**Machine Learning Course - CS-433**

# Exponential Families and Generalized Linear Models

Oct 30th, 2018

©Mohammad Emtiyaz Khan 2015

changes by Rüdiger Urbanke 2016

changes by Rüdiger Urbanke 2017

changes by Rüdiger Urbanke 2018

Last updated: October 31, 2018

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

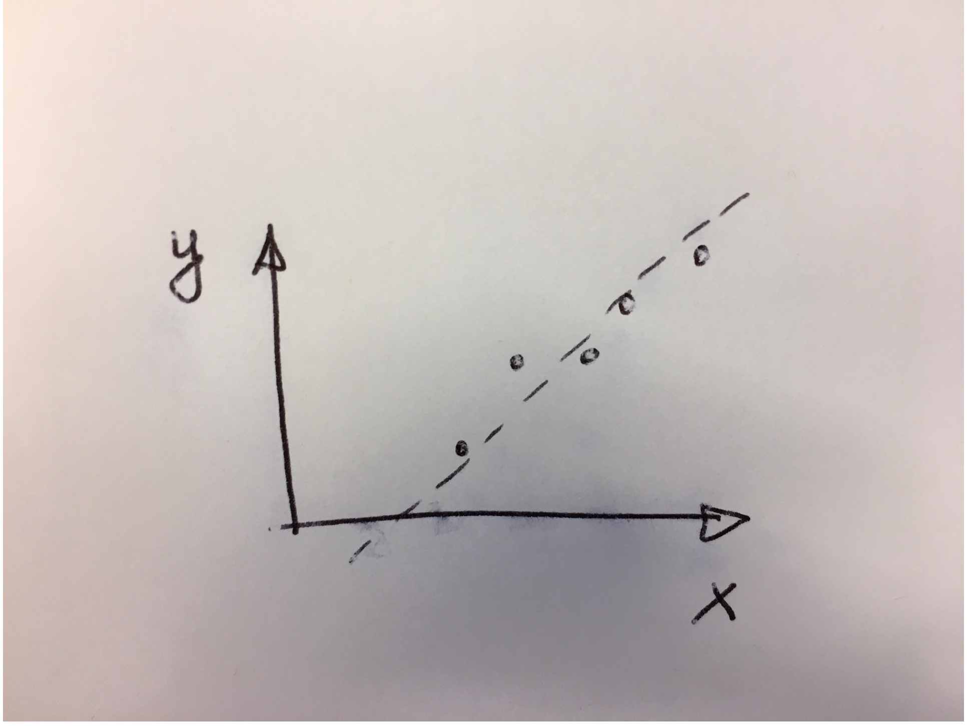

Figure 1:

## Motivation

Let us go back to regression. Consider the very simple one-dimensional example in Fig. 1. The horizontal axis represents the input $x$ and the vertical axis the output $y$. Our aim is to find a model for this data. It is very natural in this case that we try a linear model: $y = xw_1 + w_0 + Z$. I.e., we model the data as a line plus noise. Perhaps the most natural choice for the noise is a zero-mean Gaussian with some variance $\sigma^2$. As we discussed, this leads to least squares, assuming that we think of the data samples as independent and that we maximize the likelihood. This is what is typically meant when people talk about linear models (of course the data could be higher dimensional).

Now consider the data given in Fig. 2. In this case a linear

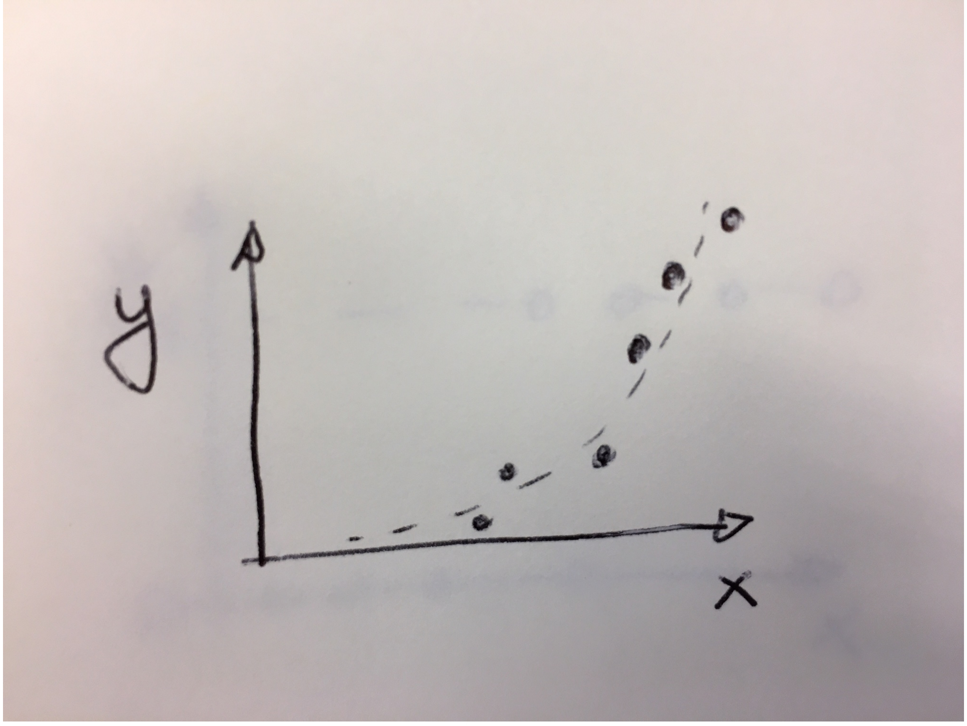

Figure 2:

model would not be a good fit. We have seen how we can get around this problem. Just add some additional features, e.g., $x^2$ and $x^3$. If we now use again a linear model, but in the extended feature space then we should be able to model the data well. So the idea was to augment or transform the feature space.

But this is not the only option we have. Note that in the example above the linear model predicts the *mean* of a distribution from which we then assume the data was sampled. Explicitly, we had $y = xw_1 + w_0 + Z$, where $xw_1 + w_0$ is the prediction of the linear model and represents the mean (i.e., the putatively "true" value for this data point) and then we get a noisy version as a sample. Here is now the extra degree of freedom we have: Instead of using the linear model to predict the mean of the distribution we can use it to predict

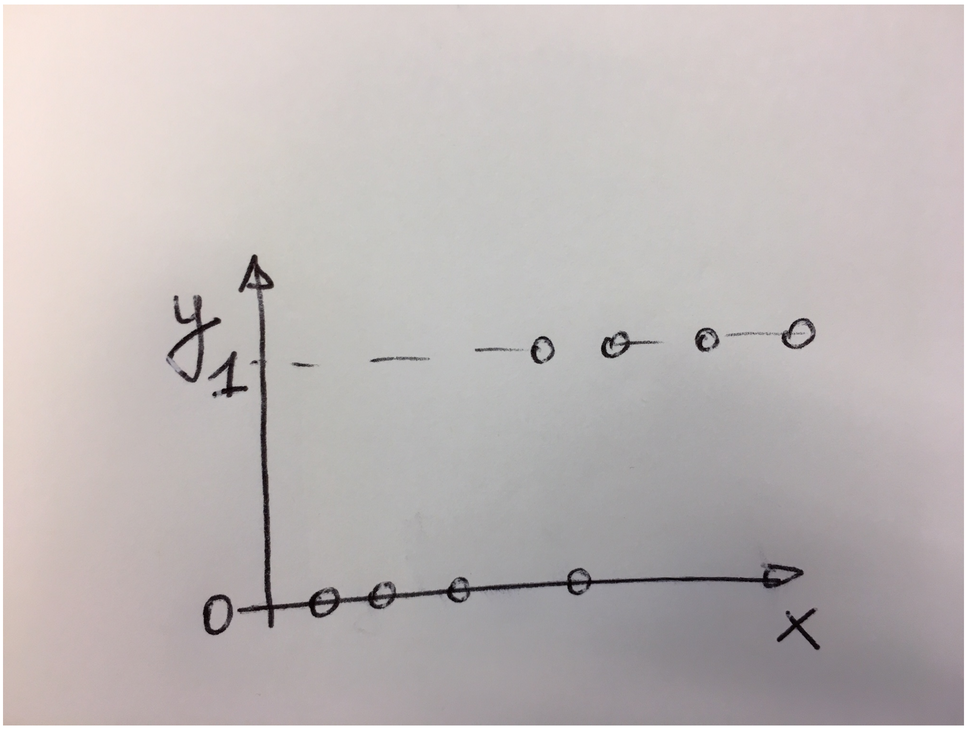

Figure 3:

a different quantity.

We have already seen an example when we talked about logistic regression. Consider the data given in Fig 3, where all the $y$ values are in $\{0, 1\}$. This might correspond to a binary classification problem. Recall that in logistic regression we model the probability of the two classes $\{0, 1\}$ given the data $\mathbf{x}$ by

$$
\begin{aligned}
p(y = 1|\eta) &= \sigma(\eta y), \\
p(y = 0|\eta) &= 1 - \sigma(\eta y),
\end{aligned}
$$

where $\eta$ as a shorthand for $\mathbf{x}^\top \mathbf{w}$. This can be written compactly as

$$
p(y|\eta) = \frac{e^{\eta y}}{1 + e^{\eta}} = \exp\left[\eta y - \log(1 + e^{\eta})\right],
$$

where $y \in \{0, 1\}$. Note that linear model predicts $\eta$, $\eta = \mathbf{x}^\top \mathbf{w}$, and that $\eta$ is *not* the mean of the distribution. Rather, $\eta$ is related to the mean $\mu$ by the non-linear relation $\eta = \ln \frac{\mu}{1-\mu}$ or $\mu = \sigma(\eta)$. This relation between the parameter we predict by the linear model and the mean is called the *link function.* It is exactly this nonlinear link function that makes it possible to use a linear model in this context.

## Outline

As you can see, we rewrote this distribution used in logistic regression in a very specific form. Our aim for today will be to generalize this form. We will see that there are many other distributions that can be written in this form. This will lead us to the class of distributions known as exponential families. We will first spend some time to talk about this family. We will see that many distributions (but not all) fit into this framework and that distributions in this family have many nice properties. We will only discuss some of these properties. Exponential families are also those distributions that have maximum entropy given some moment constraints and they are extremal also in other contexts. You are likely going to come across this family in other courses, and you will definitely see them if later on you will work in this area. As a second step we then discuss how exponential families can be used in the context of ML. In essence, by using different families we use different link functions, i.e., different relationships between the paramter $\eta$ that the linear model predicts and the mean of the distribution. And this degree of freedom can be useful when we are trying to find a good model for a given set of data.

In the subsequent discussion we consider various exponential families and then compute the corresponding link functions. But conceptually it can also be fruitful to think in the reverse way. What should the relationship be between the parameter that the linear model predicts and the mean of the distribution in order to fit the data well. I.e., perhaps we start with a desired link function and then find the exponential family that gives us this relationship.

## Exponential family – Definition

Let $y$ be a scalar and $\boldsymbol{\eta}$ be a vector. We will say that a distribution belongs to the *exponential family* if it can be written in the form

$$p(y|\boldsymbol{\eta}) = h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right]. \tag{1}$$

Let us look at the various components of this distribution.

The quantity $\boldsymbol{\phi}(y)$ is in general a vector and it is called a *sufficient statistics*. Why is $\boldsymbol{\phi}(y)$ called a sufficient statistics? Assume that we are given independent samples from this distribution. We do know $\boldsymbol{\phi}(y)$ and $h(y)$ but we do not know the parameter $\eta$. It turns out that in order to optimally estimate $\boldsymbol{\eta}$ given these samples all we need is the empirical average of the $\boldsymbol{\phi}(\mathbf{y})$. In other words, $\boldsymbol{\phi}(\mathbf{y})$ contains all the relevant information.

Note that the expression in (1) is non-negative if $h(y) \geq 0$. So we only need to ensure that it is properly normalized, i.e., we require that

$$\int_y h(y) \exp \left[ \boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta}) \right] dy] = 1.$$

Rewriting this we see that

$$\int_y h(y) \exp \left[ \boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right] dy] = e^{A(\boldsymbol{\eta})}. \tag{2}$$

We see from the last expression that the only role of $A(\boldsymbol{\eta})$ is to ensure a proper normalization. $A(\boldsymbol{\eta})$ is sometimes called the *cumulant* and some times it is called the *log partition* function. We will see shortly that despite the fact that $A(\boldsymbol{\eta})$ is *only* there for normalization purposes it plays a crucial role and contains valuable information.

If you look at the definition of the exponential family, you will see that we have several "degrees of freedom" to define an element of the family. We can choose the factor $h(y)$, we can choose the vector $\boldsymbol{\phi}(y)$, and we can choose the parameter $\boldsymbol{\eta}$. For every choice we will get an element of the exponential family. The term $A(\boldsymbol{\eta})$ is then determined for each such choice and ensures that the expression is properly normalized as dicussed. Of course it can happen that for some parameters $\boldsymbol{\eta}$, $h(y) \exp \left[ \boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right]$ is such that we cannot normalize the expression because the integral is infinity. E.g., set $h(y) = 1$, $\boldsymbol{\phi}(y) = y^2$ and $\boldsymbol{\eta} = 1$. We will exclude such parameters by only looking at the set of parameters

$$M := \{\boldsymbol{\eta} : \int_y h(y) \exp \left[ \boldsymbol{\eta}^\top \boldsymbol{\phi}(y) \right] dy] < \infty\}.$$

As a final remark concerning $A(\boldsymbol{\eta})$ note that from (2) we have

$$A(\boldsymbol{\eta}) = \ln\Big[ \int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy \Big]. \tag{3}$$

## Exponential family – Examples

Let us look at a few examples which are probably familiar to you but you might not have seen them written in this form. *Example:* We claim that the Bernoulli distribution is a member of the exponential family. We write

$$\begin{aligned} p(y|\mu) &= \mu^y (1-\mu)^{1-y}, \text{ where } \mu \in (0,1) \\ &= \exp\left[ (\ln \frac{\mu}{1-\mu}) y + \ln(1-\mu) \right] \\ &= \exp\left[ \eta \phi(y) - A(\eta) \right]. \end{aligned}$$

Mapping this to (1) we see that

$$\begin{aligned} \phi(y) &= y, \\ \eta &= \ln \frac{\mu}{1-\mu}, \\ A(\eta) &= -\ln(1-\mu) = \ln(1+e^\eta), \\ h(y) &= 1. \end{aligned}$$

In this case $\phi(y)$ is a scalar, reflecting the fact that this family only depends on a single parameter. In fact, we have a 1-1 relationship between $\eta$ and $\mu$,

$$\eta = g(\mu) = \ln \frac{\mu}{1-\mu} \iff \mu = g^{-1}(\eta) = \frac{e^\eta}{1+e^\eta}.$$

As we mentioned in the very beginning, this function $g$ is known as the *link* function (it links the mean of $\boldsymbol{\phi}(y)$ to the parameter $\eta$.)

Note that this is *exactly* the same distribution that we encountered when we discussed *logistic regression.*

*Example:* Consider the Poisson distribution with mean $\mu$. We have, for $y \in \mathbb{N}$,

$$\begin{aligned} p(y|\mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \frac{1}{y!} e^{y \ln(\mu) - \mu} \\ &= h(y) e^{\eta\phi(y) - A(\eta)}, \end{aligned}$$

where $h(y) = 1/y!$, $\phi(y) = y$, $\eta = g(\mu) = \ln(\mu)$, and $\mu = g^{-1}(\eta) = e^{\eta}$. Here again, $g(\mu)$ *links* the mean to the parameter $\eta$.

*Example:* The Gaussian distribution with mean $\mu$ and variance $\sigma^2$ as parameters is also a member of the exponential family. We write

$$\begin{aligned} p(y|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \\ &= \exp\left[(\mu/\sigma^2, -1/(2\sigma^2))(y, y^2)^\top - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]. \end{aligned}$$

Mapping this again to (1) we see that

$$\begin{aligned}
\boldsymbol{\phi}(y) &= (y, y^2)^\top \\
\boldsymbol{\eta} &= (\eta_1 = \mu/\sigma^2, \eta_2 = -1/(2\sigma^2))^\top, \\
A(\boldsymbol{\eta}) &= \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2), \\
&= -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi), \\
h(y) &= 1.
\end{aligned}$$

Note that this time $\boldsymbol{\phi}(y)$ is a vector of length two, reflecting the fact that the distribution depends on two parameters. In fact, we have the 1-1 relationship between $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and $(\mu, \sigma^2)$.

$$\eta_1 = \frac{\mu}{\sigma^2}; \eta_2 = -\frac{1}{2\sigma^2} \iff \mu = -\frac{\eta_1}{2\eta_2}; \sigma^2 = -\frac{1}{2\eta_2}.$$

# Basic Properties

## Convexity of $A(\boldsymbol{\eta})$

**Lemma.** *The cumulant $A(\boldsymbol{\eta})$ is convex as a function of $\boldsymbol{\eta}$ on $M$ (the set of parameters $\boldsymbol{\eta}$ where the cumulant is finite).*

*Proof.* Let $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ be two parameters in $M$. Define $\boldsymbol{\eta} = \lambda\boldsymbol{\eta}_1 + (1-\lambda)\boldsymbol{\eta}_2$. We start with (2) and apply Hoelder's inequality. Recall that Hoelder's inequality reads $\|fg\|_1 \leq \|f\|_p\|g\|_q$, where $p, q \in [1, \infty]$ and $1/p + 1/q = 1$. Here,

$$\|f\|_p = \Big(\int |f(y)|^p dy\Big)^{\frac{1}{p}}.$$

You might not have seen Hoelder's inequality before, but you surely have seen the special case when $p = q = 2$. In this case you get the Cauchy-Schwarz inequality.

Let us go back to the proof. Pick $p = 1/\lambda$ and $q = 1/(1-\lambda)$. Then $p, q \in [1, \infty]$ and $1/p + 1/q = \lambda + (1 - \lambda) = 1$. We have

$$\begin{aligned}
&e^{A(\boldsymbol{\eta})}\\
=&\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy\\
=&\int_y \underbrace{\left[h(y)^\lambda \exp\left[\lambda \boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right]\right]}_{f(y)} \underbrace{\left[h(y)^{1-\lambda} \exp\left[(1-\lambda)\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right]\right]}_{g(y)} dy\\
\leq&(\int_y h(y) \exp\left[\boldsymbol{\eta}_1^\top \boldsymbol{\phi}(y)\right] dy)^\lambda (\int_y h(y) \exp\left[\boldsymbol{\eta}_2^\top \boldsymbol{\phi}(y)\right] dy)^{1-\lambda}\\
=&e^{\lambda A(\boldsymbol{\eta}_1)} e^{(1-\lambda) A(\boldsymbol{\eta}_2)}.
\end{aligned}$$

Taking the log of this chain proves the claim,

$$A(\boldsymbol{\eta}) \leq \lambda A(\boldsymbol{\eta}_1) + (1 - \lambda) A(\boldsymbol{\eta}_2).$$

□

## Derivatives of $A(\boldsymbol{\eta})$ and moments

Another useful property is that the gradient and Hessian (first and second derivatives) of $A(\boldsymbol{\eta})$ are related to the mean and the variance of $\boldsymbol{\phi}(y)$.

**Lemma.**

$$\begin{aligned}
\nabla A(\boldsymbol{\eta}) &= \mathbb{E}[\boldsymbol{\phi}(y)],\\
\nabla^2 A(\boldsymbol{\eta}) &= \mathbb{E}[\boldsymbol{\phi}(y)\boldsymbol{\phi}(y)^\top] - \mathbb{E}[\boldsymbol{\phi}(y)]\mathbb{E}[\boldsymbol{\phi}(y)]^\top.
\end{aligned}$$

Note that this in particular shows that the Hessian of $A(\boldsymbol{\eta})$ is a covariance matrix and hence is positive semi-definite. This gives us a second proof that $A(\boldsymbol{\eta})$ is convex.

Before we prove this, let us check this for our two running examples. Recall that for the Bernoulli distribution $\boldsymbol{\phi}(y)$ is a scalar, namely $y$. So in this case the first derivative should be the mean of the Bernoulli distribution and the second derivative the variance. Let us verify this. We get

$$\begin{aligned} \frac{dA(\eta)}{d\eta} &= \frac{d\ln(1+e^{\eta})}{d\eta} = \frac{e^{\eta}}{1+e^{\eta}} = \sigma(\eta) = \mu, \\ \frac{d^2A(\eta)}{d\eta^2} &= \frac{d\sigma(\eta)}{d\eta} = \sigma(\eta)(1-\sigma(\eta)) = \mu(1-\mu), \end{aligned}$$

which confirms the claim.

For the Gaussian distribution our vector $\boldsymbol{\phi}(y)$ is of the form $(y, y^2)^\top$. So the first derivative (gradient) should give us the mean and the second moment of the Gaussian. The second derivative should give us the variance of various moments of $y$. We get

$$\begin{aligned} \frac{\partial A(\boldsymbol{\eta})}{d\eta_1} &= \frac{\partial(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi))}{\partial\eta_1} = -\frac{\eta_1}{2\eta_2} = \mu, \\ \frac{\partial A(\boldsymbol{\eta})}{d\eta_2} &= \frac{\partial(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-\eta_2/\pi))}{\partial\eta_2} = (\frac{\eta_1^2 - 2\eta_2}{4\eta_2^2}) = \mu^2 + \sigma^2, \end{aligned}$$

which are exactly the expected value and the second moment of $y$, as claimed. To do one more computation, let us compute

$$\frac{\partial^2 A(\boldsymbol{\eta})}{d\eta_1^2} = \frac{\partial(-\frac{\eta_1}{2\eta_2})}{\partial\eta_1} = -\frac{1}{2\eta_2} = \sigma^2,$$

which is the variance of $y$, again as expected.

*Proof.* Let us just write down the proof regarding the first derivative. The proof for the second derivative proceeds in a similar fashion. We have

$$\begin{aligned}\nabla A(\boldsymbol{\eta}) &= \nabla \ln[\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy] \\ &= \frac{\int_y \nabla h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy}{\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] dy} \\ &= \frac{\int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y)\right] \boldsymbol{\phi}(y) dy}{\exp(A(\boldsymbol{\eta}))} \\ &= \int_y h(y) \exp\left[\boldsymbol{\eta}^\top \boldsymbol{\phi}(y) - A(\boldsymbol{\eta})\right] \boldsymbol{\phi}(y) dy \\ &= \mathbb{E}[\boldsymbol{\phi}(y)].\end{aligned}$$

In the second step we have exchange the derivative with the integral. Note that the exchange of differentiation and integration is permitted if the resulting integral is finite (which it is in our case).

□

## Link function

As we have seen already in two specific cases (Bernoulli and Poisson), there is a relationship between the "mean" $\boldsymbol{\mu} :=$

$\mathbb{E}[\boldsymbol{\phi}(y)]$ and $\boldsymbol{\eta}$ defined using a so-called *link function* $\mathbf{g}$.

$$\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\mu}) \iff \boldsymbol{\mu} = \mathbf{g}^{-1}(\boldsymbol{\eta}).$$

For the Gaussian, we started with the "natural" parameters $(\mu, \sigma^2)$ and we have seen that there is a 1-1 relationship to the vector $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. But we could have started with the parameters $(\mu, \mu^2+\sigma^2)$ (which now corresponds to $\mathbb{E}[\boldsymbol{\phi}(y)] = \mathbb{E}[(y, y^2)^\top]$ instead). And again we would have found that there is a 1-1 relationship between $\mathbb{E}[\boldsymbol{\phi}(y)]$ and the vector $\boldsymbol{\eta}$. For a list of such link functions for various distributions see the chapter on "Generalized Linear Model" in the KPM book.

## Applications in ML

Let us now look at two applications of exponential families in ML.

### Maximum Likelihood Parameter Estimation

Assume that we have a set of samples $\{y_n\}_{n=1}^N$ We assume that these are independent samples from some distribution. Further, we assume that they come from some exponential family with a given $h(y)$ and sufficient statistics $\boldsymbol{\phi}(y)$ but unknown parameter $\boldsymbol{\eta}$ (or we simply want to find that element of this family of distributions that is closest). Our aim is to estimate the parameter $\boldsymbol{\eta}$. We use our maximum likelihood principle to find this parameter. Hence we minimize

$$\begin{aligned} L(\boldsymbol{\eta}) &= -\ln(p(\mathbf{y}|\boldsymbol{\eta})) \\ &= \sum_{n=1}^{N}[-\ln(h(y_n) - \boldsymbol{\eta}^\top \boldsymbol{\phi}(y_n) + A(\boldsymbol{\eta})]. \end{aligned}$$

We see that this is a convex function in $\boldsymbol{\eta}$ since $A(\boldsymbol{\eta})$ is a convex function. Further, if we assume that we can determine the link function we can derive the solution in an explicit form by taking the gradient and setting it to zero:

$$\frac{1}{N}\nabla L(\boldsymbol{\eta}) = -(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n)) + \mathbb{E}[\boldsymbol{\phi}(y)] = 0.$$

We get

$$\boldsymbol{\eta} = \mathbf{g}^{-1}(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}(y_n)).$$

We now see the justification for why we called $\boldsymbol{\phi}(y)$ a sufficient statistics.

## Generalized Linear Models

Given an element from the exponential family with a scalar $\phi(y)$, we can construct from this a data model by assuming that a sample $(\mathbf{x}, y)$ follows the distribution

$$p(y \mid \mathbf{x}, \mathbf{w}) = h(y)e^{\mathbf{x}^\top \mathbf{w}\phi(y) - A(\mathbf{x}^\top \mathbf{w})}.$$

We call such a model a *generalized linear model.* It is a generalization of the data model we used for logistic regression.

As we will now discuss, for such a model the maximum likelihood problem is particularly easy to solve. Assume that we have given a training set $S_{\text{train}}$ consisting of $N$ independent samples $(\mathbf{x}_n, y_n)$. Assume further that we fit a generalized linear model to this data. This means that we assume that samples obey a distribution of the form

$$p(y_n \mid \mathbf{x}_n, \mathbf{w}) = h(y_n) e^{\eta_n \phi(y_n) - A(\eta_n)}$$

with $\eta_n = \mathbf{x}_n^\top \mathbf{w}$. Given $S_{\text{train}}$, we then write down the likelihood and look for that weight vector $\mathbf{w}$ that maximizes this likelihood.

In more detail, we consider the cost function

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= -\sum_{n=1}^{N} \ln p(y_n | \mathbf{x}_n^\top \mathbf{w}) \\ &= -\sum_{n=1}^{N} \ln(h(y_n)) + \mathbf{x}_n^\top \mathbf{w} \phi(y_n) - A(\mathbf{x}_n^\top \mathbf{w}).\end{aligned}$$

We want to minimize this cost function (we added a minus sign). First, note that this cost function is convex, hence a greedy algorithm should work well.

Let us take the gradient of this expression,

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = -\sum_{n=1}^{N} \mathbf{x}_n \phi(y_n) - \mathbf{x}_n A(\mathbf{x}_n^\top \mathbf{w}).$$

Recall that

$$\frac{dA(\eta)}{d\eta} = \mathbb{E}[\phi(y)] = g^{-1}(\eta).$$

Hence, we get

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = -\sum_{n=1}^{N} \mathbf{x}_n\phi(y_n) - \mathbf{x}_n g^{-1}(\mathbf{x}_n^\top \mathbf{w}).$$

If we set this equation to zero we get the condition of optimality. In particular, if we rewrite this sum by using our matrix notation we get

$$\nabla\mathcal{L}(\mathbf{w}) = \mathbf{X}^\top \left[g^{-1}(\mathbf{X}\mathbf{w}) - \phi(\mathbf{y})\right] = 0,$$

where, as before, the scalar functions ($g^{-1}$ and $\phi$) are applied to each vector component-wise.
To compare, for the case of the logistic regression we got the equation

$$\nabla\mathcal{L}(\mathbf{w}) = \mathbf{X}^\top \left[\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y}\right] = 0.$$

As we have discussed, for the logistic case (Bernoulli distribution) we have the relationship $g^{-1} = \sigma$, which confirms that our previous derivation was just a special case.
Note also that we have already shown that $A(\mathbf{x}^\top\mathbf{w})$ is a convex function ($A$ is convex and $A(\mathbf{x}^\top\mathbf{w})$ is the composition of a linear function with a convex function). Therefore $\mathcal{L}(\mathbf{w})$ is convex (the other terms are constant or linear), just as we have seen this for the logistic regression. As a consequence, greedy iterative algorithms (like gradient descent) to find the optimum weight vector $\mathbf{w}$ are expected to work well in this context.