}
\end{bmatrix} = \begin{bmatrix}
\sigma^2 + \mu^2 \\
\mu
\end{bmatrix} = \begin{bmatrix}
E[X] \\
E[X^2]
\end{bmatrix}
$$

Furthermore (Exercise) the Hessian of $\psi$ will give us (on its diagonal) the variance of $X^2$, respectively the variance of $X$, which is of course $\sigma^2$.

**Example: Bernoulli** In the above, we obtained $\phi(\theta) \equiv 0$ for the Bernoulli. Hence, its gradient cannot give us the expectation of $X$. What is wrong? The problem is that the sufficient statistic $t(x) = [x_1 - x]$ is not a vector of affinely independent functions of $x$. (This happens generally for distributions over discrete sample spaces if we are not careful.) It is said that the model (28) is not in standard form.

We reparametrize $P_p$ using a single sufficient statistic and a single parameter $\theta$.

\[
x \ln p + (1 - x) \ln(1 - p) = x[\ln p - \ln(1 - p)] + \ln(1 - p)
\]

\[
\theta = \ln \frac{p}{1 - p}
\]

\[
\psi(\theta) = \ln(1 - p) = \ln \frac{1}{1 + e^\theta}
\]

Now, $\psi'(\theta) = \frac{e^\theta}{1 + e^\theta} = p = E[X]$ (by replacing $\theta$ with its value in (39)).

Let us examine ML estimation closer. Assume we have an i.i.d sample $x_1, x_2, \ldots, x_n$. The likelihood of the sample is

\[
p_\theta(x^{1:n}) = \prod_{i=1}^{n} e^{\theta^T x_i - \psi(\theta)}
\]

\[
= e^{\theta^T \sum_{i=1}^{n} x_i - n\psi(\theta)}
\]

\[
= e^{n[\theta^T \bar{x} - \psi(\theta)]}
\]
and the ML estimation equation is
\[
\max_{\theta} g(\theta, \bar{x}) = \bar{x}^T \theta - \psi(\theta)
\]  
(44)

Comparing the above equation with (1) we find that \(\theta^{ML}\) is Legendre conjugate with \(\bar{x} = \left(\sum_{i=1}^{n} x^i\right)/n\) and that the max log-likelihood (= log-likelihood at \(\theta^{ML}\)) \(\phi(\bar{x})\) is the Legendre conjugate function of \(\psi(\theta)\). Moreover, maximizing the likelihood is equivalent to solving the equations

\[
\bar{x} = \nabla \psi(\theta);
\]  
(45)

but from Property 1 we know that \(\nabla \psi(\theta) = E_\theta[X]\). Hence, the ML equations for an exponential family model amount to solving for \(\theta\) in

\[
E_\theta[X] = \frac{\sum_i x^i}{n}
\]  
(46)

In other words, \(\theta^{ML}\) is the parameter value for which the model expectation equals the sample mean of the data (=the expectation under the empirical distribution). (Exercise: Normal distribution).

6. Returning to the general expression of the log-likelihood, for any \(\theta\), the Legendre conjugate parameter \(\mu\) is given by (3) \(\mu = \nabla_\theta \psi = E_\theta[X]\). In other words, the conjugate pairs \(\theta, \mu\) represent the (parameter, mean value) pairs. The dual parametrization of the model in terms of \(\mu, \phi(\mu)\) is called the Mean value parametrization.

The domain of \(\phi(\mu)\), i.e the set \(\{E_\theta[X]\}_\theta\) is called the marginal polytope of the exponential family. Exercise: is the normal distribution’s \((\mu, \sigma^2)\) parametrization a mean value parametrization? For the Bernoulli, since \(p = E[X]\) (Exercise: check this via \(\nabla \psi(\theta)\)), the usual parametrization is the mean value parametrization.

7. The gradient of the log-likelihood w.r.t the parameters has the simple formula

\[
\nabla_\theta \frac{1}{N} \ln p_\theta(x^{1:N}) = \bar{x} - \nabla_\theta \psi(\theta) = \bar{x} - E_\theta[x]
\]  
(47)

Thus, when we fit the models by e.g gradient ascent, the direction of ascent is the difference between the data expectations and the model expectations.
Example: Generalized Linear Models (GLM)

A GLM is a regression where the “noise” distribution is in the exponential family.

- \( y \in \mathbb{R}, y \sim P_\theta \) with
  \[
  P_\theta(y) = e^{\theta y - \ln \psi(\theta)}
  \] (48)

- the parameter \( \theta \) is a linear function of \( x \in \mathbb{R}^d \)
  \[
  \theta = \beta^T x
  \] (49)

- We denote \( E_\theta[y] = \mu \). The function \( g(\mu) = \theta \) that relates the mean parameter to the natural parameter is called the link function. The link function is given by \( g(\mu) = (\nabla \psi)^{-1}(\mu) \).

The log-likelihood (w.r.t. \( \beta \)) is

\[
\ell(\beta) = \ln P_\theta(y|x) = \theta y - \psi(\theta) \quad \text{where} \quad \theta = \beta^T x
\] (50)

and the gradient w.r.t. \( \beta \) is therefore

\[
\nabla_\beta \ell = \nabla_\theta \ell \nabla_\beta (\beta^T x) = (y - \mu)x
\] (51)

This simple expression for the gradient is the generalization of the gradient expression you obtained for the two layer neural network in STAT 535. [Exercise: This means that the sigmoid function is the inverse link function defined above. Find what is the link function that corresponds to the neural network.]

8. \( H(p_\theta) \equiv H(\theta) = \psi(\theta) - \theta^T E[X] \)

\[
\text{Proof} - H(\theta) = \sum_x p_\theta(x) \ln p_\theta(x)
\] (52)

\[
= \sum_x p_\theta(x)[\theta^T x - \psi(\theta)]
\] (53)

\[
= \theta^T \sum_x p_\theta(x)x - \psi(\theta)
\] (54)

\[
= \theta^T \mu(\theta) - \psi(\theta) = \phi(\mu)
\] (55)

It follows also that \( H(\theta) = -\psi^*(\mu) \equiv -\phi(\mu) \). The conjugate of \( \psi \) is the negative entropy.
9. \( KL(\theta_1, \theta_2) = d_\psi(\theta_2, \theta_1) = d_\phi(\mu_1, \mu_2) \)

**Proof** We need to prove only one of the equalities, because the other follows from Property 7 of the Bregman divergence.

\[
KL(\theta_1, \theta_2) = \sum_x p_{\theta_1}(x) \ln \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} = \sum_x p_{\theta_1}(x) [\theta_1^T x - \psi(\theta_1) - \theta_2^T x + \psi(\theta_2)]
\]

\[
= (\theta_1 - \theta_2)^T \left[ \sum_x p_{\theta_1}(x)x \right] - \psi(\theta_1) + \psi(\theta_2)
\]

\[
= \psi(\theta_2) - \psi(\theta_1) + (\theta_1 - \theta_2)^T \nabla \psi(\theta_1)
\]

\[
= d_\psi(\theta_2, \theta_1)
\]

10. Likelihood and KL divergence (see Lecture 8)