CS 2750 Machine Learning
Lecture 7

Density estimation III

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Distribution models for random variables

Distribution models covered so far:

- **Bernoulli distribution**
  - **Model for binary random variables**

$$P(x \mid \theta) = \theta^x (1-\theta)^{(1-x)}$$

- **Binomial distribution**
  - **Model for order independent sets of binary outcomes**

$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$$

- **Multinomial distribution**
  - **Model for order independent sets of k-nary outcomes**

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

Distribution models for random variables

Models for other types of random variables:

- **Gaussian distribution**
  - **Models of real-valued random variable**
- **Gamma distribution:**
  - **Models of random variables for positive real numbers**
- **Exponential distribution**
  - **Models of random variables for positive real numbers**
- **Poisson distribution**
  - **Models of random variables for nonnegative integers**

Conjugate choices of priors for some these distributions:

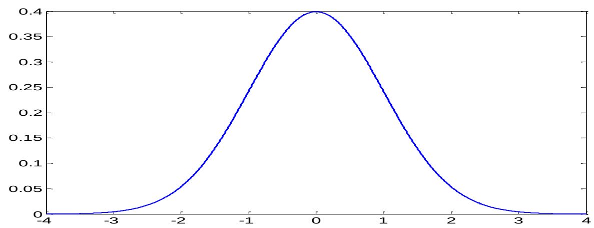
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

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:**

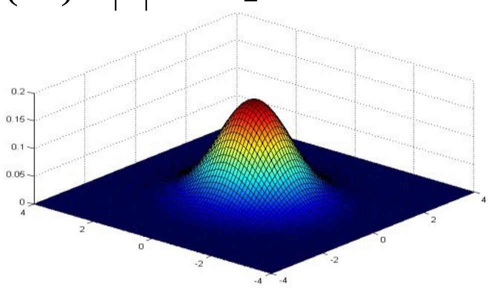
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

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:**



Partitioned Gaussian Distributions

- **Multivariate Gaussian:**

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- **Example:**

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Precision matrix

- **What are the distributions for marginals and conditionals?**

$$p(x_a) \qquad p(x_a \mid x_b)$$

Conditionals and Marginals

- **Conditional density:**

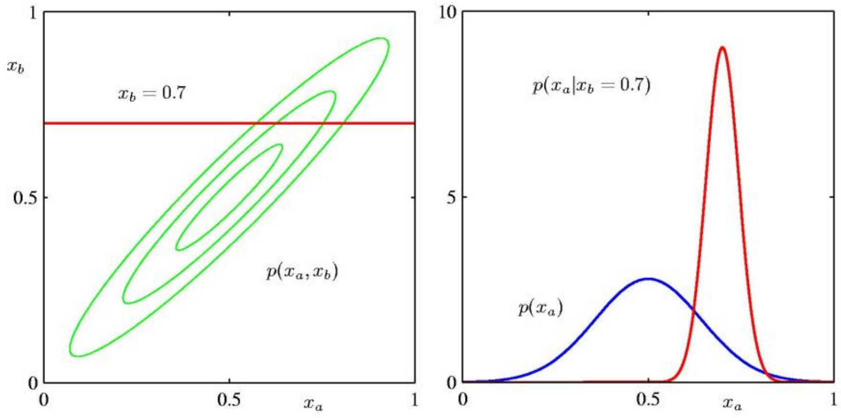
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

- **Marginal Density:**

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

Conditionals and Marginals



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 the mean of a multivariate normal

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

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

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

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

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

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

Posterior of the mean of a multivariate normal

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

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

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

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

$$\boldsymbol{\Sigma}_n = \left(\boldsymbol{\Sigma}_{\mathrm{p}}^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$$

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

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

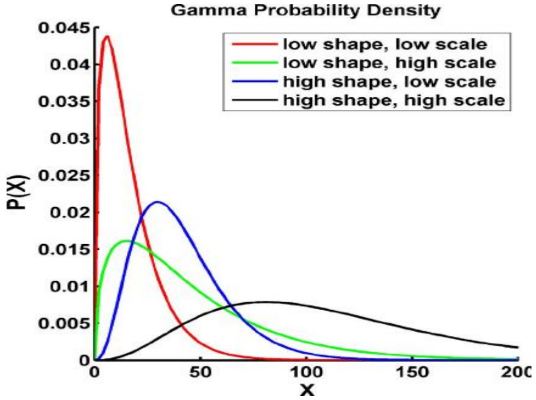
Poisson distribution:

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

Gamma distribution

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

where *a* is the shape and *b* is a scale parameter

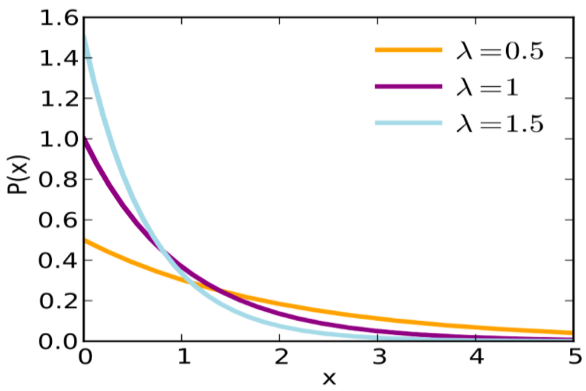


Exponential distribution

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

Alternative parameterization: $p(x \mid \lambda) = \lambda e^{-\lambda x}$

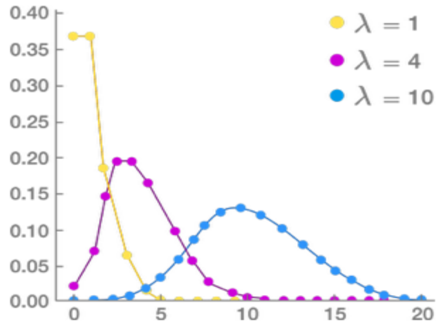
where $\lambda = 1/b$



Poisson distribution

Poisson distribution:

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



Sequential Bayesian parameter estimation

- **Sequential Bayesian approach**
  - Under the iid the estimates of the posterior can be 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

  $p(D \mid \mathbf{\Theta}) = P(x \mid \mathbf{\Theta}) P(D_{n-1} \mid \mathbf{\Theta})$

- **Then:**

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

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 (a partition function)

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

- 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**

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

- **Exponential family**

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

- **Parameters**

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

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

Exponential family: examples

- **Bernoulli distribution**

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

- **Exponential family**

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

- **Parameters**

$\boldsymbol{\eta} = \log\frac{\pi}{1-\pi}$ ← **logit function** $\quad t(\mathbf{x}) = x$

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

Exponential family: examples

- **Univariate Gaussian distribution**

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

- **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**

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

- **Exponential family**

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

- **Parameters**

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

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

- **The 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:**

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 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:**

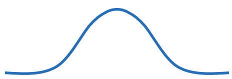
$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta} = \frac{\partial}{\partial \eta} \log(1+e^{\eta}) = \frac{e^{\eta}}{(1+e^{\eta})} = \frac{1}{(1+e^{-\eta})} = \pi$$

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

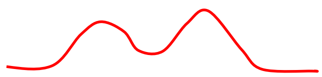
Non-parametric density estimation

Nonparametric Density Estimation

- **Parametric distribution models** are:
  - restricted to specific functional forms, which may not always be suitable;
  - **Example:** modeling a multimodal distribution with a single, unimodal model.



vs



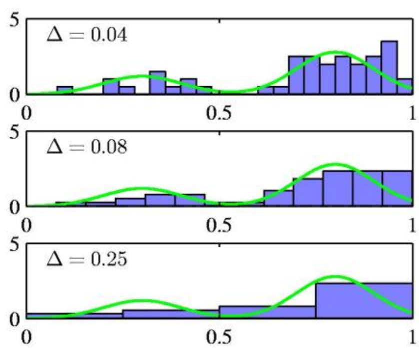
- **Nonparametric approaches:**
  - Do not make any strong assumption about the overall shape of the distribution being modelled.

Nonparametric Methods

Histogram methods:

partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

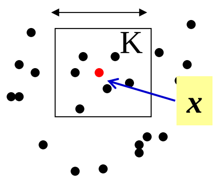


- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.

- Binning does not work well in the in a d-dimensional space,

Nonparametric Methods

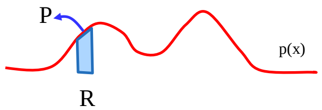
- Binning does not work well in the in a d-dimensional space,
  - M bins in each dimension will require M^d bins!
- **Solution:**
  - Build the estimates of p(**x**) by considering the data points in D and how similar (or close) they are to **x**
  - **Example: Parzen window**
    - As if we build a bin dynamically for **x** for which we need p(**x**)



Nonparametric Methods

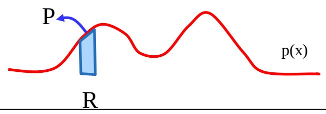
- Assume observations drawn from a density p(x) and consider a small region R containing x such that

$$P = \int_R p(x)dx$$



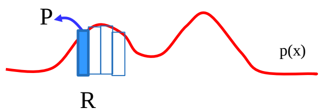
- The probability that K out of N observations lie inside R is *Bin(K,N,P)* and if N is large

$$K \cong NP$$



If the volume of R, *V*, is sufficiently small, p(x) is approximately constant over R and

$$P \cong p(x)V$$



Thus

$$p(x) = \frac{P}{V}$$

Putting things together we get:

$$\boxed{p(x) = \frac{K}{NV}}$$

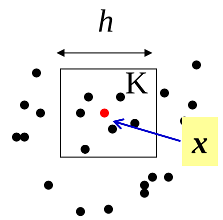
Nonparametric methods: kernel methods

Solution 1: Estimate the probability for **x** based on the fixed volume **V** built around **x**

$$p(x) = \frac{K}{NV}$$

- **Fix V, estimate K from the data**

Example: Parzen window



Nonparametric methods: kernel methods

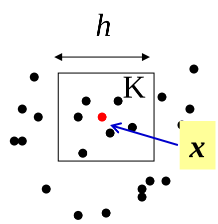
Kernel Density Estimation:

- **Parzen window:** Let R be a hypercube centred on **x** that defines the **kernel function:**

$$k\left(\frac{x-x_n}{h}\right) = \begin{cases} 1 & |(x_i - x_{ni})|/h \leq 1/2 \qquad i = 1, \ldots D \\ 0 & otherwise \end{cases}$$

- **It follows that**

$$K = \sum_{n=1}^{N} k\left(\frac{x-x_n}{h}\right)$$



- **and hence**

$$p(x) = \frac{K}{NV} = \frac{1}{Nh^D} \sum_{n=1}^{N} k\left(\frac{x-x_n}{h}\right)$$

Nonparametric Methods: smooth kernels

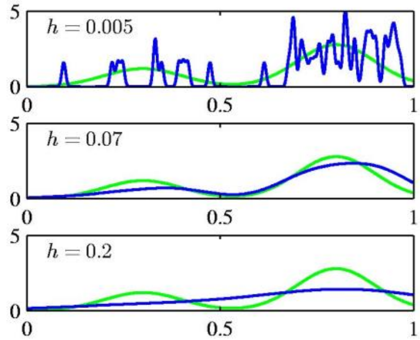
To avoid discontinuities in p(x) because of sharp boundaries we can use a **smooth kernel**, e.g. a Gaussian

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\left(2\pi h^2\right)^{D/2}}\exp\left[-\frac{\|\mathbf{x}-\mathbf{x}_n\|}{2h^2}\right]$$

- Any kernel such that

$$k(\mathbf{u}) \geq 0$$
$$\int k(\mathbf{u})d\mathbf{u} = 1$$

- will work.



h acts as a smoother.

Nonparametric Methods: kNN estimation

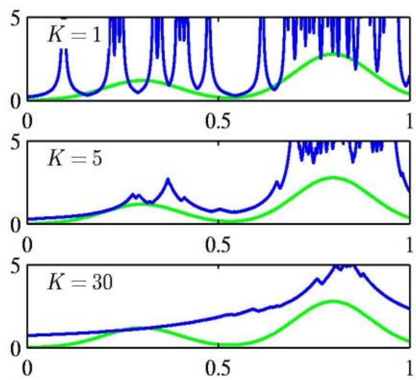
Solution 2: Estimate the probability for **x** based on a fixed count **K** for a variable volume **V** built around **x**

fix K, estimate V from the data

Nearest Neighbour Density Estimation:

Consider a hyper-sphere centred on x and let it grow to a volume, V*, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$



K acts as a smoother

Nonparametric vs Parametric Methods

Nonparametric models:

- More flexibility – no density model is needed
- But require storing the entire dataset
- and the computation is performed with all data examples.

Parametric models:

- Once fitted, only parameters need to be stored
- They are much more efficient in terms of computation
- But the model needs to be picked in advance