STAT 538 Lecture 7
Conjugate Function, Bregman Divergences, Exponential Family Models

©Marina Meilă
mmp@stat.washington.edu

Reading: Gill (Chapters 2–4)

1 Conjugate Function

The **convex conjugate** of the function f is the function

$$f^*(y) \;=\; \sup_x [y^T x - f(x)] \tag{1}$$

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.

$$\begin{aligned} \nabla_x g(x,y) &\;=\; y - \nabla f(x) \;=\; 0 & (2)\\ y &\;=\; \nabla f(x) \quad \Rightarrow \text{ solution } x^* & (3) \end{aligned}$$

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

$$f^*(y) + f(x^*) \;=\; y^T x^* \;=\; \nabla f(x^*)^T x^* \tag{4}$$

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

$$\boxed{f^*(y) + f(x) \;\geq\; y^T x \quad \text{for all}\, x, y} \tag{5}$$

This is called the **Fenchel-Legendre** inequality.

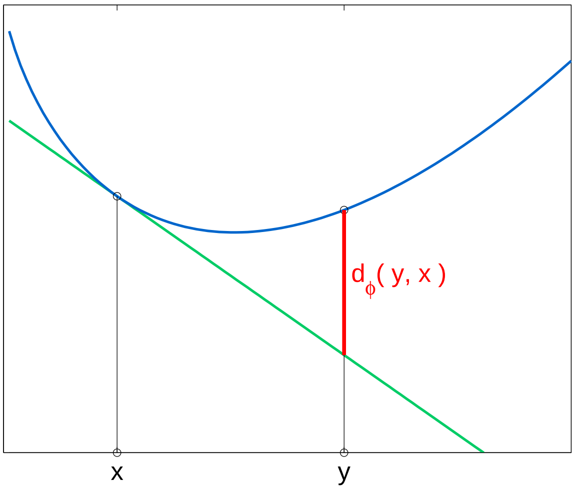
Proposition If f is convex, and $\operatorname{epi} f$ is closed, then $f^{**} = f$. In this case, we call f, f^* a **Legendre conjugate pair of functions**.

2 Bregman divergences

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

$$d_\phi(y,x) \;=\; \phi(y) - \phi(x) - \nabla\phi(x)^T(y-x) \tag{6}$$

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

ϕ	d_ϕ	
$\|\|x\|\|^2$	$\|\|x-y\|\|^2$	squared Euclidean distance
$x\ln x$	$y\ln\frac{y}{x} - (y-x)$	
$-H(p) = \sum_j p_j \ln p_j$	$KL(q\|\|p) = \sum_j q_j \ln\frac{q_j}{p_j}$	Kullbach − Leibler divergence
	$\sum p_j = \sum q_j = 1$	btw. distributions p, q

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 ϕ (easy to verify)

4. invariant to addition of affine function $d_{\phi+b^Tx+c} = d_\phi$ (easy to verify)

5. **Linear separation** $\{\, x \,|\, d_\phi(x,u) = d_\phi(x,v) \,\}$ is a hyperplane.
   Proof

$$\begin{aligned} d_\phi(x,u) &= d_\phi(x,v) & (7)\\ \phi(x) - \phi(u) - \nabla\phi(u)^T(x-u) &= \phi(x) - \phi(v) - \nabla\phi(v)^T(x-v) & (8) \end{aligned}$$

$$[\nabla\phi(u) - \nabla\phi(v)]^T x - [\nabla\phi(u)^T u - \nabla\phi(v)^T v - \phi(u) + \phi(v)] \;=\; 0 \quad (9)$$

   The last equation defines a hyperplane.

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

$$\begin{aligned} & J(u) - J(\mu) & (10)\\ =\; & \sum_x p(x) d_\phi(x,u) - \sum_x p(x) d_\phi(x,\mu) & (11)\\ =\; & \sum_x p(x)[\phi(x) - \phi(u) - \nabla\phi(u)^T(x-u) - \phi(x) + \phi(\mu) - \nabla\phi(\mu)^T(x-\mu)] & \\ =\; & \phi(\mu) - \phi(u) - \nabla\phi(u)^T[\underbrace{\sum_x p(x)x}_{\mu} - u] - \nabla\phi(\mu)^T[\sum_x p(x)x - \mu)] & (12)\\ =\; & \phi(\mu) - \phi(u) - \nabla\phi(u)^T[\mu - u] & (13)\\ =\; & d_\phi(u,\mu) \;\geq\; 0 & (14) \end{aligned}$$

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

$$\begin{aligned} d_\phi(\mu_1,\mu_2) &= \phi(\mu_1)-\phi(\mu_2)-(\mu_1-\mu_2)^T\underbrace{\nabla\phi(\mu_2)}_{\theta_2} && (15)\\ &= \phi(\mu_1)-\phi(\mu_2)-(\mu_1-\mu_2)^T\theta_2+\mu_1^T\theta_1-\mu_1^T\theta_1 && (16)\\ &= [-\mu_1^T\theta_1+\phi(\mu_1)]+[\mu_2^T\theta_2-\phi(\mu_2)]-\mu_1^T\theta_2+\mu_1^T\theta_1 && (17)\\ &= -\psi(\theta_1)+\psi(\theta_2)-\underbrace{\mu_1^T}_{\nabla\psi(\theta_1)^T}(\theta_2-\theta_1) && (18)\\ &= d_\psi(\theta_2,\theta_1) && (19)\end{aligned}$$

3 Exponential family models

A family of probability distributions that can be put in the form below is called and **exponential family** model[1].

$$p_\theta(x) = \frac{1}{Z(\theta)}e^{\theta^T x} \tag{20}$$

In the above, $x \in \mathbb{R}^n$ are the **natural coordinates**, $\theta \in \mathbb{R}^n$ are the **natural parameters** of the exponential family, and Z is the normalization constant. We will find it useful to work with $\ln Z(\theta)$ which is the **partition function** or the **cumulant function**.

$$\psi(\theta) = \ln Z(\theta) = \ln\sum_x e^{\theta^T x} \tag{21}$$

This function is always convex in θ 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 (implictly) 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.

[1] The general form of an exponential family model is $\log p(x) = \frac{\theta^T f(x)-\log Z_(\theta)}{c(\gamma)} + \log c'(x,\gamma)$, with γ another parameter called *nuisance parameter* and c, c' some 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 a finite number of sufficient statistics[2]; they have **conjugate priors**; from the differential geometry p.o.v, exponential families repreent **flat manifolds**, i.e affine function spaces spanned by the vectors θ_i. We will show some of these properties here.

Using (21) we can express the distribution $p(x)$ as

$$p_\theta(x) \;=\; e^{\theta^T x - \psi(\theta)} \tag{22}$$

3.1 Examples

TBW

3.2 Expectations, moments and covexity

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

$$\begin{aligned}
\nabla\psi(\theta) \;&=\; \frac{\nabla_\theta\left(\sum_x e^{\theta^T x}\right)}{Z(\theta)} && (23)\\
&=\; \frac{\sum_x x e^{\theta^T x}}{Z(\theta)} && (24)\\
&=\; \sum_x x \frac{e^{\theta^T x}}{Z(\theta)} && (25)\\
&=\; \sum_x x p(x) \;=\; E_\theta[X] && (26)
\end{aligned}$$

[2] Distributions that are piecewise uniform may also have finite sufficient statistics. In their case, the sufficient statistics are intervals in which the data lie.

2. $\boxed{Var_\theta[X] \ = \ \nabla^2\psi(\theta)}$
Proof

$$\begin{aligned}
\nabla^2\psi(\theta) &= \nabla_\theta^T\left[\frac{\sum_x xe^{\theta^T x}}{Z(\theta)}\right] && (27)\\
&= \sum_x\left\{xx^T\frac{e^{\theta^T x}}{Z(\theta)} + xe^{\theta^T x}\left[-\frac{\nabla^T Z(\theta)}{Z^2(\theta)}\right]\right\} && (28)\\
&= \left\{\sum_x xx^T p(x) - xe^{\theta^T x}\left[-\frac{\sum_{x'} x'e^{\theta^T x'}}{Z^2(\theta)}\right]^T\right\} && (29)\\
&= E_\theta[xx^T] - E_\theta[x](E_\theta[x])^T \ = \ Var_\theta X && (30)
\end{aligned}$$

3. From Property 2, because the variance is always positive definite, we conclude that $\boxed{\psi(\theta) \text{ is convex}}$.

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

$$\begin{aligned}
p_\theta(x^{1:n}) &= \prod_{i=1}^n e^{\theta^T x^i - \psi(\theta)} && (31)\\
&= e^{\theta^T \sum_{i=1}^n x^i - n\psi(\theta)} && (32)\\
&= e^{n[\theta^T \bar{x} - \psi(\theta)]} && (33)
\end{aligned}$$

and the ML estimation equation is

$$\max_\theta g(\theta, \bar{x}) = \bar{x}^T\theta - \psi(\theta) \qquad (34)$$

Comparing the above equation with (1) we find that
$\boxed{\theta^{ML} \text{ is Legendre conjugate with } \bar{x} = (\sum_{i=1}^n x^i)/n}$ and that the max

log-likelihood (= log-likelihood at θ^{ML}) $\boxed{\phi(\bar{x}) \text{ is the Legendre conjugate function of } \psi(\theta)}$. Moreover, maximizing the likelihood is equivalent to solving the equations $\bar{x} = \nabla\psi(\theta)$; 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 θ in

$$E_\theta[X] \;=\; \frac{\sum_i x^i}{n} \tag{35}$$

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

6. TBW
   Returning to the general expression of the log-likelihood, for any θ, the Legendre conjugate parameter μ is given by (3) $\mu = \nabla_\theta\psi = E_\theta[X]$. In other words, the conjugate pairs θ, μ 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. *Examples Normal, binomial*

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

$$\nabla_\theta \frac{1}{N} \ln p_\theta(x^{1:N}) \;=\; \bar{x} - \nabla_\theta\psi(\theta) \;=\; \bar{x} - E_\theta[x] \tag{36}$$

   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

$$P_\theta(y) \;=\; e^{\theta y - \ln\psi(\theta)} \tag{37}$$

   - the parameter θ is a linear function of $x \in \mathbb{R}^d$

$$\theta \;=\; \beta^T x \tag{38}$$

- 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 log-likelihood (w.r.t. β) is

$$l(\beta) \;=\; \ln P_\theta(y|x) \;=\; \theta y - \psi(\theta) \quad \text{where } \theta = \beta^T x \tag{39}$$

and the gradient w.r.t. β is therefore

$$\nabla_\beta l \;=\; \nabla_\theta l \nabla_\beta(\beta^T x) \;=\; (y - \mu)x \tag{40}$$

This simple expression for the gradient is the generalization of the gradient expression you obtained for the two layer neural network in the homework. [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. $\boxed{\psi^*(\mu) \equiv \phi(\mu) = -H(\theta)}$. The dual of ψ is the negative entropy.
Proof

$$\begin{aligned} -H(\theta) &= \sum_x p_\theta(x) \ln p_\theta(x) && (41) \\ &= \sum_x p_\theta(x)[\theta^T x - \psi(\theta)] && (42) \\ &= \theta^T \sum_x p_\theta(x) x - \psi(\theta) && (43) \\ &= \theta^T \mu(\theta) - \psi(\theta) \;=\; \phi(\mu) && (44) \end{aligned}$$

9. $\boxed{KL(\theta_1, \theta_2) = d_\psi(\theta_2, \theta_1) = d_\phi(\mu_1, \mu_2)}$
Proof We need to prove only one of the equalities, because the other

follows from Property 7 of the Bregman divergence.

$$\begin{aligned}
KL(\theta_1, \theta_2) &= \sum_x p_{\theta_1}(x) \ln \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} && (45)\\
&= \sum_x p_{\theta_1}(x)[\theta_1^T x - \psi(\theta_1) - \theta_2^T x + \psi(\theta_2)] && (46)\\
&= (\theta_1 - \theta_2)^T \underbrace{[\sum_x p_{\theta_1}(x)x]}_{\mu(\theta_1)=\nabla\psi(\theta_1)} - \psi(\theta_1) + \psi(\theta_2) && (47)\\
&= \psi(\theta_2) - \psi(\theta_1) + (\theta_1 - \theta_2)^T \nabla\psi(\theta_1) && (48)\\
&= d_\psi(\theta_2, \theta_1) && (49)
\end{aligned}$$

10. Likelihood and KL divergence (see Lecture 8)