# Machine Learning for Engineers: Exponential Family of Distributions

Isaac Osei Nyantakyi,PH.D

# This Chapter

- A rare example of a distribution that does not belong to this class is given by a uniform distribution in an interval dependent on model parameters.
- We will discover that all distributions in the exponential family share the useful properties mentioned above in terms of log-loss, ML learning, and information-theoretic measures.
- As a result, the methods studied in the previous chapters can be extended to a much larger class of problems by replacing Bernoulli, categorical, and Gaussian distributions with another model in the exponential family.

# Overview

- Exponential family: definitions and examples
- Gradient of the log-loss, or score vector
- ML learning
- Information-theoretic metrics
- Fisher information matrix
- Generalized linear models

# Exponential Family: Definitions and Examples

# Exponential Family

- Bernoulli, categorical, and Gaussian distributions have the useful property that the log-loss $-\log p(x|\eta)$ is a convex function of the model parameters $\eta$.
- Note that in this chapter we will introduce two different definitions for model parameters, which will be denoted as $\eta$ (natural parameters) and $\mu$ (mean parameters). Accordingly, we will not adopt the notation $\theta$ used thus far when we need to identify which type of model parameter is being considered.
- Ex.: For a Gaussian rv $\mathrm{x} \sim \mathcal{N}(\eta, 1)$, the log-loss is given as

$$-\log p(x|\eta) = \frac{1}{2}(x-\eta)^2 + \text{const. indep. of } \eta,$$

  which is quadratic, and hence convex in the parameter $\eta$.
- All distributions in the exponential family share this key property, which simplifies optimization (see Chapter 5), and they are defined as follows.

# Exponential Family

- A probabilistic model $p(x|\eta)$ in the exponential family is described by
  - a $K \times 1$ vector of *sufficient statistics*

$$s(x) = \left[ \begin{array}{c} s_1(x) \\ \vdots \\ s_K(x) \end{array} \right],$$

  - as well as by a "log-base measure" function $M(x)$.
- This is in the sense that the distribution $p(x|\eta)$ depends on $x$ only through vector $s(x)$ and function $M(x)$. One can hence think of $s(x)$ and $M(x)$ as defining a vector of features that determines the distribution of $x$.

# Exponential Family

- Specifically, the exponential family contains discrete and continuous distributions whose log-loss can be written as

$$-\log p(x|\eta) = -\underbrace{\eta^T s(x)}_{=\sum_{k=1}^{K}\eta_k s_k(x)} - \underbrace{M(x)}_{\text{log-base measure}} + \underbrace{A(\eta)}_{\text{log-partition function}},$$

where we have defined

  - the $K \times 1$ natural parameters vector $\eta = [\eta_1, ..., \eta_K]^T$,
  - and the log-partition function $A(\eta)$.

# Log-Partition Function

- The log-partition function $A(\eta)$ is fixed once functions $s(\cdot)$ and $M(\cdot)$ are specified, and is convex in $\eta$.
- To see this, we can write the distributions in the exponential family as

$$\begin{aligned} p(x|\eta) &= \exp\left(\eta^T s(x) - A(\eta) + M(x)\right) \\ &= \frac{1}{\exp(A(\eta))} \exp\left(\eta^T s(x) + M(x)\right). \end{aligned}$$

- In order to guarantee the normalization $\int p(x|\eta)dx = 1$ for continuous variables and $\sum_x p(x|\eta) = 1$ for discrete variables, we need to set

$$A(\eta) = \log \int \exp\left(\eta^T s(x) + M(x)\right) dx$$

  for continuous rvs and

$$A(\eta) = \log \sum_x \exp\left(\eta^T s(x) + M(x)\right)$$

  for discrete rvs.

# Log-Partition Function

- The log-partition function has a “log-sum-exp” form and is hence convex in $\eta$.
- Note that the function $\exp(A(\eta))$ is known as partition function – whence the name log-partition function for $A(\eta)$.
- The log-loss is the sum of a linear function in $\eta$, namely $-\eta^T s(x)$, and of a convex function in $\eta$, namely $A(\eta)$. This implies that the log-loss is convex in $\eta$.
- The vector of natural parameters $\eta$ can take any value that ensures that the distribution can be normalized, i.e., $A(\eta) < \infty$. This feasible set is convex (i.e., it contains all segments between any two points in the set) and is taken to be open, that is, not to include its boundary.
- (Technically, this defines the class of distributions in the *regular* exponential family, which we will focus on).

# Exponential Family

- To summarize, the exponential family contains discrete and continuous distributions that are specified by sufficient statistics $s(x)$ and log-base measure function $M(x)$ as

$$p(x|\eta) \propto \exp\left( \underbrace{\eta^T s(x)}_{\text{linear function of } \eta} + \underbrace{M(x)}_{\text{general function of } x} \right),$$

  where the "proportional to" sign $\propto$ makes the normalizing constant $\frac{1}{\exp(A(\eta))}$ implicit.

- When no confusion can arise, we will write a distribution in the exponential family as

$$p(x|\eta) = \mathrm{ExpFam}(x|\eta),$$

  where the notation hides the dependence on $s(x)$ and $M(x)$.

# Example 1: Gaussian Distribution with Fixed Variance

- For the Gaussian distribution $\mathcal{N}(\nu, \beta^{-1})$ with a fixed precision $\beta$, the log-loss can be written as

$$-\log \mathcal{N}(x|\nu, \beta^{-1}) = -\underbrace{\beta\nu}_{\eta}\underbrace{x}_{s(x)} - \left(\underbrace{-\frac{\beta}{2}x^2 - \frac{1}{2}\log(2\pi\beta^{-1})}_{M(x)}\right)$$

$$+ \left(\underbrace{\frac{\beta\nu^2}{2}}_{A(\eta)}\right).$$

# Example 1: Gaussian Distribution with Fixed Variance

- The log-partition function can be expressed in terms of the natural parameter $\eta = \beta\nu$ as

$$A(\eta) = \frac{\beta\nu^2}{2} = \frac{1}{2}\frac{\eta^2}{\beta}.$$

  - As expected, this is a (strictly) convex function of $\eta$ for all $\eta \in \mathbb{R}$.

- In the previous chapters, we would have parametrized the distribution $\mathcal{N}(\nu, \beta^{-1})$ through the mean parameter $\nu = \mathrm{E}_{\mathrm{x}\sim\mathcal{N}(\nu,\beta^{-1})}[s(\mathrm{x})] = \mathrm{E}_{\mathrm{x}\sim\mathcal{N}(\nu,\beta^{-1})}[\mathrm{x}]$.
  - We now have an alternative parametrization in terms of the natural parameter $\eta$.
  - The natural parameter $\eta$ and the mean parameter $\mu$ are in a one-to-one relationship as one can be recovered from the other through the equality $\nu = \eta/\beta$ (recall that $\beta$ is fixed and hence it should be considered as a numerical value).

# Example 2: Bernoulli Distribution

- The Bernoulli distribution can be written as

$$\mathrm{Bern}(x|\mu) = \mu^{x}(1-\mu)^{1-x},$$

where $x \in \{0,1\}$ and we have the mean parameter $\mu = \mathrm{E}_{\mathrm{x}\sim\mathrm{Bern}(x|\mu)}[\mathrm{x}] = \mathrm{Pr}[\mathrm{x} = 1]$.

- Therefore, the log-loss is

$$-\log \mathrm{Bern}(x|\mu) = -\underbrace{\log\left(\frac{\mu}{1-\mu}\right)}_{\eta}\underbrace{x}_{s(x)} + \underbrace{(-\log(1-\mu))}_{A(\eta)},$$

where $M(x) = 0$.

# Example 2: Bernoulli Distribution

- The Bernoulli distribution can be written as

$$\mathrm{Bern}(x|\mu) = \mu^{x}\,(1-\mu)^{1-x}\,,$$

  where $x \in \{0, 1\}$ and we have the mean parameter $\mu = \mathrm{E}_{\mathrm{x}\sim\mathrm{Bern}(x|\mu)}[\mathrm{x}] = \mathrm{Pr}[\mathrm{x} = 1]$.

- Therefore, the log-loss is

$$-\log\mathrm{Bern}(x|\mu) = -\underbrace{\log\left(\frac{\mu}{1-\mu}\right)}_{\eta}\underbrace{x}_{s(x)} + \underbrace{(-\log(1-\mu))}_{A(\eta)},$$

  where $M(x) = 0$.

# Example 2: Bernoulli Distribution

- It follows that the natural parameter is the logit or log-odds (see Chapter 6)
$$\eta = \log\left(\frac{\mathrm{Bern}(1|\mu)}{\mathrm{Bern}(0|\mu)}\right) = \log\left(\frac{\mu}{1-\mu}\right).$$
- The mean parameter $\mu$ is in a one-to-one relationship with the natural parameter $\eta$: inverting the equality above, we have
$$\mu = \sigma(\eta) = \frac{1}{1+e^{-\eta}}.$$
- Therefore, the log-partition function can be expressed in terms of the natural parameter $\eta$ as
$$A(\eta) = \log(1+e^{\eta}),$$
which is a (strictly) convex function of $\eta \in \mathbb{R}$.

# Example 3: General Gaussian Distribution

- For a Gaussian distribution $\mathcal{N}(\nu, \beta^{-1})$ with parameters $(\nu, \beta)$, we have the log-loss

$$-\log \mathcal{N}(x|\nu, \beta^{-1}) = -\left( \underbrace{\beta\nu}_{\eta_1} \underbrace{x}_{s_1(x)} + \left( \underbrace{-\frac{\beta}{2}}_{\eta_2} \right) \underbrace{x^2}_{s_2(x)} \right) + \left( \underbrace{\frac{\nu^2\beta}{2} + \frac{1}{2}\log(2\pi\beta^{-1})}_{A(\eta)} \right),$$

where $M(x) = 0$.

- Note that we now have a two-dimensional vector of sufficient statistics, i.e., $K = 2$, namely $s(x) = \left[\, x\, x^2 \,\right]^T$.

# Example 3: General Gaussian Distribution

- Following the previous examples, the two-dimensional vector of mean parameters is defined as the vector of averages of the sufficient statistics under the model, i.e.,

$$\mu = \left[ \begin{array}{c} \mathrm{E}_{\mathrm{x}\sim\mathcal{N}(\nu,\beta^{-1})}[s_1(\mathrm{x})] \\ \mathrm{E}_{\mathrm{x}\sim\mathcal{N}(\nu,\beta^{-1})}[s_2(\mathrm{x})] \end{array} \right] = \left[ \begin{array}{c} \mathrm{E}_{\mathrm{x}\sim\mathcal{N}(\nu,\beta^{-1})}[\mathrm{x}] \\ \mathrm{E}_{\mathrm{x}\sim\mathcal{N}(\nu,\beta^{-1})}[\mathrm{x}^2] \end{array} \right] = \left[ \begin{array}{c} \nu \\ \nu^2 + \beta^{-1} \end{array} \right].$$

- This vector is in a one-to-one correspondence with the vector of natural parameters

$$\eta = \left[ \begin{array}{c} \beta\nu \\ -\frac{\beta}{2} \end{array} \right],$$

and we can write the log-partition function as

$$\begin{aligned} A(\eta) &= \frac{\nu^2\beta}{2} + \frac{1}{2}\log(2\pi\beta^{-1}) \\ &= -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right), \end{aligned}$$

# Exponential Family

- This is a much larger family!
- In fact, any distribution that can be described by a finite-dimensional vector of parameters and whose support does not depend on the parameters is in the exponential family.
- Among others, apart from the mentioned distributions, it includes the following distributions:
  - discrete: binomial, negative binomial, geometric, Poisson;
  - continuous: lognormal, gamma, inverse gamma, chi-squared, exponential, beta, Dirichlet, Pareto, Laplace.

# Example 4: Poisson Distribution

- As an example of a distribution that we have not considered before consider the Poisson distribution, which is used extensively in fields as diverse as neuroscience and communication network design.
- The Poisson distribution can be written as

$$\mathrm{Poiss}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

  where $x \in \{0, 1, 2, ...\}$.

- The log-loss is

$$-\log \mathrm{Poiss}(x|\lambda) = -\underbrace{\log(\lambda)}_{\eta}\underbrace{x}_{s(x)} + \underbrace{\lambda}_{A(\eta)} + \underbrace{\log(x!)}_{M(x)}.$$

# Example 4: Poisson Distribution

- As an example of a distribution that we have not considered before consider the Poisson distribution, which is used extensively in fields as diverse as neuroscience and communication network design.
- The Poisson distribution can be written as

$$\mathrm{Poiss}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

where $x \in \{0, 1, 2, ...\}$.

- The log-loss is

$$-\log \mathrm{Poiss}(x|\lambda) = -\underbrace{\log(\lambda)}_{\eta}\underbrace{x}_{s(x)} + \underbrace{\lambda}_{A(\eta)} + \underbrace{\log(x!)}_{M(x)}.$$

# Example 4: Poisson Distribution

- The mean parameter $\mu = \mathrm{E}_{\mathrm{x} \sim \mathrm{Poiss}(x|\lambda)}[x] = \lambda$ is in a one-to-one relationship with the natural parameter $\eta = \log(\lambda)$.
- Therefore, the log-partition function can be expressed in terms of the natural parameter $\eta$ as

$$A(\eta) = \exp(\eta),$$

which is a (strictly) convex function of $\eta \in \mathbb{R}$.

# Natural vs Mean Parameters

- Generalizing the examples above, distributions in the exponential family can be specified by the vector $\eta$ of natural parameters or by the vector $\mu$ of *mean parameters*

$$\mu = \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}[s(\mathrm{x})],$$

  which is the vector of averages of the sufficient statistics.

- Therefore, we can write a distribution in the exponential family as a function of $\eta$ as $p(x|\eta) = \mathrm{ExpFam}(x|\eta)$ or as a function of $\mu$ as $p(x|\mu) = \mathrm{ExpFam}(x|\mu)$.
- We have used, and we will use, both notations, hence overloading the notation $p(x|\cdot)$.
- As we will see below, there may be multiple natural parameter vectors $\eta$ yielding the same distribution and hence the same mean parameters $\mu$.

# Minimal Exponential Family

- A class of distributions $p(x|\eta) = \mathrm{ExpFam}(x|\eta)$ in the exponential family is said to be *minimal* if no two natural parameter vectors yield the same distribution, or more precisely if there is no $\eta$ in the domain for which $\eta^T s(x)$ is constant.
    - If there were such a value of $\eta$, then we could add it to any other value of $\eta$ without changing the distribution.
- For minimal classes of distributions, there is a one-to-one correspondence between natural and mean parameters:
    - In this case, each natural parameter vector yields a different distribution $p(x|\eta)$;
    - each mean parameter vector $\mu$ is associated with a single natural parameter vector $\eta$ (the vice versa is always true);
    - and there exists a general explicit expression for the log-loss as a function of the mean parameters (see Appendix).
- For minimal classes of distributions, the log-partition function is strictly convex, and hence we have $\nabla^2 A(\eta) \succ 0$.

# Example 5: Categorical (or Multinoulli) Distribution

- Not all classes of distributions in the exponential family are minimal.
- Consider the categorical distribution for a rv $\mathrm{x}$ that can take $C$ values $\{0, 1, .., , C-1\}$.
- The distribution can be written as

$$\mathrm{Cat}(x|\mu) = \prod_{k=0}^{C-1} \mu_k^{\mathbb{1}(x=k)} = \frac{1}{a} \prod_{k=0}^{C-1} (a\mu_k)^{\mathbb{1}(x=k)},$$

with any $a > 0$ and probabilities

$$\mu_k = \mathrm{Pr}[\mathrm{x} = k].$$

# Example 5: Categorical (or Multinoulli) Distribution

- The log-loss is

$$-\log(\mathrm{Cat}(x|\mu)) = -\sum_{k=1}^{C-1} \underbrace{\log(a\mu_k)}_{\eta_k} \underbrace{\mathbb{1}(x=k)}_{s_k(x)} + \underbrace{(\log a)}_{A(\eta)=\log\left(\sum_{k=0}^{C-1} e^{\eta_k}\right)} .$$

- Therefore, we have
  - sufficient statistics $s(x) = [\mathbb{1}(x=0), \ldots, \mathbb{1}(x=C-1)]^T$, which is the $C \times 1$ one-hot vector $x^{OH}$ (see Chapter 6);
  - natural parameters $\eta = [\eta_0, ..., \eta_{C-1}]^T$, which will be seen below to correspond to the logits discussed in Chapter 6;
  - mean parameter: $\mu = [\mu_0, ..., \mu_{C-1}]^T$ with $\mu_k = \mathrm{E}_{\mathrm{x}\sim\mathrm{Cat}(x|\mu)}[s_k(\mathrm{x})] = \Pr[\mathrm{x}=k]$.

# Example 5: Categorical (or Multinoulli) Distribution

- This parameterization is not minimal since we can always add a constant vector to the logit vector $\eta$ without changing the distribution:
  - the parameter $a > 0$ is arbitrary in the parametrization of the logits $\eta_k = \log(a\mu_k)$.

# Example 5: Categorical (or Multinoulli) Distribution

- Even for non-minimal families, there is a single mean parameter vector for each natural parameter vector $\eta$.
- In the case of a categorical distribution, the relationship is given by the softmax function (see Chapter 6)

$$\mu = \mathrm{softmax}(\eta) = \begin{bmatrix} \frac{e^{\eta_0}}{\sum_{k=0}^{C-1} e^{\eta_k}} \\ \vdots \\ \frac{e^{\eta_{C-1}}}{\sum_{k=0}^{C-1} e^{\eta_k}} \end{bmatrix}.$$

- But there are infinitely many logit vectors associated to each mean vector namely

$$\eta = \begin{bmatrix} \log(\mu_0) \\ \vdots \\ \log(\mu_{C-1}) \end{bmatrix} + b \times 1_C,$$

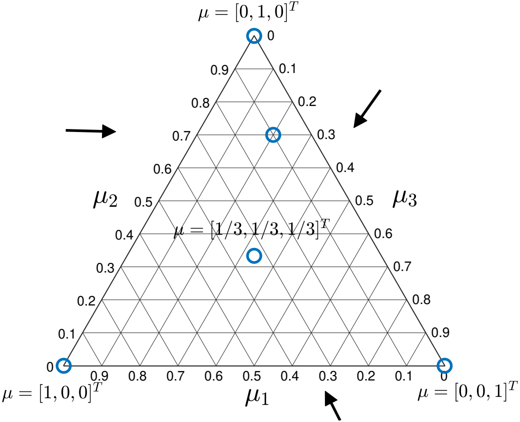
where $b$ is arbitrary.

# Example 5: Categorical (or Multinoulli) Distribution

- Example with $C = 3$ :
  - $\mathrm{x} \sim \mathrm{Cat}(x|[0.1, 0.8, 0.1]^T)$
  - mean parameters: $\mu_0 = \Pr[\mathrm{x} = 0] = 0.1$, $\mu_1 = \Pr[\mathrm{x} = 1] = 0.8$, $\mu_2 = \Pr[\mathrm{x} = 2] = 0.1$
  - natural parameters (logits) with $a = 1$: $\eta_0 = \eta_2 = \log(a \cdot 0.1) = -2.30$ and $\eta_1 = \log(a \cdot 0.8) = -0.22$.
  - we have the equality
    $$\mathrm{softmax}\left(\begin{bmatrix}-2.30\\-0.22\\-2.30\end{bmatrix}\right) = \frac{1}{\sum_{k=0}^{2} e^{\eta_k}}\begin{bmatrix}e^{\eta_1}\\e^{\eta_2}\\e^{\eta_3}\end{bmatrix} = \begin{bmatrix}0.1\\0.8\\0.1\end{bmatrix} = \mu.$$

# Example 5: Categorical (or Multinoulli) Distribution

- Categorical distributions can be conveniently represented on a simplex (or ternary) plot for $C = 3$ (arrows represent reading directions).



# Example 6: Joint Bernoulli-Gaussian Distribution

- Consider the rv $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2]^T$ with $\mathrm{x}_1 \sim \mathrm{Bern}(x|\mu)$ and $\mathrm{x}_2 \sim \mathcal{N}(\nu_{\mathrm{x}_1}, \beta^{-1})$, where the precision $\beta$ is fixed and the model parameters are $(\mu, \nu_0, \nu_1)$. Note that, when both rvs $\mathrm{x}_1$and $\mathrm{x}_2$ are observed, this corresponds to a generative model of the type studied in Chapter 6.
- Following the same steps as in the examples above, one can see that this joint distribution is in the exponential family, with sufficient statistics given by $s(x) = [x_1, x_2(1 - x_1), x_2 x_1]^T$ and natural parameter vector $\eta = [\log(\mu/(1-\mu)), \beta\nu_0, \beta\nu_1]$ .
- The marginal distribution of $\mathrm{x}_1$ under this joint distribution is a mixture of Gaussians, which is not in the exponential family.
- This example shows that the exponential family is not “closed” with respect to the operation of marginalization.

# Alternative Formulation of Exponential Family

- As we have seen, the log-loss for the exponential family is given as

$$-\log p(x|\eta) = -\underbrace{\eta^T s(x)}_{\sum_{k=1}^K \eta_k s_k(x)} - \underbrace{M(x)}_{\text{log-base measure}} + \underbrace{A(\eta)}_{\text{log-partition function}} .$$

- For the purpose of analytical calculations, it is often convenient to write the distribution in terms of an augmented natural parameter vector $\tilde{\eta} = \begin{bmatrix} \eta \\ 1 \end{bmatrix}$ and augmented sufficient statistics $\tilde{u}(x) = \begin{bmatrix} s(x) \\ M(x) \end{bmatrix}$, yielding

$$-\log p(x|\eta) = -\underbrace{\tilde{\eta}^T \tilde{u}(x)}_{\sum_{k=1}^K \tilde{\eta}_k \tilde{u}_k(x)} + \underbrace{A(\eta)}_{\text{convex function of } \eta}$$

or equivalently

$$p(x|\eta) \propto \exp\left(\tilde{\eta}^T \tilde{u}(x)\right).$$

# Exponential Family

- This a useful reference table. Note that, when $s(x)$ is a square matrix, the corresponding natural parameters $\eta$ are also in the form a matrix of the same dimension and the inner product is written as the trace $\mathrm{tr}(\eta s(x))$.

| distribution | $s(x)$ | $\eta$ | $\mu$ |
|---|---|---|---|
| Bern($\mu$) | $x$ | $\log\left(\frac{p}{1-p}\right)$ (logit) | $\sigma(\eta)$ |
| Cat($\mu$) | $x^{OH}$ (one-hot vector) | $\eta_k = \log(a\mu_k)$ for $a > 0$ (logits) | $\mu = \mathrm{softmax}(\eta)$ |
| $\mathcal{N}(\nu, \Theta^{-1})$, fixed $\Theta$ | $x$ | $\Theta\nu$ | $\nu$ |
| $\mathcal{N}(\nu, \Theta^{-1})$ | $x$ and $xx^T$ | $\Theta\nu$ and $-\frac{1}{2}\Theta$ | $\nu$ and $\Theta^{-1} + \nu\nu^T$ |

# Exponential Family via Maximum-Entropy Modelling

- To conclude this first section, we ask: How can we justify the use of the exponential family apart from analytical tractability?
- Suppose that the only information available about some data $x$ is given by the means $\mathrm{E}_{\mathrm{x}\sim p(x)}[s_k(\mathrm{x})] = \mu_k$ of given functions, or statistics, $s_k(x)$ for $k = 1, ..., K$
  - How should we choose $p(x)$?
  - Note that we cannot use density estimation since we do not have samples from $\mathrm{x}$.
- Ex.: We measure empirical mean average lifetime of all the nuclei of a radioactive atomic species – how should we model their distribution?
- One well-established principle is to choose the distribution $p(x)$ that is least predictable, or "more random", under the given average constraints.

# Exponential Family via Maximum-Entropy Modelling

- Recall that the entropy is a measure of “unpredictability” of a random variable, i.e., it measures the minimum average prediction log-loss when all that is known is the distribution $p(x)$.
- So the outlined problem can be formulated as the optimization

$$\max_{p(x)} \mathrm{H}(p(x)) \text{ s.t. } \mathrm{E}_{\mathrm{x}\sim p(x)}[s_k(\mathrm{x})]{=}\mu_k \text{ for } k = 1, ..., K.$$

- It can be proved that the distribution

$$p(x|\eta) = \exp\left(\eta^T s(x) + M(x) - A(\eta)\right),$$

  from the exponential family solves this problem, where each natural parameter $\eta_k$ is the optimal Lagrange multipliers associated with the $k$th constraint. This provides another interesting link between mean and natural parameters.
- This result offers a theoretical justification for the use of distributions in the exponential family:
  - the exponential family “makes the least assumptions” while being consistent with the available data.

# Gradient, or Score Vector

# Gradient of the Log-Loss, or Negative Score Vector

- As we will prove, the partial derivative of the log-loss with respect to each natural parameter $\eta_k$ is

$$\frac{\partial(-\log p(x|\eta))}{\partial\eta_k} = \underbrace{\mu_k - s_k(x)}_{\text{mean error for } s_k(x)}.$$

- Equivalently, the gradient with respect to the natural parameters is

$$\nabla_\eta(-\log p(x|\eta)) = \underbrace{\mu - s(x)}_{\text{mean error for } s(x)}.$$

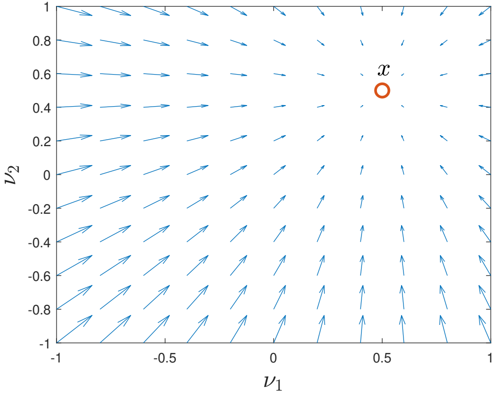
- In other words, the score vector is given as the negative mean error

$$\nabla_\eta \log p(x|\eta) = s(x) - \mu.$$

- As we have already seen in Chapter 6, this formula underlies many machine learning algorithms based on gradient descent.

# Gradient, or Score Vector

- The score vector $\nabla \log p(x|\eta)$ points to the direction in natural parameter space that locally maximizes the log-probability of $x$.
- Example: The score vector $\nabla_\nu \log \mathcal{N}(x|\nu, I) = x - \nu$ is illustrated in the figure for $x = [0.5, 0.5]^T$.



# Gradient, or Score Vector

- Proof of the gradient formula: By using the expression of the log-loss, we directly have

$$\frac{\partial(-\log p(x|\eta))}{\partial\eta_k} = -s_k(x) + \frac{\partial A(\eta)}{\partial\eta_k}.$$

- Moreover, we have the relationship (see Appendix):

$$\frac{\partial A(\eta)}{\partial\eta_k} = \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}\left[s_k(\mathrm{x})\right] = \mu_k$$

or, in vector form,

$$\nabla_\eta A(\eta) = \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}\left[s\left(\mathrm{x}\right)\right] = \mu.$$

- This concludes the proof.

# From Natural Parameters to Mean Parameters

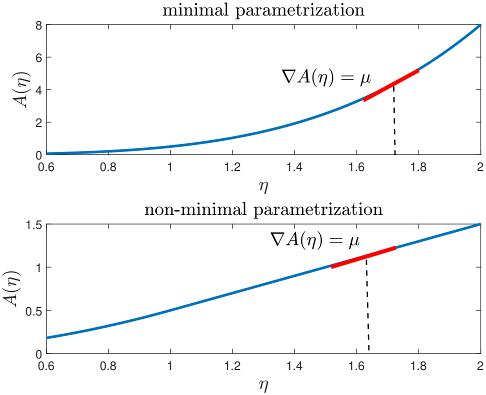
- The identity

  $$\nabla_\eta A(\eta) = \mu$$

  is a key result:
  - The gradient $\nabla_\eta A(\eta)$ of the log-partition function maps natural parameter vector $\eta$ to the corresponding mean parameter vector.
- The inverse mapping from $\mu$ to $\eta$ exists only if the distribution is minimal, in which case the mapping between $\mu$ and $\eta$ is one-to-one. We refer to the Appendix for further discussion on this point.

# From Natural Parameters to Mean Parameters

- The figure illustrates the one-to-one mapping between natural and mean parameters for minimal parametrizations, in which case the log-partition function is strictly convex (i.e., it is a strictly positive curvature), and the many-to-one mapping between natural and mean parameters for non-minimal parametrizations, for which the log-partition function is convex (i.e., it has zero curvature for some values of $\eta$).


minimal parametrization

non-minimal parametrization

# ML Learning

# Training Models from the Exponential Family

- Let us now consider the problem of ML learning for probabilistic models in the exponential family.
- Given training data $\mathcal{D} = \{x_n\}_{n=1}^{N}$, the training log-loss is given as

$$\begin{aligned} L_{\mathcal{D}}(\eta) &= -\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\eta) \\ &= -\eta^T \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N} s(x_n)\right)}_{:=s(\mathcal{D}),\ \text{empirical average of the suff. statistics}} \\ &\quad -\frac{1}{N}\sum_{n=1}^{N} M(x_n) + A(\eta). \end{aligned}$$

# Training Models from the Exponential Family

- Therefore, the training log-loss, seen as a function of the model parameter vector $\eta$, depends on the training set $\mathcal{D}$ only through the empirical average of the sufficient statistics

$$s(\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N} s(x_n).$$

- Note, in fact, that the term $N^{-1}\sum_{n=1}^{N} M(x_n)$ does not depend on the parameter vector $\eta$.
- It follows that the training does not require the entire data set and its complexity – in terms of computation and storage – does not increase with the size of the data set $N$.

# Training Models from the Exponential Family

- Using the formula derived above for the score function, the gradient of the log-loss can be directly computed as

$$\begin{aligned}\nabla L_{\mathcal{D}}(\eta) &= \frac{1}{N}\sum_{n=1}^{N}\nabla_{\eta}(-\log p(x_n|\eta)) \\ &= \frac{1}{N}\sum_{n=1}^{N}\underbrace{(\mu - s(x_n))}_{\text{mean error for } s(x_n)} \\ &= \underbrace{\mu - s(\mathcal{D})}_{\text{mean error for } s(\mathcal{D})},\end{aligned}$$

  where we recall that $\mu = \nabla_{\eta}A(\eta)$ is the mean parameter associated to the natural parameter vector $\eta$.

- The gradient is hence given by mean error signal obtained as the difference between the ensemble average under the model $\mu$ and the empirical average $s(\mathcal{D}) = N^{-1}\sum_{n=1}^{N}s(x_n)$ of the observations for the sufficient statistics.

# Training Models from the Exponential Family

- Since the log-loss is convex, the stationarity condition $\nabla L_{\mathcal{D}}(\eta) = 0$ is necessary and sufficient for global optimality:
  - It follows that the ML estimate of the mean parameters is given as

    $$\mu^{ML} = s(\mathcal{D}),$$

    that is, as the empirical average of the sufficient statistics;
  - This moment matching condition can be written more explicitly as

    $$\underbrace{\mathrm{E}_{\mathrm{x}\sim p(x|\mu)}[s(\mathrm{x})]}_{=\mu} = \underbrace{\mathrm{E}_{\mathrm{x}\sim p_{\mathcal{D}}(x)}[s(\mathrm{x})]}_{=s(\mathcal{D})},$$

    where $p_{\mathcal{D}}(x)$ is the empirical distribution of the data;
  - For minimal distributions, we can obtain the ML estimate $\eta^{ML}$ from $\mu^{ML}$ using the one-to-one mapping; for non-minimal families there will be multiple equivalent solutions $\eta^{ML}$ for the ML problem.

# Example

- We have already encountered the moment matching condition in Chapter 6 when discussing the training of generative models, which required the ML training of Bernoulli, categorical, and Gaussian distributions.
- As a reminder and a simple example, consider the problem of ML training for a Bernoulli distribution Bern($\mu$) using training set $\mathcal{D}=\{0,1,1,0,0,0,1,0,0,0\}$ with $N=10$.
- Using the moment matching condition, we have the ML estimate of the mean parameter

  $$\mu^{ML} = \frac{1}{N}\sum_{n=1}^{N} s(x_n) = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{N[1]}{N} = \frac{3}{10},$$

  where

  $$N[1] = |\{n:\ x_n = 1\}| = 3$$

  is the count of observations equal to 1.
- Since this is a minimal distribution, we can obtain the unique ML estimate of the logit, i.e., of the natural parameter, as $\eta^{ML} = \log(\mu^{ML}/(1-\mu^{ML}))$.

# Example

- As another example, consider a categorical distribution $\mathrm{Cat}(\mu)$ with $C = 4$, and a data set $\mathcal{D} = \{0, 1, 2, 0, 0, 0, 2, 0, 0, 0\}$ with $N = 10$.
- Using moment matching, the ML estimate is

$$\mu^{ML} = \frac{1}{N} \begin{bmatrix} N[0] \\ N[1] \\ N[2] \\ N[3] \end{bmatrix},$$

  which is the standard histogram of the observations.
- Note the "black-swan problem": The value $x = 3$ was never observed, and the model assigns it zero probability!
- Since the parametrization is not minimal, there is an infinity of possible solutions for the logits, i.e., for the natural parameters, namely any vector

$$\eta^{ML} = \begin{bmatrix} \log(\mu_0^{ML}) \\ \vdots \\ \log(\mu_{C-1}^{ML}) \end{bmatrix} + b \times 1_C$$

  for some constant $b$.

# Training Models from the Exponential Family

- It is important to realize that the moment matching condition is not always tractable.
- This is because it may difficult to compute the moment parameters $\mu$ as a function of the model parameters, and hence this condition cannot be solved explicitly.
- An example is given by the Boltzmann distribution

$$\begin{aligned} -\log p(x|\eta) &= -x^T W x + a^T x + A(\eta), \\ &= -\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j - \sum_{i=1}^{D} a_i x_i + A(\eta), \end{aligned}$$

  where $\eta = (a, W)$ are the natural parameters and the sufficient statistics are given as

$$s(x) = [x, xx^T].$$

- For cases such as this one, as we have seen in the special case of RBM in Chapter 7, one can leverage SGD based on the formula for the gradient obtained above.

# Information-Theoretic Metrics for the Exponential Family

# Information-Theoretic Metrics for the Exponential Family

- Distributions in the exponential family have the useful property that information-theoretic metrics can be efficiently evaluated.
- This is important in many learning methods that rely on information-theoretic metrics such as KL divergence and entropy.
- Using the augmented formulation $p(x|\eta) \propto \exp\left(\tilde{\eta}^T \tilde{u}(x)\right)$ seen above, the general form of the entropy for distributions in the exponential family is (see Appendix for a proof)

$$\begin{aligned} \mathrm{H}(\mathrm{ExpFam}(x|\eta)) &= \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}\left[-\log p(\mathrm{x}|\eta)\right] \\ &= -\tilde{\eta}^T \tilde{\mu} + A(\eta), \end{aligned}$$

  where we have defined the augmented mean vector as $\tilde{\mu} = \left[ \begin{array}{c} \mu \\ \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}[M(\mathrm{x})] \end{array} \right]$.
- This relation shows that (negative) entropy and log-partition function are “dual” to each other, in a sense that is made precise in the Appendix.

# Information-Theoretic Metrics for the Exponential Family

- This general expression can be evaluated explicitly as a function of $\mu$ or $\eta$ for each distribution in the exponential family. Examples are given in the table.

| distribution | $\mathrm{H}(\mathrm{ExpFam}(x\|\eta))$ |
|---|---|
| Bern($\mu$) | $-\mu \log \mu - (1-\mu) \log(1-\mu)$ |
| Cat($\mu$) | $-\sum_{k=0}^{C-1} \mu_k \log(\mu_k)$ |
| $\mathcal{N}(\nu, \Theta^{-1})$ | $\frac{1}{2} \log \det((2\pi e)\Theta^{-1})$ |

# Information-Theoretic Metrics for the Exponential Family

- The general form of the KL divergence for distributions in the same class within the exponential family can be computed as

$$
\begin{aligned}
\mathrm{KL}(\mathrm{ExpFam}(x|\eta_1)||\mathrm{ExpFam}(x|\eta_2)) &= \mathrm{E}_{\mathrm{x}\sim p(x|\eta_1)}\left[\log\frac{p(\mathrm{x}|\eta_1)}{p(\mathrm{x}|\eta_2)}\right] \\
&= A(\eta_2) - A(\eta_1) - (\eta_2 - \eta_1)^T \mu_1,
\end{aligned}
$$

  where $\mu_1$ is the mean parameter vector corresponding to $\eta_1$.
- This formula shows that the KL divergence can be computed as a distance beween $\eta_1$ and $\eta_2$, in a sense that is again made precise in the Appendix.

# Information-Theoretic Metrics for the Exponential Family

- The formula can be computed explicitly for all distributions in the exponential family, and some examples can be found in the table.

| $p(x)$ | $q(x)$ | $\mathrm{KL}(p\|\|q)$ |
|---|---|---|
| $\mathsf{Bern}(\mu)$ | $\mathsf{Bern}(\bar{\mu})$ | $\mu \log \frac{\mu}{\bar{\mu}} + (1-\mu) \log \frac{1-\mu}{1-\bar{\mu}}$ |
| $\mathsf{Cat}(\mu)$ | $\mathsf{Cat}(\bar{\mu})$ | $\sum_{k=0}^{C-1} \mu_k \log \frac{\mu_k}{\bar{\mu}_k}$ |
| $\mathcal{N}(\nu, \Sigma)$ | $\mathcal{N}(\bar{\nu}, \bar{\Sigma})$ | $\frac{1}{2}\left[\mathrm{tr}\left(\bar{\Sigma}^{-1}\Sigma\right) + \log\left(\frac{\det(\bar{\Sigma})}{\det(\Sigma)}\right) + (\bar{\nu}-\nu)^T \bar{\Sigma}^{-1}(\bar{\nu}-\nu) - D\right]$ |

# Fisher Information Matrix

# Score Vector and Fisher Information Matrix

- Related to information-theoretic measures is an important metric known as Fisher information matrix (FIM).
- To describe it, let us focus first on a general probabilistic models $p(x|\theta)$, not necessarily from the exponential family.
- The cross entropy

$$\mathrm{H}(p(x|\theta), p(x|\theta')) = \mathrm{E}_{\mathrm{x} \sim p(x|\theta)}[-\log p(x|\theta')].$$

  can be interpreted as the population log-loss when the population distribution is $p(x|\theta)$ for some ground-truth value $\theta$ and $\theta'$ is the model parameter vector.

- We know that the minimum of $\mathrm{H}(p(x|\theta), p(x|\theta'))$ over $\theta'$ – and equivalently the minimum of $\mathrm{KL}(p(x|\theta)||p(x|\theta'))$ over $\theta'$ – is given by $\theta' = \theta$.
- This has two useful consequences.

# Score Vector and Fisher Information Matrix

- 1) The first-order optimality condition requires the equality $\nabla_{\theta'} \mathrm{H}(p(x|\theta), p(x|\theta'))|_{\theta'=\theta} = 0$, which implies

$$\nabla_{\theta'} \mathrm{H}(p(x|\theta), p(x|\theta'))|_{\theta'=\theta} = \mathrm{E}_{\mathrm{x} \sim p(x|\theta)} \left[ - \underbrace{\nabla_\theta \log p(\mathrm{x}|\theta)}_{\text{score vector}} \right] = 0 :$$

  - The score vector $\nabla_\theta \log p(x|\theta)$ has zero mean when averaged over $p(x|\theta)$.

# Score Vector and Fisher Information Matrix

- 2) The second-order optimality condition $\nabla^2_{\theta'} \mathrm{H}(p(x|\theta), p(x|\theta'))|_{\theta'=\theta} \succeq 0$ implies

$$\nabla^2_{\theta'} \mathrm{H}(p(x|\theta), p(x|\theta'))|_{\theta'=\theta} = \underbrace{\mathrm{E}_{\mathrm{x} \sim p(x|\theta)}[-\nabla^2_{\theta} \log p(\mathrm{x}|\theta)]}_{:=\mathrm{FIM}(\theta),\ \text{Fisher Information Matrix (FIM)}} \succeq 0 :$$

  - The FIM $\mathrm{FIM}(\theta)$ measures the (non-negative) curvature of the population log-loss $\mathrm{H}(p(x|\theta), p(x|\theta'))$ at the optimal point $\theta' = \theta$:
    - ⋆ The "larger" the FIM $\mathrm{FIM}(\theta^*)$ is, the "easier" it is to obtain the optimal value $\theta$ when minimizing $\mathrm{H}(p(x|\theta), p(x|\theta'))$ over $\theta'$.

# Fisher Information Matrix

- The FIM $\mathrm{FIM}(\theta)$ quantifies the amount of information that data generated from the model $p(x|\theta)$ provides about the value of the model parameter $\theta$.
- Ex.: For a $\mathrm{Bern}(\mu)$ rv, the FIM is $\mathrm{FIM}(\mu) = \frac{1}{\mu(1-\mu)}$ :
  - When data is generated as $\mathrm{x} \sim \mathrm{Bern}(\mu)$, it is easier to estimate the parameter values $\mu = 0$ and $\mu = 1$, and most difficult to estimate $\mu = 0.5$, at which point the observations are maximally random.
- Ex.: For a $\mathcal{N}(\nu, \beta^{-1})$ rv with a fixed precision $\beta$, the FIM is $\mathrm{FIM}(\nu) = \beta$:
  - When data is generated as $\mathrm{x} \sim \mathcal{N}(\nu, \beta^{-1})$ all values $\nu$ are equally difficult to estimate, and the amount of information we have about $\nu$ increases with the precision $\beta$.

# Fisher Information Matrix

- The FIM can also be written as the covariance of the score vector, i.e.,

$$\begin{aligned}\mathrm{FIM}(\theta) =& \mathrm{E}_{\mathrm{x}\sim p(x|\theta)}[-\nabla_\theta^2 \log p(\mathrm{x}|\theta)]\\ =& \mathrm{E}_{\mathrm{x}\sim p(x|\theta)}[(\nabla_\theta \log p(\mathrm{x}|\theta))(\nabla_\theta \log p(\mathrm{x}|\theta))^T].\end{aligned}$$

- Note that the above is the covariance matrix since the mean of the score vector is zero.
- A proof of this equality and further discussion on the FIM can be found in the Appendix.

# Fisher Information Matrix

- The same argument can also be applied by swapping the role of $\theta$ and $\theta'$.
- Putting together the resulting first and second-order derivatives, we have the following useful Taylor second-order approximation of the KL divergence

  $$\mathrm{KL}(p(x|\theta), p(x|\theta_0)) \simeq \frac{1}{2}(\theta - \theta_0)^T \mathrm{FIM}(\theta_0)(\theta - \theta_0)$$

  around any point $\theta_0$.
- So the FIM $\mathrm{FIM}(\theta_0)$ describes the curvature of the KL divergence around $\theta_0$.

# Fisher Information Matrix for Exponential Family

- Having discussed the FIM in the context of general probability models, we now specialize the results to the exponential family.
- Using the general definition above, the FIM for the natural parameters can be directly computed as

$$\begin{aligned}\mathrm{FIM}(\eta) &= \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}[-\nabla_\eta^2 \log p(\mathrm{x}|\eta)],\\ &= \nabla_\eta^2 A(\eta),\end{aligned}$$

  that is the FIM is the Hessian of the log-partition function.
- This implies that the FIM is positive definite $\mathrm{FIM}(\eta) \succ 0$, and hence invertible, if the class of distributions is minimal.

# Fisher Information Matrix for Exponential Family

- As a result of the equality above, by computing the Hessian of the log-partition function, the FIM can also be written as the covariance of the sufficient statistic vector $s(\mathrm{x})$, i.e.,

$$\mathrm{FIM}(\eta) = \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}\left[(s(\mathrm{x}) - \mu)(s(\mathrm{x}) - \mu)^T\right].$$

- Proofs are in the Appendix.
- The FIM can also be used to define an alternative to gradient descent that is known as natural gradient descent. Unlike gradient descent, natural gradient descent operates in the geometry implied by the KL divergence in the space of distributions.
- Intuitively, and informally, natural gradient descent can be seen as a way to approximate Newton's method. This is because the Hessian $N^{-1}\sum_{n=1}^{N}(-\nabla_{\eta}^{2}\log p(\mathrm{x}_n|\eta))$ of the log-loss $N^{-1}\sum_{n=1}^{N}(-\log p(\mathrm{x}_n|\eta))$ for general probabilistic models tends to the FIM as $N$ grows large. Note that this is true if the data is assumed to be generated by the model, i.e., if we have i.i.d. samples $\mathrm{x}_n \sim p(x|\eta)$.
- More discussion can be found in the Appendix.

# Fisher Information Matrix for Exponential Family

- The general formula above can be computed explicitly for all distributions in the exponential family, as per the examples in the table.

| distribution | $\mathrm{FIM}(\mu)$ |
|---|---|
| Bern$(\mu)$ | $\frac{1}{\mu(1-\mu)}$ |
| $\mathcal{N}(\nu, \Theta^{-1})$, fixed $\Theta$ | $\Theta^{-1}$ |
| $\mathcal{N}(\nu, \Theta^{-1})$ | $\begin{bmatrix} \Theta^{-1} & 0 \\ 0 & \frac{1}{2}\Theta^{-2} \end{bmatrix}$ |

# Generalized Linear Models

# Generalized Linear Models (GLM)

- The exponential family provides a flexible class of distributions to model densities.
- In many problems in machine learning, we need to model conditional distributions.
- A useful extension of exponential-family distributions to conditional models is given by GLMs.
- We have already encountered several GLMs in Chapter 4 and 6, namely polynomial regression in Chapter 4 and logistic and softmax regression models in Chapter 6.
- A GLM defines a model class of conditional probabilities defined as

$$p(t|x, W) = \mathrm{ExpFam}(t|\mu = g(Wu(x)))$$

  where
  - $\mathrm{ExpFam}(t|\mu)$ represents any distribution in the exponential family with mean parameter vector $\mu$;
  - $u(x)$ is a vector of features, as defined in previous chapters;
  - $W$ is a matrix defining the model parameters;
  - $g(\cdot)$ is an invertible function – the inverse $g^{-1}(\cdot)$ is known as link function.

# Generalized Linear Models (GLM)

- Intuitively, GLMs generalize deterministic linear models of the form $t = Wu(x)$ by "adding noise" around the mean $\mu = g(Wu(x))$ that is drawn from a distribution in the exponential family.
- GLMs in the canonical form are written as

  $$p(t|x, W) = \mathrm{ExpFam}(t|\eta = Wu(x)),$$

  so that the natural parameter vector is the linear function $\eta = Wu(x)$.
- Note that here the rv is $\mathrm{t}$, while $x$ is the fixed input.
- This corresponds to setting the inverse of the link function as $g(\eta) = \nabla_\eta A(\eta)$.
- As two examples, we will now see that logistic and softmax regression are special cases of GLMs.

# Logistic Regression as a GLM

- As we discussed in Chapter 6, logistic regression assumes that the label is conditionally distributed as

$$(\mathrm{t}|\mathrm{x}{=}x,\theta){\sim}\mathrm{Bern}(\sigma(\theta^T u(x))),$$

  so that we have the predictive distribution $p(\mathrm{t}=1|x,\theta)=\sigma(\theta^T u(x))$ for model parameter $\theta$.
- Therefore, logistic regression is a GLM with exponential-family distribution given by the Bernoulli distribution and natural parameter given by the logit $\eta=\theta^T u(x)$.
- Note that the gradient with respect to $\theta$ derived in Chapter 6 can be directly obtained from the score function derived above via the chain rule as

$$\begin{aligned}\nabla_\theta(-\log\mathrm{Bern}(\sigma(\theta^T u(x)))) &= \frac{\partial}{\partial\eta}(-\log\mathrm{Bern}(\eta)|_{\eta=\theta^T u(x)})\times\nabla_\theta\eta\\ &= \underbrace{(\sigma(\theta^T u(x))-t)}_{:=\delta(x,t),\ \text{mean error}}\times\underbrace{u(x)}_{\text{feature vector}}.\end{aligned}$$

# Softmax Regression as GLM

- Softmax regression assumes that the one-hot label vector is distributed as

$$(\mathrm{t}|\mathrm{x} = x, W) \sim \mathsf{Cat}(t|\eta = Wu(x))$$

  for model parameter matrix $W$.
- Therefore, the conditional distribution $p(t|x, W)$ is a GLM with exponential-family distribution given by the categorical distribution and natural parameter given by the logit vector $\eta = Wu(x)$.
- Again, it can be readily checked that the gradient with respect to $W$ derived in Chapter 6 can be directly obtained form the score function derived above via the chain rule.

# Summary

# Summary

- The exponential family contains a large number of discrete and continuous parametric distributions:
  - it contains all distributions with finite-dimensional parameterization and fixed support.
- Distributions in the exponential family have convex log-loss, easy-to-compute score vectors (i.e., gradients of the log-loss) and information-theoretic measures, including FIMs.
- A distribution in the exponential family can be expressed in terms of natural parameters $\eta$ or mean parameters $\mu$; we can write

  $$p(x|\eta) = \mathrm{ExpFam}(x|\eta)$$

  or

  $$p(x|\mu) = \mathrm{ExpFam}(x|\mu)$$

  where $\mathrm{ExpFam}$ denotes any distribution in the exponential family.

# Summary

- For every natural parameter vector $\eta$, there is a unique mean vector $\mu = \mathrm{E}_{\mathrm{x} \sim \mathrm{ExpFam}(x|\eta)}[s(\mathrm{x})]$.
- For a mean vector $\mu$, there may be more than one natural parameters $\eta$ unless the distribution class is minimal.
- For all distributions in the exponential family, by definition, the log-loss is a convex function

$$-\log(\mathrm{ExpFam}(x|\eta)) = -\eta^T s(x) + A(\eta) + \text{terms indep. of } \eta$$

- Furthermore, for a training set $\mathcal{D}$, the training log-loss

$$L_{\mathcal{D}}(\eta) = \frac{1}{N} \sum_{n=1}^{N} (-\log(\mathrm{ExpFam}(x_n|\eta))) = -\eta^T s(\mathcal{D}) + A(\eta),$$

  and hence also the ML estimates of the parameters, depend only on the empirical average of the sufficient statistics $s(\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} s(x_n)$.

# Summary

- The gradient of the log-loss, or negative score vector, is given as

$$\nabla(-\log(\mathrm{ExpFam}(x|\eta))) = \underbrace{\mu - s(x)}_{\text{mean error for } s(x)}$$

  and the gradient of the training loss is given as

$$\nabla L_{\mathcal{D}}(\eta) = \underbrace{\mu - s(\mathcal{D})}_{\text{mean error for } s(\mathcal{D})} .$$

- ML is obtained via moment matching, or, when not feasible, via gradient descent.
- Information-theoretic measures and FIM can be easily computed.

# Appendix

# Gradient of the Log-Partition Function

- Here, we validate the relationship between mean parameters and gradient of the log-partition function introduced in the text.
- This is done through the direct calculation

$$\begin{aligned}\frac{\partial A(\eta)}{\partial \eta_k} &= \frac{\sum_x s_k(x) \exp\left(\eta^T s(x) + M(x)\right)}{\sum_{x'} \exp\left(\eta^T s(x') + M(x')\right)} \\ &= \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}\left[s_k(\mathrm{x})\right] = \mu_k .\end{aligned}$$

- The same derivation applies for continuous rvs by replacing sums with integrals.

# Duality and Exponential Family

- Exponential family distributions have interesting properties in terms of duality, as we explore next.
- To start, let us first define the Bregman divergence:
  - given a convex function $f(\cdot)$, the Bregman divergence is defined as

$$B_f(x, y) = f(x) - f(y) - (x - y)^T \nabla f(x);$$

  - using the convexity of $f(x)$, it can be directly proved that we have $B_f(x, y) \geq 0$ and $B_f(x, x) = 0$;
  - it can also be seen that $B_f(x, y)$ is convex in $x$, but not necessarily in $y$.

# Convex Duality

- We also need the definition of convex dual, or Fenchel dual, of a convex function $f(x)$ as
$$f^*(y) = \max_x \left\{ x^T y - f(x) \right\}.$$
- Geometrically, the convex dual finds the negative intercept of the tangent to the function $f(x)$ with gradient $y$.
- In fact, since $f(x)$ is convex, global optimal solutions can be found by solving the equation $\nabla f(x) = y$. Accordingly, we have $f^*(y) = x^T \nabla f(x) - f(x)$ where $x$ is such that we have $y = \nabla f(x)$. The intercept is defined as the value at the origin of the domain of the first-order approximation, i.e.,
$$f(x) + \nabla f(x)^T (0 - x).$$
- A strictly convex function can be fully described by the set of intercepts for each possible gradient value $y$. In fact, for every possible gradient value $y$, there is at most one value of $x$ for which we have $\nabla f(x) = y$.
- For a convex function, there may be multiple such values of $x$.
- In either case, we also have $f^{**}(x) = f(x)$, that is, the convex dual of the convex dual is the function itself.

# Duality and Exponential Family

- Since the log-partition function is convex, we can define its convex dual as
$$A^*(y) = \max_{\eta} \left\{ \eta^T y - A(\eta) \right\}.$$
- Under the assumption of minimality, the log-partition function is strictly convex and hence its global optimum can be obtained by applying the first-order optimality solution $\nabla_\eta A(\eta) = y$:
  - It follows that the optimal value of $\eta$ is the natural parameter vector corresponding to mean parameter $\mu$.
- We conclude that we have the equality
$$A^*(\mu) = \eta^T \mu - A(\eta),$$
where $(\mu, \eta)$ is the pair of mean and natural parameters.

# Duality and Exponential Family

- A useful way to recall this relationship is

$$A^*(\mu) + A(\eta) = \eta^T \mu.$$

- This relationship can also be used to prove that, for minimal distributions, we have the inverse mapping between natural and mean parameters

$$\nabla_\mu A^*(\mu) = \eta.$$

- We now discuss some important implications of duality.

# Distribution and Log-Loss as a Function of the Mean Parameters

- With this background, we can first prove that, under the assumption of minimality, the distribution can be expressed as

  $$p(x|\mu) = \exp(-B_{A^*}(s(x),\mu) + G(x)),$$

  for a suitable function $G(x)$ independent of the model parameters.
- Hence all minimal distributions in the exponential family have associated a measure of distance in the mean parameter space:
  - The probability of $x$ depends on how far $s(x)$ is from the mean vector $\mu$.
- Ex.: For the Gaussian distribution $\mathcal{N}(\mu,1)$, we have $p(x|\mu) \propto \exp(-\underbrace{||x-\mu||^2/2}_{B_{A*}(s(x),\mu)})$.

# Distribution and Log-Loss as a Function of the Mean Parameters

- From the table, the Gaussian distribution is associated with the (weighted) squared Euclidean distance, while Bernoulli and categorical variables are associated with the KL divergence.
- We have defined $\mathrm{KL}(x||\mu) = x \log\left(\frac{x}{\mu}\right) + (1-x)\log\left(\frac{1-x}{1-\mu}\right)$ for the Bernoulli distribution and $\mathrm{KL}(x||\mu) = \sum_{k=0}^{C-1} x_k \log\left(\frac{x_k}{\mu_k}\right)$ for the categorical distribution.

| distribution | $A(\eta)$ | $A^*(\mu)$ | $B_{A^*}(s(x), \mu)$ |
|---|---|---|---|
| Bern$(\mu)$ | $\log(1+e^{\eta})$ | $\mu \log \mu + (1-\mu)\log(1-\mu)$ | $\mathrm{KL}(x\|\|\mu)$ |
| Cat$(\mu)$ | $\log\left(\sum_{k=0}^{C-1} e^{\eta_k}\right)$ | $\sum_{k=0}^{C-1} \mu_k \log \mu_k$ | $\mathrm{KL}(x\|\|\mu)$ |
| $\mathcal{N}(\nu, \Theta^{-1})$ with fixed $\Theta^{-1}$ | $\frac{1}{2}\eta^T\Theta^{-1}\eta$ | $\frac{1}{2}\mu^T\Theta\mu$ | $\frac{1}{2}\|\|s(x)-\mu\|\|_{\Theta}^2$ |

# Distribution and Log-Loss as a Function of the Mean Parameters

- Proof of the equality $p(x|\mu) = \exp(-B_{A^*}(s(x), \mu) + G(x))$: We have

$$
\begin{aligned}
p(x|\eta) &= \exp\left(\eta^T s(x) - A(\eta) + M(x)\right) \\
&= \exp\left(\eta^T s(x) - (\eta^T \mu - A^*(\mu)) + M(x)\right) \\
&= \exp\left(A^*(\mu) + \eta^T (s(x) - \mu) + M(x)\right) \\
&= \exp\left(-\underbrace{(A^*(s(x)) - A^*(\mu) - (s(x) - \mu)^T \nabla A^*(\mu))}_{B_{A^*}(s(x),\mu)} + \underbrace{(A^*(s(x)) + M(x))}_{=:G(x)}\right).
\end{aligned}
$$

# Duality and Entropy

- As another implication of duality, we show now that negative entropy and log-partition function are dual to one another.
- As we have seen in the main text, the entropy of an exponential family distribution can be written as

$$\begin{aligned} \mathrm{H}(p(x|\eta)) &= \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}[-\log p(\mathrm{x}|\eta)] \\ &= -\eta^T \mu + A(\eta) - \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}[M(\mathrm{x})]. \end{aligned}$$

- So, we have the duality relationship between entropy and log-partition function

$$A^*(\mu) = -\mathrm{H}(p(x|\eta)) - \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}[M(\mathrm{x})].$$

# Duality and KL Divergence

- We now relate the KL divergence between two distributions $p(x|\eta_1) = \mathrm{ExpFam}(x|\eta_1)$ and $p(x|\eta_2) = \mathrm{ExpFam}(x|\eta_2)$ from the same exponential family with log-partition function $A(\cdot)$, and sufficient statistics $s(\cdot)$, to the Bregman divergence. This will allow us to obtain an explicit expression as a function of the moment parameters $\mu_1$ and $\mu_2$.
- As we have seen in the text, we have the following identity

$$\begin{aligned}\mathrm{KL}(\mathrm{ExpFam}(x|\eta_1)||\mathrm{ExpFam}(x|\eta_2)) &= B_A(\eta_2, \eta_1)\\ &= A(\eta_2) - A(\eta_1) - (\eta_2 - \eta_1)^T \mu_1.\end{aligned}$$

- Using duality for minimal families, we can also write

$$\begin{aligned}\mathrm{KL}(\mathrm{ExpFam}(x|\mu_1)||\mathrm{ExpFam}(x|\mu_2)) &= B_{A^*}(\mu_2, \mu_1)\\ &= A^*(\mu_1) - A^*(\mu_2) - (\mu_1 - \mu_2)^T \eta_2.\end{aligned}$$

# Fisher Information Matrix

- We now prove the equality

$$\begin{aligned}\mathrm{FIM}(\theta) =& \mathrm{E}_{\mathrm{x}\sim p(x|\theta)}[-\nabla_\theta^2 \log p(\mathrm{x}|\theta)] \\ =& \mathrm{E}_{\mathrm{x}\sim p(x|\theta)}[(\nabla_\theta \log p(\mathrm{x}|\theta))(\nabla_\theta \log p(\mathrm{x}|\theta))^T].\end{aligned}$$

- To prove this result, we first note that the Hessian $\nabla_\theta^2 \log p(x|\theta)$ can be written as

$$\begin{aligned}\nabla_\theta^2 \log p(x|\theta) &= \nabla_\theta(\nabla_\theta \log p(x|\theta)) \\ &= \nabla_\theta \left(\frac{\nabla_\theta p(x|\theta)}{p(x|\theta)}\right) \\ &= -\left(\frac{\nabla_\theta p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla_\theta p(x|\theta)}{p(x|\theta)}\right)^T + \frac{\nabla_\theta^2 p(x|\theta)}{p(x|\theta)} \\ &= -(\nabla_\theta \log p(x|\theta))(\nabla_\theta \log p(x|\theta))^T + \frac{\nabla_\theta^2 p(x|\theta)}{p(x|\theta)}.\end{aligned}$$

# Fisher Information Matrix

- Now, taking the expectation, the first term recovers the desired result, while the second equals an all-zero vector since

$$
\begin{aligned}
\mathrm{E}_{\mathrm{x}\sim p(x|\theta)}\left[\frac{\nabla_\theta^2 p(\mathrm{x}|\theta)}{p(\mathrm{x}|\theta)}\right] &= \sum_x \nabla_\theta^2 p(x|\theta) \\
&= \nabla_\theta^2 \underbrace{\sum_x p(x|\theta)}_{=1} = 0.
\end{aligned}
$$

# FIM and Estimation

- We have said in the text that $\mathrm{FIM}(\theta)$ measures how "easy" it is to estimate $\theta$ based on data from the model $p(x|\theta)$. We now discuss a formal statement of this property.
- In statistics, and sometimes also in machine learning, it is assumed that the model class

$$\mathcal{H} = \{p(x, t|\theta) : \theta \in \Theta\}$$

  includes the (unknown) population distribution $p(x, t)$ for some true (unknown) value of the parameter vector $\theta_0$, i.e., $p(x, t|\theta_0) = p(x, t)$.
- This is also referred to as a realizabilty assumption since the true distribution can be "realized" by the model.
- In this case, one can ask how well the true parameter $\theta_0$ is estimated.

# FIM and Estimation

- Under the realizability assumption, it is common in statistics to provide properties of an estimator of $\theta_0$ based on data.
- Most notably, as the number of data points goes to infinity, i.e., $N \to \infty$, one can prove the following properties for the ML estimator $\theta^{ML}$:
  - ML provides a consistent estimate, that is, we have $\theta^{ML} \to \theta_0$ with high probability;
  - ML is asymptotically Gaussian, that is, the rv $\sqrt{N}(\theta^{ML} - \theta_0)$ tends to have distribution $\mathcal{N}(0, \mathrm{FIM}(\theta_0)^{-1})$.

# Fisher Information Matrix for Exponential Family

- Here, we prove the equalities $\mathrm{FIM}(\eta) = \nabla_\eta^2 A(\eta) = \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}[(s(\mathrm{x}) - \mu)(s(\mathrm{x}) - \mu)^T]$.
- Using the expression for the score vector $\nabla_\eta(\log p(x|\eta)) = s(x) - \mu$, we have

$$
\begin{aligned}
\mathrm{FIM}(\eta) &= \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}[-\nabla_\eta^2 \log p(\mathrm{x}|\eta)] \\
&= \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}[-\nabla_\eta(\nabla_\eta \log p(\mathrm{x}|\eta))] \\
&= \mathrm{E}_{\mathrm{x} \sim p(x|\eta)}[-\nabla_\eta s(\mathrm{x}) + \nabla_\eta \mu] \\
&= \nabla_\eta \mu \\
&= \nabla_\eta(\nabla_\eta A(\eta)) \\
&= \nabla_\eta^2 A(\eta).
\end{aligned}
$$

# Fisher Information Matrix for Exponential Family

- Using the expression for the score vector $\nabla_\eta(\log p(x|\eta)) = s(x) - \mu$, we also have

$$\begin{aligned}\mathrm{FIM}(\eta) &= \mathrm{E}_{\mathrm{x}\sim p(x|\eta)}[(\nabla_\eta \log p(\mathrm{x}|\eta))(\nabla_\eta \log p(\mathrm{x}|\eta))^T] \\ &= \mathrm{E}_{\mathrm{x}\sim p(x|\mu)}[(s(\mathrm{x}) - \mu)(s(\mathrm{x}) - \mu)^T].\end{aligned}$$

# Natural Gradient Descent

- Recalling that we have the mapping $\nabla_\eta A(\eta) = \mu$, the FIM can also be expressed in terms of the mean parameters as the Jacobian

$$\mathrm{FIM}(\eta) = \nabla^2_\eta A(\eta) = \nabla_\eta \mu.$$

- As we will see next, this is useful to relate derivatives with respect to mean and natural parameters.
- To see this, consider a function $g(\eta)$ of the natural parameters, e.g., $g(\eta) = \mathrm{E}_{x \sim p(x|\eta)}[f(\mathrm{x})]$ for some function $f(\cdot)$. We have the following relationship between the gradients $\nabla_\eta f(\eta)$ and $\nabla_\mu f(\eta)$:

$$\nabla_\eta f(\eta) = \nabla_\eta \mu \cdot \nabla_\mu f(\eta) = \mathrm{FIM}(\eta) \cdot \nabla_\mu f(\eta).$$

- Furthermore, if the class of distributions is minimal, we can also write

$$\nabla_\mu f(\eta) = (\mathrm{FIM}(\eta))^{-1} \nabla_\eta f(\eta).$$

# Natural Gradient Descent

- In gradient descent, we minimize at each step the following strictly convex approximant of the cost function

$$\tilde{g}_\gamma(\theta;\theta^{(i)}) = \underbrace{g(\theta^{(i)}) + \nabla g(\theta^{(i)})^T(\theta - \theta^{(i)})}_{\text{first-order Taylor approximation}} + \underbrace{\frac{1}{2\gamma}||\theta - \theta^{(i)}||^2}_{\text{proximity penalty}}.$$

- Therefore, the distance between $\theta$ and the previous iterate is measured by the Euclidean distance $||\theta - \theta^{(i)}||^2$:
  - With probabilistic models, the Euclidean distance may not reflect actual changes in the distribution.
- Therefore, the choice of the learning rate must be conservative in order to avoid unstable updates.

# Natural Gradient Descent

- As an example, consider the optimization over the mean $\mu$ for a Gaussian distribution with fixed variance $\sigma^2$.
- For a fixed squared Euclidean distance $||\mu - \mu^{(i)}||^2$, the two distributions $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\mu^{(i)}, \sigma^2)$ may be more or less distinct depending on the value of $\sigma^2$ :
  - if $\sigma^2$ is large compared to $||\mu - \mu^{(i)}||^2$, the two distributions are very similar;
  - while the opposite is true when $\sigma^2$ is sufficiently smaller.

# Natural Gradient Descent

- In light of this, natural gradient descent replaces the square Euclidean distance $||\theta - \theta^{(i)}||^2$ with a more relevant measure of the distance between the two distributions $p(x|\theta)$ and $p(x|\theta^{(i)})$.
- As we know, a relevant measure is the KL divergence $\mathrm{KL}(p(x|\theta)||p(x|\theta^{(i)}))$.
- Using directly the KL divergence as a penalty yields the mirror descent method (to be discussed).
- Natural gradient descent approximates the KL divergence assuming that $\Delta\theta^{(i)} = \theta - \theta^{(i)}$ is sufficiently small by using the discussed second-order Taylor approximation with respect to $\Delta\theta^{(i)}$

$$\begin{aligned}\mathrm{KL}(p(x|\theta)||p(x|\theta^{(i)})) &\simeq \frac{1}{2}(\theta - \theta^{(i)})^T \mathrm{FIM}(\theta^{(i)})(\theta - \theta^{(i)}) \\ &= \frac{1}{2}||\theta - \theta^{(i)}||^2_{\mathrm{FIM}(\theta^{(i)})}.\end{aligned}$$

- Note that, if $\mathrm{FIM}(\theta^{(i)}) = I$, we recover the standard squared Euclidean distance.

# Natural Gradient Descent

- Overall, natural gradient descent minimizes at each iteration

$$\tilde{g}_{\gamma}(\theta;\theta^{(i)}) = \underbrace{g(\theta^{(i)}) + \frac{dg(\theta^{(i)})}{d\theta}(\theta-\theta^{(i)})}_{\text{first-order Taylor approximation}} + \underbrace{\frac{1}{2\gamma}||\theta-\theta^{(i)}||^2_{\mathrm{FIM}(\theta^{(i)})}}_{\text{proximity penalty}}.$$

- This yields the update

$$\theta^{(i+1)} = \theta^{(i)} - \gamma(\mathrm{FIM}(\theta^{(i)}))^{-1}\nabla g(\theta^{(i)}).$$

- Note that this applies to any distribution parametrized by a vector $\theta$ and not only to the exponential family.
- Natural gradient descent is hence generally more complex because it requires to invert the FIM.
- However, by operating using a metric that is tailored to the space of distributions, it generally allows to use of larger step size, potentially reducing the number of iterations needed to obtain a desirable value of the cost function.

# Natural Gradient Descent for the Exponential Family

- But the natural-gradient update simplifies for minimal exponential-family distributions.
- To see this, consider natural gradient descent with respect to the natural parameters, whose update is given as

$$\eta^{(i+1)} = \eta^{(i)} - \gamma(\mathrm{FIM}(\eta^{(i)}))^{-1}\nabla_\eta g(\eta^{(i)}).$$

- As we have seen in the text, we have the equality $(\mathrm{FIM}(\eta^{(i)}))^{-1}\nabla_\eta g(\eta^{(i)}) = \nabla_\mu g(\mu^{(i)})$, which yields the equivalent simplified update

$$\eta^{(i+1)} = \eta^{(i)} - \gamma\nabla_\mu g(\mu^{(i)}).$$

- So, natural gradient descent on the natural parameters follows the gradient with respect to the mean parameters.

# Mirror Descent

- In contrast to natural gradient descent, mirror descent minimizes at each iteration the convex approximation

$$\tilde{g}_\gamma(\theta;\theta^{(i)}) = \underbrace{g(\theta^{(i)}) + \nabla g(\theta^{(i)})^T(\theta - \theta^{(i)})}_{\text{first-order Taylor approximation}} + \underbrace{\frac{1}{\gamma}\mathrm{KL}(p(x|\theta)||p(x|\theta^{(i)}))}_{\text{proximity penalty}}.$$

- Note that this applies to any distribution parametrized by a vector $\theta$ and not only to the exponential family.
- Consider the important case of a categorical distribution

$$\tilde{g}_\gamma(\mu;\mu^{(i)}) = \underbrace{g(\mu^{(i)}) + \nabla g(\mu^{(i)})^T(\mu - \mu^{(i)})}_{\text{first-order Taylor approximation}} + \underbrace{\frac{1}{\gamma}\mathrm{KL}(\mathrm{Cat}(x|\mu)||\mathrm{Cat}(x|\mu^{(i)}))}_{\text{proximity penalty}}.$$

- This yields the exponentiated gradient update

$$\mu_k^{(i+1)} = \frac{\mu_k^{(i)} \exp\left(-\gamma[\nabla g(\mu^{(i)})]_k\right)}{\sum_{c=0}^{K-1} \mu_c^{(i)} \exp\left(-\gamma[\nabla g(\mu^{(i)})]_c\right)}.$$