CS 2750 Machine Learning
Lecture 5

Density estimation II.

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Announcements

- **Homework 1 in**
- **Homework 2 out**
  - Due on Wednesday before the class
  - **Reports:** hand in before the class
  - **Programs:** submit electronically

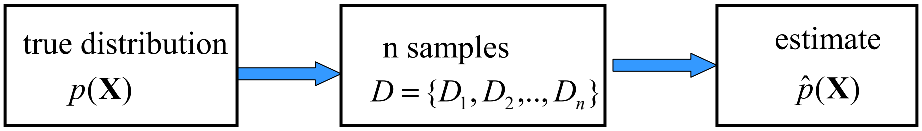
- **Collaborations on homeworks:**
  - You may discuss material with your fellow students, but the report and programs should be written individually

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ 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



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

n samples $D = \{D_1, D_2, .., D_n\}$

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

Standard (iid) assumptions: Samples

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

Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in $\boldsymbol{X}$ with parameters Θ :
$$\hat{p}(\mathbf{X} \mid \Theta)$$
- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: find the description of parameters Θ so they fit the observed data

Parameter estimation.

- **Maximum likelihood (ML)**

  maximize $p(D \mid \Theta, \xi)$
  - yields: one set of parameters Θ_{ML}
  - the target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{ML})$$

- **Bayesian parameter estimation**
  - uses the posterior distribution over possible parameters

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

  - Yields: all possible settings of Θ (and their "weights")
  - The target distribution is approximated as:

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

Parameter estimation.

Other possible criteria:

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

  maximize $p(\mathbf{\Theta} \mid D, \xi)$ (mode of the posterior)
  - Yields: one set of parameters $\mathbf{\Theta}_{MAP}$
  - Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{MAP})$$

- **Expected value of the parameter**

  $\hat{\mathbf{\Theta}} = E(\mathbf{\Theta})$ (mean of the posterior)
  - Expectation taken with regard to posterior $p(\mathbf{\Theta} \mid D, \xi)$
  - Yields: one set of parameters
  - Approximation:

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

Multinomial distribution

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

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

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

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

θ_i - probability of an outcome i

Probability of data (likelihood)

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

Multinomial distribution

ML estimate:

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

Bayesian estimate

Choice of the prior: **Dirichlet distribution**

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

Dirichlet is a conjugate choice for the multinomial

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

Posterior density

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

Hyper-parameters defining the posterior density:

$$\alpha_1' = \alpha_1 + N_1, \alpha_2' = \alpha_2 + N_2, .., \alpha_k' = \alpha_k + N_k$$

MAP estimate

Posterior density

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

MAP estimate:

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

Expected value

The result is analogous to the result for binomial

$$E(\boldsymbol{\theta}) = \int_{0 \le \theta_i \le 1, \sum \theta_i = 1} \boldsymbol{\theta} \; Dir(\boldsymbol{\theta} \mid \boldsymbol{\eta}) d\boldsymbol{\theta} = \left(\frac{\eta_1}{\eta_1 + \eta_2 + \eta_k}, \ldots \frac{\eta_i}{\eta_1 + \eta_2 + \eta_k}, \ldots \frac{\eta_k}{\eta_1 + \eta_2 + \eta_k} \right)$$

Expectation based parameter estimate

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

Directly represents the predictive probability of an event x=i

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

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 –Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

Other distributions

Gamma distribution:

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

Exponential distribution:

- A special case of Gamma for a=1

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

Poisson distribution:

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

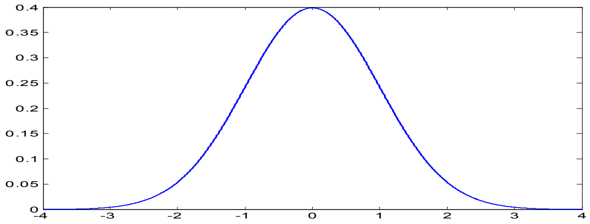
Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean

  σ - standard deviation
- **Density function:**

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

- **Example:**



$N(0,1)$

Parameter estimates

- **Loglikelihood** $$l(D, \mu, \sigma) = \log \prod_{i=1}^{n} p(x_i \mid \mu, \sigma)$$
- **ML estimates of the mean and variance:**

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

  – ML variance estimate is biased

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

- **Unbiased estimate:**

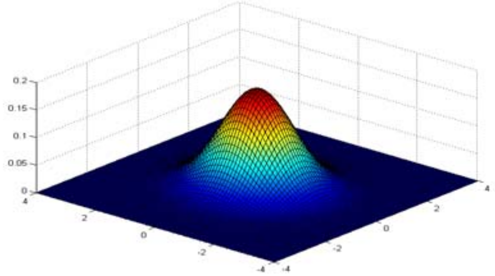
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Multivariate normal distribution

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

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

- **Example:**



Parameter estimates

- **Loglikelihood** $l(D, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **ML estimates of the mean and covariances:**

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

  – 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:**

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

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**

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

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

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

Posterior of a multivariate normal

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

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

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

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

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

Recursive Bayesian parameter estimation.

- **Recursive Bayesian approach**
  - Estimates of the posterior can be sometimes computed incrementally for a sequence of data points

$$p(\Theta \mid D,\xi) = \frac{p(D \mid \Theta,\xi)\, p(\Theta \mid \xi)}{\int_{\Theta} p(D \mid \Theta,\xi)\, p(\Theta \mid \xi)\, d\Theta}$$

- If we use a conjugate prior we get back the same posterior
- Assume we split the data D in the last element x and the rest $p(D \mid \boldsymbol{\Theta}) = P(x \mid \boldsymbol{\Theta}) P(D_{n-1} \mid \boldsymbol{\Theta})$
- **Then:**

$$p(\Theta \mid D,\xi) = \frac{P(x \mid \boldsymbol{\Theta}) \overbrace{P(D_{n-1} \mid \boldsymbol{\Theta})\, p(\Theta \mid \xi)}^{\textbf{A “new” prior}}}{\int_{\Theta} P(x \mid \boldsymbol{\Theta}) P(D_{n-1} \mid \boldsymbol{\Theta})\, p(\Theta \mid \xi)\, d\Theta}$$

Exponential family

Exponential family:

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

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

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

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

- Other common form:

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})\right] \qquad \log Z(\boldsymbol{\eta}) = A(\boldsymbol{\eta})$$

Exponential family: examples

- **Bernoulli distribution**

$$p(x \mid \pi) = \pi^{x}(1-\pi)^{1-x}$$

$$= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}$$

$$= \exp\{\log(1-\pi)\}\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x\right\}$$

- **Exponential family**

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

- **Parameters**

$\boldsymbol{\eta} = ?$

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

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

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

Exponential family: examples

- **Bernoulli distribution**

$$p(x \mid \pi) = \pi^{x}(1-\pi)^{1-x}$$

$$= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}$$

$$= \exp\{\log(1-\pi)\}\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x\right\}$$

- **Exponential family**

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

- **Parameters**

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

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

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

$h(\mathbf{x}) = 1$

Exponential family: examples

- **Univariate Gaussian distribution**

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

$$= \frac{1}{2\pi} \exp\left(-\frac{\mu}{2\sigma^2} - \log\sigma\right) \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right\}$$

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

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

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

Exponential family: examples

- **Univariate Gaussian distribution**

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

$$= \frac{1}{2\pi} \exp\left(-\frac{\mu}{2\sigma^2} - \log\sigma\right) \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right\}$$

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

$$\boldsymbol{\eta} = \begin{bmatrix} \mu / 2\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \qquad t(\mathbf{x}) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$Z(\boldsymbol{\eta}) = \exp\left\{\frac{\mu}{2\sigma^2} + \log\sigma\right\} = \exp\left\{-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right\}$$

$$h(\mathbf{x}) = 1/\sqrt{2\pi}$$

Exponential family

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

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

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

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

- **Important:**
  - the dimensionality of the sufficient statistic remains the same with the number of samples

---

Exponential family

- **log likelihood of data is**

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

$$= \log\left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] + \left[\boldsymbol{\eta}^T \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

- **Optimizing the loglikelihood**

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

- **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

- **Rewritting the gradient:**

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

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

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

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

- **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(x) is be 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** $p(x \mid \pi) = \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}$

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

- **Derivatives:**

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

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