**Computer Vision**

# Exponential Families and Bayesian Inference

*Lecture 2*

## Exponential Families

An exponential family of distributions is a $d$-parameter family $f(x;\eta)$ having the following form:

$$f(x;\eta) = h(x)e^{\ g(\eta)^t\ T(x)\ -\ B(\eta)}, \tag{2.1}$$

where $\eta = (\eta_1, \ldots, \eta_d) \in R^d$, and $g(\eta) = [g_1(\eta), \ldots, g_d(\eta)]$, for $d$ functions $g_i : R^d \to R$, and $T(x) = [T_1(x), \ldots, T_d(x)]$.

### Some examples with $d = 1$.

Many well-known distributions belong to this family. Let us look at some examples:

1. *Bernoulli distribution* The bernoulli distribution characterizes coin tosses:

   $$P(X;p) = p^X\ (1-p)^{1-X} = e^{X\ \log p + (1-X)\ \log(1-p)}.$$

   Comparing with equation (2.1):
   $\eta = p,\ T(x) = x,\ g(p) = \log \frac{p}{1-p},\ B(p) = log(1-p),\ h(x) = 1.$

2. *Binomial Distribution* The binomial distribution characterizes the number of success (e.g heads) in $n$ trials (coin tosses) i.e $X \in \{0, 1, \ldots, n\}$.

   $$P(X;p) = \left(\begin{array}{c} n \\ x \end{array}\right) p^X (1-p)^{n-X} = \left(\begin{array}{c} n \\ x \end{array}\right) e^{x\ \log \frac{p}{1-p} + \eta\ \log(1-p)}.$$

   Comparing with Eqn 3.1:
   $\eta = p,\ T(x) = x,\ g(p) = \log \frac{p}{1-p},\ B(p) = n\log(1-p),\ h(x) = \left(\begin{array}{c} n \\ x \end{array}\right).$

3. *Poisson Distribution* The poisson distribution is given by:

   $$f(x;\eta) = \frac{\eta^x e^{-\eta}}{x!} = \frac{1}{x!} e^{x\ \log \eta - \eta}.$$

   Comparing with Eqn 3.1:
   $\eta = p,\ T(x) = x,\ g(\eta) = \log \eta,\ B(\eta) = \eta,\ h(x) = \frac{1}{x!}.$

**An example with $d > 1$**

*Normal Distribution* The general univariate normal density is given by:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2-\mu^2+2x\mu}{2\sigma^2} - \frac{1}{2}\log\sigma^2}.$$

which is of the form above, setting
$\eta = [\mu \;\; \sigma]^T$, $T(x) = [x^2 \;\; x]^T$, $g(\eta) = [-\frac{1}{2\sigma^2} \;\; \frac{\mu}{\sigma^2}]^T$, $B(\eta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2$ and $h(x) = \frac{1}{\sqrt{2\pi}}$.

*Exponential Families are closed under Sampling*

If $X_1, \ldots, X_n$ are sampled i.i.d from an exponential family, the joint density has the form:

$$f(X_1, \ldots, X_n; \eta) = \prod_{i=1}^{n} h(X_i) e^{g(\eta)\sum_{i=1}^{n} T(X_i) - nB(\eta)}. \tag{2.2}$$

which also belongs has an exponential form, with $h^n(X_1, \ldots, X_n) e^{\; g^n(\eta)\; T^n(X_1,\ldots,X_n) \; - \; B^n(\eta)}$, $T^n(X_1, \ldots, X_n) = \sum_i T(X_i)$, $h^n(X_1, \ldots, X_n) = \prod_i h(x_i)$ and $B^n(\eta) = nB(\eta)$.

## $B(\eta)$ and Normalization

Consider the form $h(x)e^{\; g(\eta)T(x)}$. To turn this exponential form into a density, we need to divide by the normalizing constant

$$\int h(x) e^{g(\eta)T(x)} dx.$$

Define:

$$B(\eta) = \log \int h(x) e^{\; g(\eta)T(x)} dx.$$

so that $\int h(x) e^{g(\eta)T(x)} = e^{B(\eta)}$. Now, the exponential form becomes a density that integrates to 1:

$$f(x;\eta) = \frac{h(x)e^{\; g(\eta)T(x)}}{e^{B(\eta)}} = h(x) e^{g(\eta)T(x) - B(\eta)}.$$

So $B(\eta)$ is the log of the normalizing constant.

## Derivatives of $B(\eta)$ and Moments of $T$

$$\text{Define: } A(g) = \log \int h(x) e^{\; g\; T(x)} \; dx$$
$$\text{so that } \; B(\eta) = A(g(\eta))$$

Taking the derivative of $A$ with respect to $g$, we have:

$$A'(g) = \frac{\int T(x)\, h(x)\, e^{\,g\, T(x)}\, dx}{\int h(x)\, e^{\,g\, T(x)}\, dx}.$$

This shows that the derivative of the normalizing constant gives the Expectation of $T$.

$$A'(g(\eta)) = E_\eta T(X). \tag{2.3}$$

One can also verify:

$$A''(g(\eta)) = Var_\eta\, T(X).$$

More generally, a connection between $m^{th}$ derivative and $m^{th}$ moment of $T(X)$ can be established. This is a very useful result since the problem of estimating moments which involves computing integrals has been turned into a problem of differentiating a function.

## Maximum Likelihood Estimation

We now use the above properties for Maximum likelihood estimation based on i.i.d samples $X_1, \ldots, X_n$. The joint density is given in Eqn 3.2. The log-likelihood function is given by taking logs in Eqn 3.2:

$$l(X_1, \ldots, X_n; \eta) = g(\eta) \sum_{i=1}^{n} T(X_i) - nB(\eta) + \sum_{i=1}^{n} h(X_i).$$

The MLE estimate is obtained by maximizing the function above:

$$\begin{aligned}
\hat{\eta} &= \operatorname*{argmax}_{\eta}\ l(X_1, \ldots, X_n|\eta) \\
&= \operatorname*{argmax}_{\eta} g(\eta) \sum_{i=1}^{n} T(X_i) - nB(\eta) + \sum_{i=1}^{n} h(X_i) \\
&= \operatorname*{argmax}_{\eta} g(\eta) \sum_{i=1}^{n} T(X_i) - nB(\eta) \\
&= \operatorname*{argmax}_{\eta} g(\eta) \sum_{i=1}^{n} T(X_i) - nA(g(\eta)).
\end{aligned}$$

Now

$$\frac{dl(X_1, \ldots, X_n; \eta)}{d\eta} = \frac{dl(X_1, \ldots, X_n; g(\eta))}{dg} \eta'(g).$$

Setting the derivative equal to 0 we get

$$\sum_{i=1}^{n} T(x_i) - nA'(g(\eta)) = 0,$$

which we rewrite using (2.3)

$$\frac{1}{n}\sum_{i=1}^{n} T(x_i) = E_\eta T(X).$$

Thus, $\eta$ that maximizes likelihood is the $\eta$ for which the true expectation of $T(X)$ equals the sample expectation.

The only way in which the data is involved in the estimation of $\eta$ is via the sample mean $\frac{1}{n}\sum_{i=1}^{n} T(X_i)$, which is refered to as a *sufficient statistic* for inference about $\eta$.

*Multivariate Exponential Family*

The observations above also hold for a multivariate $d$-parameter exponential family,

$$f(x;\eta) = h(x)e^{\,g(\eta)^t T(x) - B(\eta)}$$

with $\eta = [\eta_1 \dots \eta_d]^T$,$T(X) = [T(X_1) \dots T(X_d)]$ $g(\eta) = [g_1(\eta) \dots g_d(\eta)]$. Again defining

$$A(g) = \log \int h(x) e^{g^t T(x)} dx,$$

the following results corresponding to the one-parameter case can be established

$$\nabla_g A(g(\eta)) = \begin{pmatrix} E_\eta T_1(X) \\ \vdots \\ E_\eta T_k(X) \end{pmatrix}.$$

$$\frac{\partial^2 A}{\partial g_i \partial g_j}(g(\eta)) = \mathrm{cov}_\eta(T_i(X), T_j(X)).$$

The maximum likelihood estimate of $\eta$ is made by solving the following set of equations:

$$\frac{1}{n}\sum_{i=1}^{n} T_j(x) = E_\eta T_j \quad j = 1 \dots d.$$

Defining the discrete empirical distribution which is uniform over the values $X_1 \dots X_n$:

$$R_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

we can express the above equalities as:

$$E_{R_X}\, T_j = E_\eta\, T_j.$$

At the ML estimate of $\eta$, the expectations under the empirical distribution equals the true expectation of $T$.

## The Bayesian Approach

So far we have used the maximum likelihood method for defining estimators for $\eta$ which is thought of as a parameter. The Bayesian approach treats parameters $\eta$ as random variables that can be described by probabilistic statements.

Bayesian inference is carried out in the following way:

1. We choose a probability density $P(\eta)$ – called the prior distribution – that expresses our prior beliefs about $\eta$ before we see any data.
2. Define the family of *conditional* distributions $P(X|\eta)$. *Note that since now $\eta$ is a random variable we write $P(X|\eta)$ as opposed to $P(X;\eta)$.*
3. After observing data $X_1 \dots X_n$, we compute the posterior distribution $P(\eta|X_1 \dots X_n)$.

For the thid step, we exploy the Bayes rule:

$$P(\eta|X_1 \dots X_n) = \frac{P(X_1 \dots X_n|\eta)P(\eta)}{P(X_1 \dots X_n)} = \frac{P(X_1|\eta)P(X_2|\eta) \dots P(X_n|\eta)P(\eta)}{P(X_1 \dots X_n)}$$

What can we do with the posterior? Two options are to estimate $\eta$ via the mode or the mean of the posterior distribution. From Bayesian Decision theory, these options correspond to optimizing with respect to a zero-one cost or a squared cost respectively.

To maximize the posterior:

$$\begin{aligned} \hat{\eta} &= \operatorname*{argmax}_{\eta} P(\eta|X_1 \dots X_n) = \operatorname*{argmax}_{\eta} \log P(\eta|X_1 \dots X_n) \\ &= \operatorname*{argmax}_{\eta} \sum_{i=1}^{n} \log P(X_i|\eta) + \log P(\eta). \end{aligned}$$

Note that the normalizing term $P(X_1 \dots X_n)$ can be ignored.

To estimate via the mean of the posterior.

$$\hat{\eta} = \int \eta P(\eta|X_1 \dots X_n)$$

In this case, the normalizing term $P(X_1 \dots X_n)$ cannot be ignored.

## Conjugate Priors

In Bayesian statistics a prior distribution is multiplied by the likelihood function and then normalized to produce a posterior distribution. A conjugate prior is one which, when combined with the likelihood and normalized, produces a posterior distribution which is

of the same family as the prior. In most cases once the unnormalized posterior is known the normalization follows directly from the form of the distribution.

*Example*

If one is estimating the parameter (the success probability) of a Bernoulli distribution, and if one chooses to use a beta distribution as one's prior, then the posterior is always another beta distribution. This allows us to figure out the normalizing constants bypassing their actual computation.

The Bernoulli distribution is given by:

$$P(X|p) = p^X(1-p)^{(1-X)}$$

We put a Beta distribution $B(\alpha, \beta)$ on $p$:

$$P(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where the $\Gamma$ function is a generalization of the factorial to complex and real-valued arguments:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$$

which for integers $\alpha = n$ gives the factorial $\Gamma(n) = (n-1)!$. We know that for a Beta distribution, the Expectation is given by:

$$E\ B(\alpha, \beta) = \frac{\alpha}{\alpha+\beta}$$

Now consider the posterior distribution of $p$ given the i.i.d sampled data:

$$P(p|X_1 \dots X_n) = P(p) \prod_{i=1}^{n} \frac{P(X_i|p)}{P(X_1 \dots X_n)} = \frac{C_{\alpha,\beta} p^{\alpha-1}(1-p)^{\beta-1} \prod_{i=1}^n p^{X_i}(1-p)^{(1-X_i)}}{P(X_1 \dots X_n)}$$

where $C_{\alpha,\beta} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the normalizing constant for the $B(\alpha, \beta)$ distribution. The above expression can be written as:

$$P(p|X_1 \dots X_n) = \frac{C_{\alpha,\beta} p^{\alpha-1}(1-p)^{\beta-1} p^s (1-p)^{(n-s)}}{P(X_1 \dots X_n)} = C\ p^{s+\alpha-1}(1-p)^{n-s+\beta-1}$$

where $s = \sum_{i=1}^n X_i$ is the number of sucesses and $C$ is the normalizing constant for the posterior.

Note that from the form of the posterior we already know it is a Beta distribution $B(s+\alpha, n-s+\beta)$ and the normalizing constant $C$ is given by

$$C = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(s+\alpha)\Gamma(n-s+\beta)}.$$

The posterior mean estimate $\hat{p}$, therefore, is:

$$\hat{p} = E\ B(s+\alpha, n-s+\beta) = \frac{s+\alpha}{n+\alpha+\beta} \tag{2.4}$$

Recall that the ML estimate $\hat{p}_{ML}$ was:

$$\hat{p}_{ML} = \frac{s}{n}$$

The posterior estimate in (2.4) and the maximum likelihood estimate are the same asymptotically. However, for small sample sizes (2.4) has a smoothing effect. It disallows zero probability inferences when the sucess count is zero, and enforces the influence of a prior estimate. For $\alpha = \beta = 2$, the posterior mean is $\hat{p} = \frac{s+2}{n+4}$ which is the so called Wilson's estimate of $p$.

## Conjugate Priors and the Normal Density

Consider observations $X_1, \ldots, X_n$ i.i.d $N(\mu, \sigma)$ where we assume $\sigma$ to be known, and $\mu$ to be the only unknown parameter.

$$P(X_1 \ldots X_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ -\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2} \right\}$$

Assume a Gaussian prior $N(\mu_0, \sigma_0)$ on the mean i.e our prior belief is to see the mean $\mu$ around some value $\mu_0$ with variance $\sigma_0$ distributed normally:

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} exp^{-\frac{(\mu-\mu_0)^2}{2{\sigma_0}^2}}$$

The posterior has the following form:

$$P(\mu|X_1 \ldots X_n) = C\ P(\mu)P(X_1 \ldots X_n|\mu) = C\ \frac{1}{\sqrt{2\pi\sigma_0^2}} \frac{1}{\sqrt{2\pi\sigma^{2n}}}\ \exp\left\{ -\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} \right\}$$

$$= C'\ \exp\left\{ -\frac{\left[\mu - \frac{(\frac{\sum_{i=1}^{n} X_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})}{\frac{n}{\sigma^2}+\frac{1}{\sigma^2}}\right]^2}{2\frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}} \right\}$$

where $C, C'$ are appropriate normalization constants. From the last expression it follows

that the posterior is also normal with mean $\mu_{post}$ and variance $\sigma^2_{post}$ given by:

$$\mu_{post} = \frac{(\frac{\sum_{i=1}^n X_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}\left(\frac{1}{\sigma^2}\sum_{i=1}^n X_i + \frac{\mu_0}{\sigma_0^2}\right)$$

$$\frac{1}{\sigma^2_{post}} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \implies \sigma^2_{post} = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

Recall that the maximum likelihood estimate of the mean was $\mu_{ML} = \frac{\sum_{i=1}^n X_i}{n}$. The expression for $\mu_{post}$ above can be written as:

$$\mu_{post} = \frac{\sigma_0^2 n}{n\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma_2}{n\sigma_0^2 + \sigma^2}\mu_0$$

Thus in both examples the posterior mean is a weighted average of the sample mean (Maximum likelihood estimate) and the prior mean. Asymptotically, the posterior mean and the sample mean are identical. In the small sample case, the prior belief can strongly influence the choice of $\mu$ in the manner expressed above.

The posterior mean in the multivariate case has the same form. For estimation of the covariance of a multivariate normal it is also possible to define a conjugate prior - the inverse Wishart distribution on positive definite matrices. We omit the precise form of the distribution. For our purposes it suffices to note that the distribution depends on a 'central' covariance $C_0$ and a concentration parameter $a$, and prefers covariances close to $C_0$. The final posterior mean again has the form of a weighted average of the empirical covariance matrix and $C_0$:

$$C_{post} = \frac{n\hat{C} + aC_0}{n + a},$$

where $\hat{C}$ is the empirical covariance which is the maximum likelihood estimate (see lecture 1).