

# CS 3750 Machine Learning

## Lecture 3

### Density estimation II

Milos Hauskrecht  
[milos@cs.pitt.edu](mailto:milos@cs.pitt.edu)  
5329 Sennott Square

---

CS 3750 Advanced Machine Learning

### Outline

**Outline:**

- **Density estimation:**
  - Maximum likelihood (ML)
  - Bayesian parameter estimates
  - MAP
- **Bernoulli distribution.**
- **Binomial distribution**
- **Multinomial distribution**
- **Normal distribution**

---

CS 3750 Advanced Machine Learning

## Density estimation

**Data:**  $D = \{D_1, D_2, \dots, D_n\}$   
 $D_i = \mathbf{x}_i$  a vector of attribute values

### Attributes:

- modeled by random variables  $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$  with:

- **Continuous values**

- **Discrete values**

E.g. *blood pressure* with numerical values

or *chest pain* with discrete values

[no-pain, mild, moderate, strong]

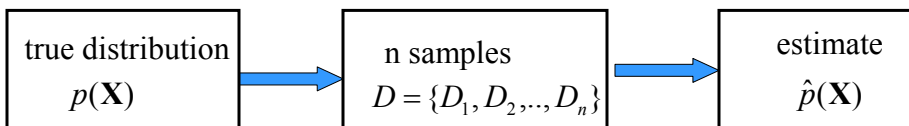
### Underlying true probability distribution:

$$p(\mathbf{X})$$

## Density estimation

**Data:**  $D = \{D_1, D_2, \dots, D_n\}$   
 $D_i = \mathbf{x}_i$  a vector of attribute values

**Objective:** try to estimate the underlying ‘true’ probability distribution over variables  $\mathbf{X}$ ,  $p(\mathbf{X})$ , using examples in  $D$



### Standard (iid) assumptions: Samples

- are **independent** of each other
- come from the same (**identical**) **distribution** (fixed  $p(\mathbf{X})$ )

## Density estimation

### Types of density estimation:

#### Parametric

- the distribution is modeled using a set of parameters  $\Theta$

$$p(\mathbf{X} | \Theta)$$

- Example:** mean and covariances of a multivariate normal
- Estimation:** find parameters  $\Theta$  describing data  $D$

#### Non-parametric

- The model of the distribution utilizes all examples in  $D$
- As if all examples were parameters of the distribution
- Examples:** Nearest-neighbor

## Parameter estimation

- Maximum likelihood (ML)**

maximize  $p(D | \Theta, \xi)$

- yields: one set of parameters  $\Theta_{ML}$
- the target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | \Theta_{ML})$$

- Bayesian parameter estimation**

- uses the posterior distribution over possible parameters

$$p(\Theta | D, \xi) = \frac{p(D | \Theta, \xi) p(\Theta | \xi)}{p(D | \xi)}$$

- Yields: all possible settings of  $\Theta$  (and their “weights”)
- The target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | D) = \int_{\Theta} p(\mathbf{X} | \Theta) p(\Theta | D, \xi) d\Theta$$

## Parameter estimation

### Other possible criteria:

- **Maximum a posteriori probability (MAP)**

maximize  $p(\Theta | D, \xi)$  (mode of the posterior)

– Yields: one set of parameters  $\Theta_{MAP}$

– Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | \Theta_{MAP})$$

- **Expected value of the parameter**

$\hat{\Theta} = E(\Theta)$  (mean of the posterior)

– Expectation taken with regard to posterior  $p(\Theta | D, \xi)$

– Yields: one set of parameters

– Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | \hat{\Theta})$$

## Binomial distribution

**Example problem:** a biased coin

**Outcomes:** two possible values -- head or tail

**Data:**  $D$  a set of order-independent outcomes

**We treat  $D$  as a multi-set !!!**

$N_1$  - number of heads seen     $N_2$  - number of tails seen

**Model:** probability of a head     $\theta$   
probability of a tail     $(1 - \theta)$

**Objective:**

We would like to estimate the probability of a head  $\hat{\theta}$

**Probability of an outcome**

$$P(D | \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N_2} \quad \text{Binomial distribution}$$

## Maximum likelihood (ML) estimate.

### Likelihood of data:

$$P(D | \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N_2} = \frac{N!}{N_1! N_2!} \theta^{N_1} (1 - \theta)^{N_2}$$

### Log-likelihood

$$l(D, \theta) = \log \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N_2} = \log \frac{N!}{N_1! N_2!} + N_1 \log \theta + N_2 \log(1 - \theta)$$

Constant from the point of optimization !!!

$$\text{ML Solution: } \theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

The same as for an iid sequence of Bernoulli trials

## Posterior density

### Posterior density

$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) p(\theta | \xi)}{P(D | \xi)} \quad (\text{via Bayes rule})$$

### Prior choice

$$p(\theta | \xi) = \text{Beta}(\theta | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_2 - 1}$$

### Likelihood

$$P(D | \theta) = \frac{\Gamma(N_1 + N_2)}{\Gamma(N_1) \Gamma(N_2)} \theta^{N_1} (1 - \theta)^{N_2}$$

### Posterior

$$p(\theta | D, \xi) = \text{Beta}(\alpha_1 + N_1, \alpha_2 + N_2)$$

### MAP estimate

$$\theta_{MAP} = \arg \max_{\theta} p(\theta | D, \xi)$$
$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

## Expected value of the parameter

The result is the same as for the Bernoulli distribution

$$E(\theta) = \int_0^1 \theta \text{Beta}(\theta | \eta_1, \eta_2) d\theta = \frac{\eta_1}{\eta_1 + \eta_2}$$

Expected value of the parameter

$$E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

Predictive probability of event  $x=1$

$$P(x = 1 | \theta, \xi) = E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

## Multinomial distribution

**Example: Multi-way coin toss, roll of a dice**

- **Data:** a set of  $N$  outcomes (multi-set)

$N_i$  - a number of times an outcome  $i$  has been seen

**Model parameters:**  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  s.t.  $\sum_{i=1}^k \theta_i = 1$   
 $\theta_i$  - probability of an outcome  $i$

**Probability of data** (likelihood)

$$P(N_1, N_2, \dots, N_k | \theta, \xi) = \frac{N!}{N_1! N_2! \dots N_k!} \theta_1^{N_1} \theta_2^{N_2} \dots \theta_k^{N_k} \quad \text{Multinomial distribution}$$

**ML estimate:**

$$\theta_{i,ML} = \frac{N_i}{N}$$

## Posterior density and MAP estimate

**Choice of the prior: Dirichlet distribution**

$$Dir(\boldsymbol{\theta} | \alpha_1, \dots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}$$

**Dirichlet is the conjugate choice for multinomial**

$$P(D | \boldsymbol{\theta}, \xi) = P(N_1, N_2, \dots, N_k | \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \dots N_k!} \theta_1^{N_1} \theta_2^{N_2} \dots \theta_k^{N_k}$$

**Posterior density**

$$p(\boldsymbol{\theta} | D, \xi) = \frac{P(D | \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} | \alpha_1, \alpha_2, \dots, \alpha_k)}{P(D | \xi)} = Dir(\boldsymbol{\theta} | \alpha_1 + N_1, \dots, \alpha_k + N_k)$$

**MAP estimate:**

$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1, \dots, k} (\alpha_i + N_i) - k}$$

CS 3750 Advanced Machine Learning

## Expected value

**The result is analogous to the result for binomial**

$$E(\boldsymbol{\theta}) = \int_{0 \leq \theta_i \leq 1, \sum \theta_i = 1} \boldsymbol{\theta} Dir(\boldsymbol{\theta} | \boldsymbol{\eta}) d\boldsymbol{\theta} = \left( \frac{\eta_1}{\eta_1 + \eta_2 + \eta_k}, \dots, \frac{\eta_i}{\eta_1 + \eta_2 + \eta_k}, \dots, \frac{\eta_k}{\eta_1 + \eta_2 + \eta_k} \right)$$

**Expectation based parameter estimate**

$$E(\boldsymbol{\theta}) = \left( \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \dots + \alpha_k + N_k}, \dots, \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \dots + \alpha_k + N_k}, \dots, \frac{\alpha_k + N_k}{\alpha_1 + N_1 + \dots + \alpha_k + N_k} \right)$$

**Represents the predictive probability** of an event  $x=i$

$$P(x=i | \boldsymbol{\theta}, \xi) = \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \dots + \alpha_k + N_k}$$

CS 3750 Advanced Machine Learning

## Other distributions

### The same ideas can be applied to other distributions

- Typically we choose distributions that behave well so that computations lead to “nice” solutions

- **Exponential family of distributions**

**Conjugate choices** for some of the distributions from the exponential family:

- **Binomial – Beta**
- **Multinomial - Dirichlet**
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

## Other distributions

### Gamma distribution:

$$p(x | a, b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}} \quad \text{for } x \in [0, \infty]$$

### Exponential distribution:

- A special case of Gamma for  $a=1$

$$p(x | b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}}$$

### Poisson distribution:

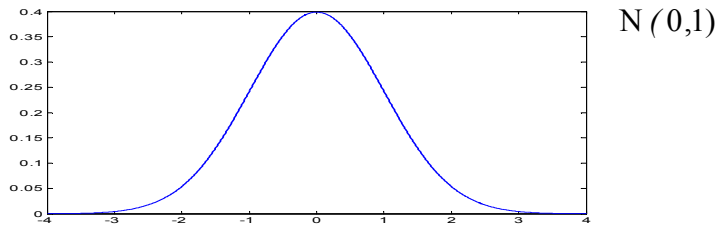
$$p(x | \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x \in \{0, 1, 2, \dots\}$$

## Gaussian (normal) distribution

- **Gaussian:**  $x \sim N(\mu, \sigma)$
- **Parameters:**  $\mu$  - mean  
 $\sigma$  - standard deviation
- **Density function:**

$$p(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right]$$

- **Example:**



CS 3750 Advanced Machine Learning

## Parameter estimates

- **Loglikelihood**  $l(D, \mu, \sigma) = \log \prod_{i=1}^n p(x_i | \mu, \sigma)$

- **ML estimates of the mean and variance:**

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

– ML variance estimate is biased

$$E_n(\hat{\sigma}^2) = E_n\left(\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2\right) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

- **Unbiased estimate:**

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

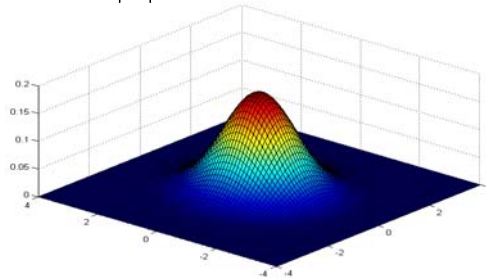
CS 3750 Advanced Machine Learning

## Multivariate normal distribution

- **Multivariate normal:**  $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Parameters:**  $\boldsymbol{\mu}$ - mean  
 $\boldsymbol{\Sigma}$ - covariance matrix
- **Density function:**

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

- **Example:**



CS 3750 Advanced Machine Learning

## Parameter estimates

- **Loglikelihood**  $l(D, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{i=1}^n p(\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$

- **ML estimates of the mean and covariances:**

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

– Covariance estimate is biased

$$E_n(\hat{\boldsymbol{\Sigma}}) = E_n \left( \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T \right) = \frac{n-1}{n} \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

- **Unbiased estimate:**

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

CS 3750 Advanced Machine Learning

## Posterior of a multivariate normal

- Assume a prior on the mean  $\boldsymbol{\mu}$  that is normally distributed:

$$p(\boldsymbol{\mu}) \approx N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

- Then the posterior of  $\boldsymbol{\mu}$  is normally distributed

$$\begin{aligned} p(\boldsymbol{\mu} | D) &\approx \left( \prod_{i=1}^n \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right) \\ &\quad * \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_p|^{1/2}} \exp \left[ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) \right] \\ &= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp \left[ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right] \end{aligned}$$

CS 3750 Advanced Machine Learning

## Posterior of a multivariate normal

- Then the posterior of  $\boldsymbol{\mu}$  is normally distributed

$$p(\boldsymbol{\mu} | D) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp \left[ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_p^{-1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_p \left( \boldsymbol{\Sigma}_p + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right) + \frac{1}{n} \boldsymbol{\Sigma} \left( \boldsymbol{\Sigma}_p + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_p$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_p \left( \boldsymbol{\Sigma}_p + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \frac{1}{n} \boldsymbol{\Sigma}$$

CS 3750 Advanced Machine Learning

## Exponential family

### Exponential family:

- all probability mass / density functions that can be written in the exponential normal form

$$f(\mathbf{x} | \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp[\boldsymbol{\eta}^T t(\mathbf{x})]$$

- $\boldsymbol{\eta}$  a vector of natural (or canonical) parameters
- $t(\mathbf{x})$  a function referred to as a sufficient statistic
- $h(\mathbf{x})$  a function of  $\mathbf{x}$  (it is less important)
- $Z(\boldsymbol{\eta})$  a normalization constant (a partition function)

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T t(\mathbf{x})\} d\mathbf{x}$$

- Other common form:

$$f(\mathbf{x} | \boldsymbol{\eta}) = h(\mathbf{x}) \exp[\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})] \quad \log Z(\boldsymbol{\eta}) = A(\boldsymbol{\eta})$$

## Exponential family: examples

- Bernoulli distribution**

$$\begin{aligned} p(x | \pi) &= \pi^x (1 - \pi)^{1-x} \\ &= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\} \\ &= \exp\{\log(1-\pi)\} \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x\right\} \end{aligned}$$

- Exponential family**

$$f(\mathbf{x} | \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp[\boldsymbol{\eta}^T t(\mathbf{x})]$$

- Parameters**

$$\boldsymbol{\eta} = ?$$

$$t(\mathbf{x}) = ?$$

$$Z(\boldsymbol{\eta}) = ?$$

$$h(\mathbf{x}) = ?$$

## Exponential family: examples

- **Bernoulli distribution**

$$\begin{aligned} p(x | \pi) &= \pi^x (1 - \pi)^{1-x} \\ &= \exp \left\{ \log \left( \frac{\pi}{1 - \pi} \right) x + \log(1 - \pi) \right\} \\ &= \exp \{ \log(1 - \pi) \} \exp \left\{ \log \left( \frac{\pi}{1 - \pi} \right) x \right\} \end{aligned}$$

- **Exponential family**

$$f(\mathbf{x} | \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp [\boldsymbol{\eta}^T t(\mathbf{x})]$$

- **Parameters**

$$\boldsymbol{\eta} = \log \frac{\pi}{1 - \pi} \quad (\text{note } \pi = \frac{1}{1 + e^{-\eta}}) \quad t(\mathbf{x}) = x$$

$$Z(\boldsymbol{\eta}) = \frac{1}{1 - \pi} = 1 + e^{\eta} \quad h(\mathbf{x}) = 1$$

CS 3750 Advanced Machine Learning

## Exponential family: examples

- **Univariate Gaussian distribution**

$$\begin{aligned} p(x | \mu, \sigma) &= \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma^2} (x - \mu)^2 \right] \\ &= \frac{1}{2\pi} \exp \left( -\frac{\mu}{2\sigma^2} - \log \sigma \right) \exp \left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 \right\} \end{aligned}$$

- **Exponential family**

$$f(\mathbf{x} | \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(x) \exp [\boldsymbol{\eta}^T t(x)]$$

- **Parameters**

$$\boldsymbol{\eta} = ? \quad t(\mathbf{x}) = ?$$

$$Z(\boldsymbol{\eta}) = ? \quad h(\mathbf{x}) = ?$$

CS 3750 Advanced Machine Learning

## Exponential family: examples

- **Univariate Gaussian distribution**

$$p(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} (x - \mu)^2\right]$$
$$= \frac{1}{2\pi} \exp\left(-\frac{\mu^2}{2\sigma^2} - \log \sigma\right) \exp\left\{\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2\right\}$$

- **Exponential family**  $f(\mathbf{x} | \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(x) \exp[\boldsymbol{\eta}^T t(x)]$

- **Parameters**

$$\boldsymbol{\eta} = \begin{bmatrix} \mu / 2\sigma^2 \\ -1 / 2\sigma^2 \end{bmatrix} \quad t(\mathbf{x}) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$
$$Z(\boldsymbol{\eta}) = \exp\left\{\frac{\mu^2}{2\sigma^2} + \log \sigma\right\} = \exp\left\{-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \log(-2\eta_2)\right\}$$
$$h(\mathbf{x}) = 1/\sqrt{2\pi}$$

## Exponential family

- **For iid samples, the likelihood of data is**

$$P(D | \boldsymbol{\eta}) = \prod_{i=1}^n p(\mathbf{x}_i | \boldsymbol{\eta}) = \prod_{i=1}^n h(\mathbf{x}_i) \exp[\boldsymbol{\eta}^T t(\mathbf{x}_i) - A(\boldsymbol{\eta})]$$
$$= \left[ \prod_{i=1}^n h(\mathbf{x}_i) \right] \exp\left[ \sum_{i=1}^n \boldsymbol{\eta}^T t(\mathbf{x}_i) - nA(\boldsymbol{\eta}) \right]$$
$$= \left[ \prod_{i=1}^n h(\mathbf{x}_i) \right] \exp\left[ \boldsymbol{\eta}^T \left( \sum_{i=1}^n t(\mathbf{x}_i) \right) - nA(\boldsymbol{\eta}) \right]$$

- **Important:**

- the dimensionality of the sufficient statistic remains the same with the number of samples

## Exponential family

- The log likelihood of data is

$$\begin{aligned}l(D, \boldsymbol{\eta}) &= \log \left[ \prod_{i=1}^n h(\mathbf{x}_i) \right] \exp \left[ \boldsymbol{\eta}^T \left( \sum_{i=1}^n t(\mathbf{x}_i) \right) - nA(\boldsymbol{\eta}) \right] \\ &= \log \left[ \prod_{i=1}^n h(\mathbf{x}_i) \right] + \left[ \boldsymbol{\eta}^T \left( \sum_{i=1}^n t(\mathbf{x}_i) \right) - nA(\boldsymbol{\eta}) \right]\end{aligned}$$

- Optimizing the loglikelihood

$$\nabla_{\boldsymbol{\eta}} l(D, \boldsymbol{\eta}) = \left( \sum_{i=1}^n t(\mathbf{x}_i) \right) - n \nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \mathbf{0}$$

- For the ML estimate it must hold

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \frac{1}{n} \left( \sum_{i=1}^n t(\mathbf{x}_i) \right)$$

## Exponential family

- Rewriting the gradient:

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \log Z(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \log \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T t(\mathbf{x}) \} d\mathbf{x}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \frac{\int t(\mathbf{x}) h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T t(\mathbf{x}) \} d\mathbf{x}}{\int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T t(\mathbf{x}) \} d\mathbf{x}}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \int t(\mathbf{x}) h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta}) \} d\mathbf{x}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = E(t(\mathbf{x}))$$

- Result: 
$$E(t(\mathbf{x})) = \frac{1}{n} \left( \sum_{i=1}^n t(\mathbf{x}_i) \right)$$

- For the ML estimate the parameters  $\boldsymbol{\eta}$  should be adjusted such that the expectation of the statistic  $t(\mathbf{x})$  is equal to the observed sample statistics

## Moments of the distribution

- **For the exponential family**

- The k-th moment of the statistic corresponds to the k-th derivative of  $A(\boldsymbol{\eta})$
- If  $x$  is a component of  $t(x)$  then we get the moments of the distribution by differentiating its corresponding natural parameter

- **Example: Bernoulli**  $p(x | \pi) = \exp \left\{ \log \left( \frac{\pi}{1 - \pi} \right) x + \log(1 - \pi) \right\}$

$$A(\boldsymbol{\eta}) = \log \frac{1}{1 - \pi} = \log(1 + e^\eta)$$

- **Derivatives:**

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta} = \frac{\partial}{\partial \eta} \log(1 + e^\eta) = \frac{e^\eta}{(1 + e^\eta)} = \frac{1}{(1 + e^{-\eta})} = \pi$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta^2} = \frac{\partial}{\partial \eta} \frac{1}{(1 + e^{-\eta})} = \pi(1 - \pi)$$