

Class Summaries

Classes 1-10

Introduction class

- Statistics is about making inferences from empirical datasets. It is the core of the scientific method.
- Statistics is increasing in importance in engineering as large systems and large data-sets proliferate.
- Histograms are useful for summarizing one dimensional data, but the number and size of bins should be chosen carefully based on the amount of available data and properties of the distribution.

One dimensional distributions

- Computing CDF's requires sorting.
- If the CDF is known, sorting can be done in linear time.
- The Glivenko - Cantelli theorem: a uniform bound on the maximal difference between the true CDF and the empirical CDF.
- CDFs are more reliable than histograms when visualizing one dimensional distributions.
- The Kolmogorov-Smirnoff test compares two 1D samples using the maximal difference between their CDFs.

Means and Variances

$$\text{Var}(X) = E((X - \mu_X)^2) = E(X^2) - E(X)^2$$

If X, Y are independent $E(XY) = E(X)E(Y)$

If S_n is a sum of independent RVs $S_n = \sum_{i=1}^n X_i$

$$E(S_n) = nE(X_i)$$

$$\sigma_n = \sqrt{\text{Var}(S_n)} = \sqrt{n\text{Var}(X_i)}$$

The Binomial Distribution

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Mean: $\mu(S_n) = np$

Standard Deviation: $\sigma(S_n) = \sqrt{np(1-p)}$

Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be IID Random variables with common mean μ and variance σ^2

Define:
$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then the CDF of Z_n converges to the standard normal CDF:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx,$$

In the sense that

$$\forall z, \quad \lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z)$$

Statistical Tests

- **Null Hypothesis** = represents the skeptic: “nothing is new under the sun”
- A statistical test attempts to **reject** the null hypothesis. Non-rejection conveys no information.
- **Significance** of the statistical test = probability that the test rejects the null hypothesis when the null hypothesis is correct.
- **Non-parametric tests** assume only that data drawn IID. - **KS-test**
- require more data.
- **Parametric tests** assume that data drawn IID from distribution of particular form (usually normal). Null hypothesis can make assumptions on parameters (example: bound on variance). - **Chi-square tests, t-tests** ... require less data ... Many variants corresponding to different null hypotheses.

Small and large deviations

Small Deviations $\left| \frac{1}{n} S_n - \mu \right| \approx \frac{c}{\sqrt{n}}$

Small deviations are characterized by
the central limit theorem

Large deviations $\left| \frac{1}{n} S_n - \mu \right| > \epsilon$

Large deviation bounds

$$\hat{p} > q > p \text{ or } \hat{p} < q < p$$

$$P[\hat{p} > q] \leq e^{-m D_{\text{KL}}(q||p)} \quad D_{\text{KL}}(q||p) = q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p}$$

Hoeffding: good for any p

$$P[|\hat{p} - p| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

Chernoff: good when p close to zero or one

if $q \leq p$,

$$D_{\text{KL}}(q||p) \geq \frac{1}{2} \frac{(p - q)^2}{p} \Rightarrow P[\hat{p} < q] \leq e^{-\frac{1}{2}m \frac{(p - q)^2}{p}}$$

if $q \geq p$,

$$D_{\text{KL}}(q||p) \geq \frac{1}{3} \frac{(p - q)^2}{p} \Rightarrow P[\hat{p} > q] \leq e^{-\frac{1}{3}m \frac{(p - q)^2}{p}}$$

PDFs vs CDFs

$$P(X \leq a) = \int_{-\infty}^a p(x) dx$$

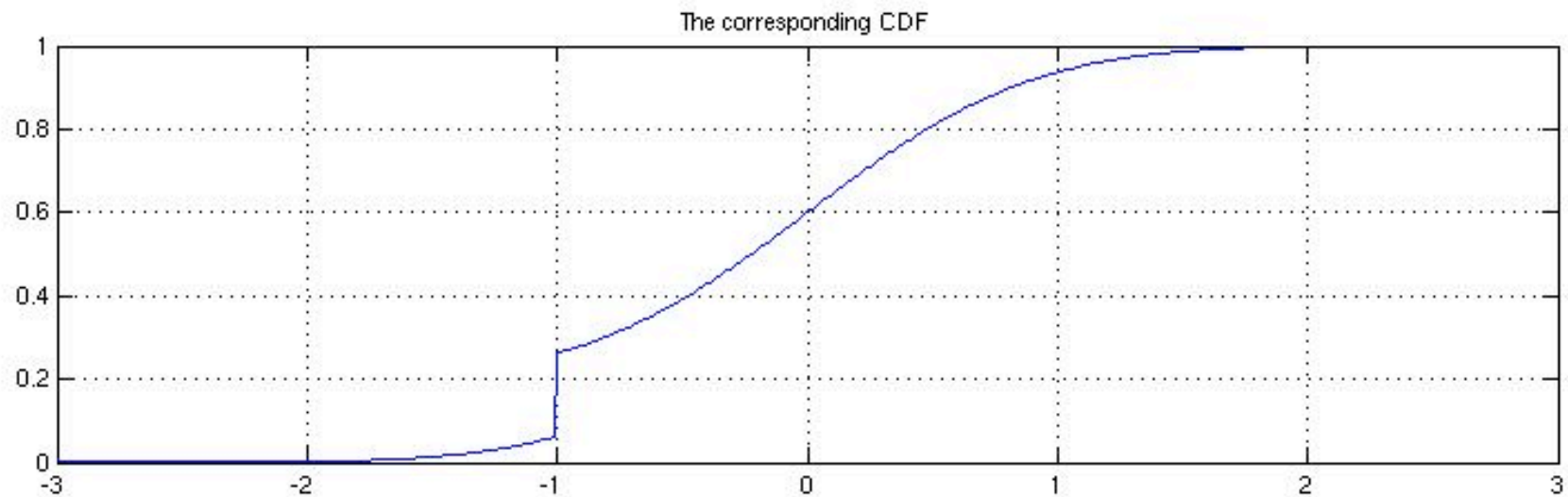
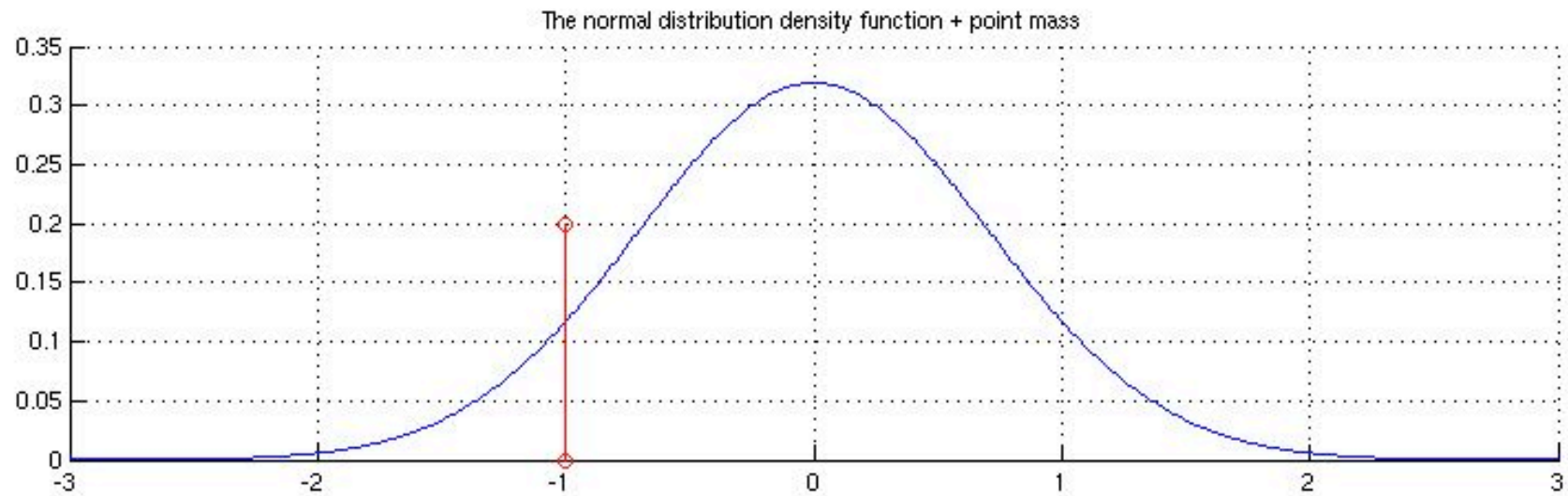
$$p(x) = \frac{d}{da} P(X \leq a)$$

Might be undefined



$P(A)$ is always ≤ 1 , $p(x)$ can be larger than 1.

CDFs are more expressive



Poisson Distribution

- Poisson approximates Binomial for small p
- Number of events per unit time (# cars crossing bridge per minute).

$$P(N_\tau = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$E(N_\tau) = \lambda$$

$$\text{Var}(N_\tau) = \lambda$$

Exponential Distribution

- Time between cars crossing bridge.

$$p(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

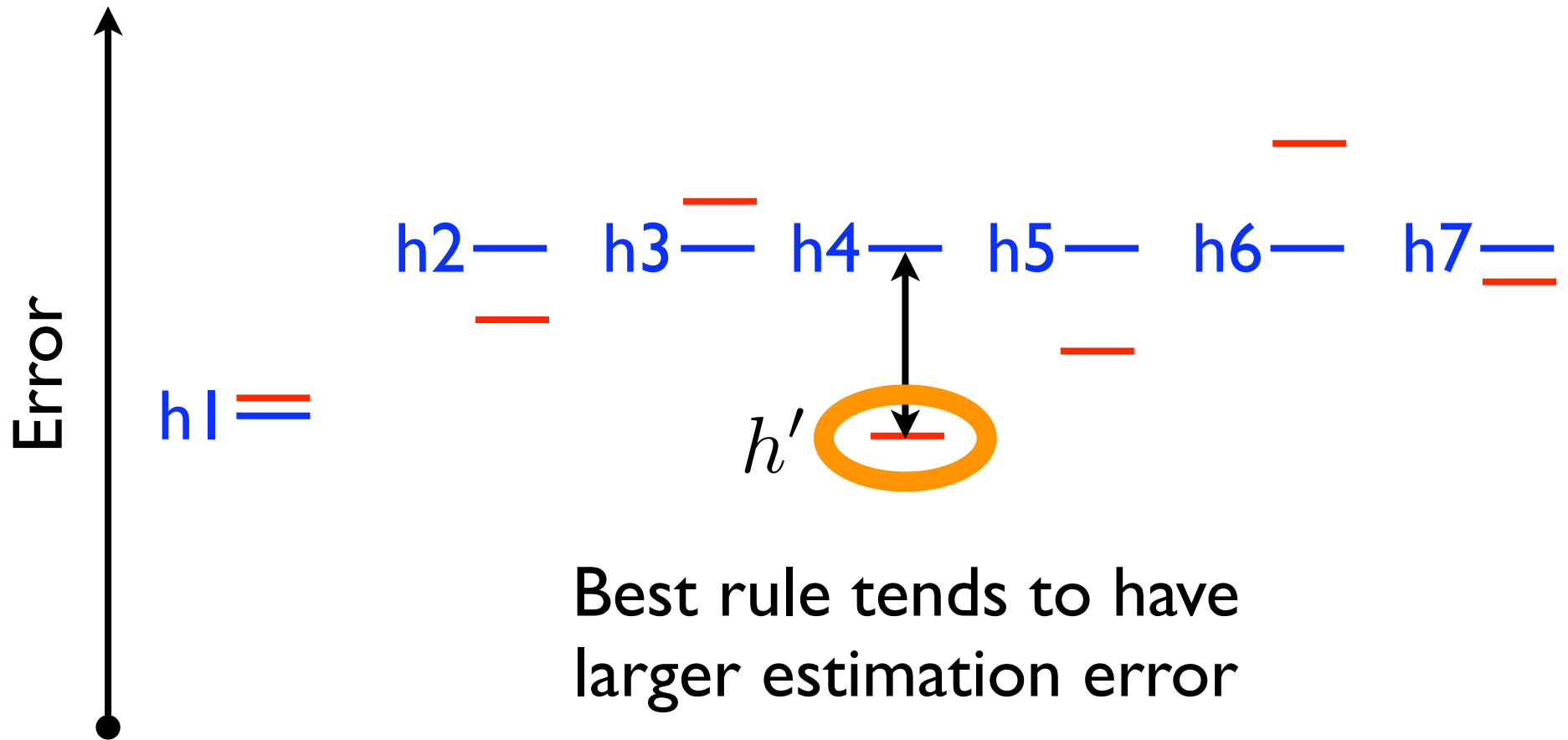
$$E(t) = \int_0^{\infty} t \lambda e^{-\lambda t} dt = \frac{1}{\lambda}$$

$$\text{Var}(t) = \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

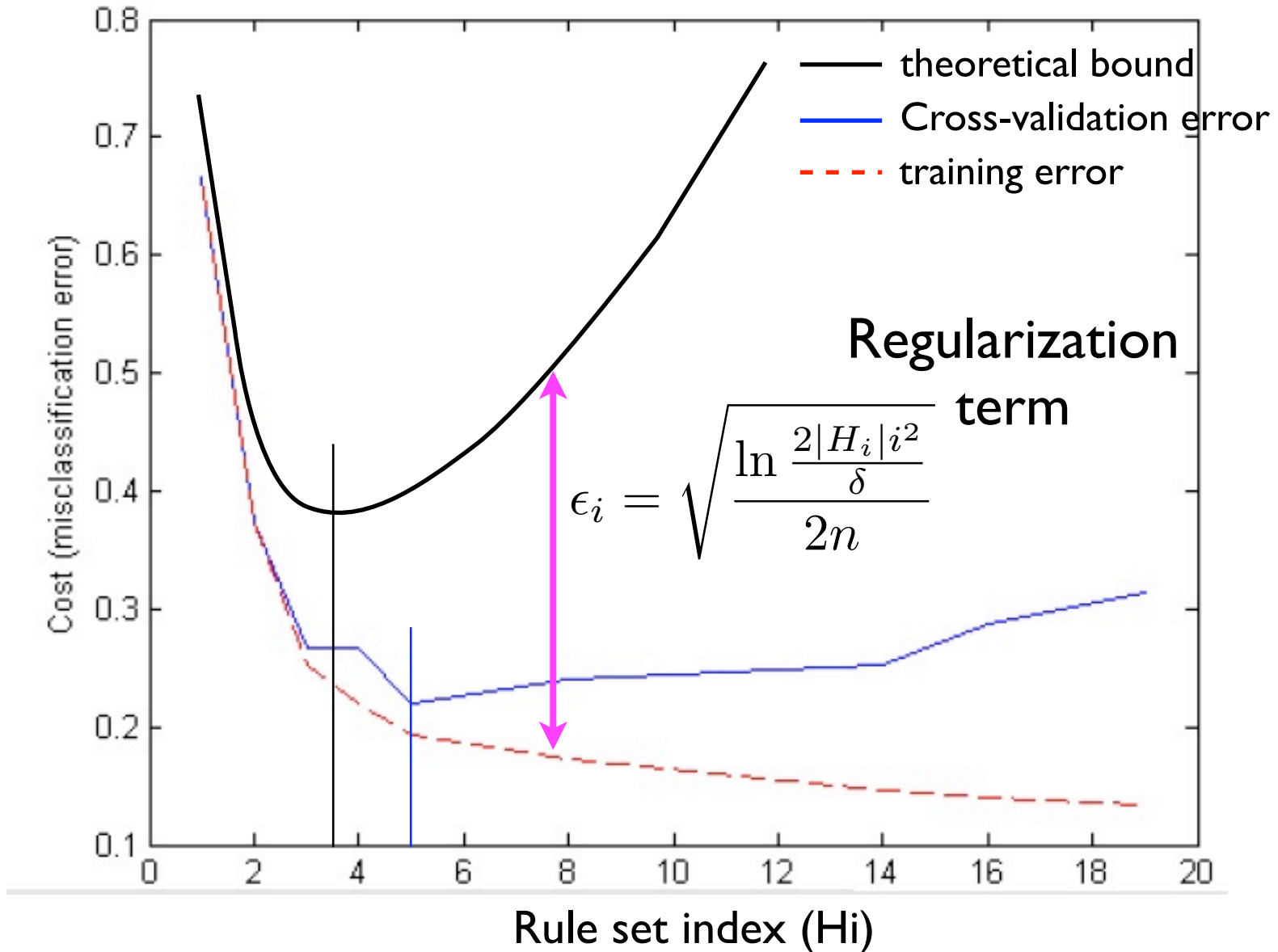
Learning classifiers

- Algorithm finds rule with minimal number of mistakes on training set.
- **Model Selection:** Choosing the rule complexity so as to minimize test error.
- **Validation set:** subset of training set left aside for evaluating learned rule.

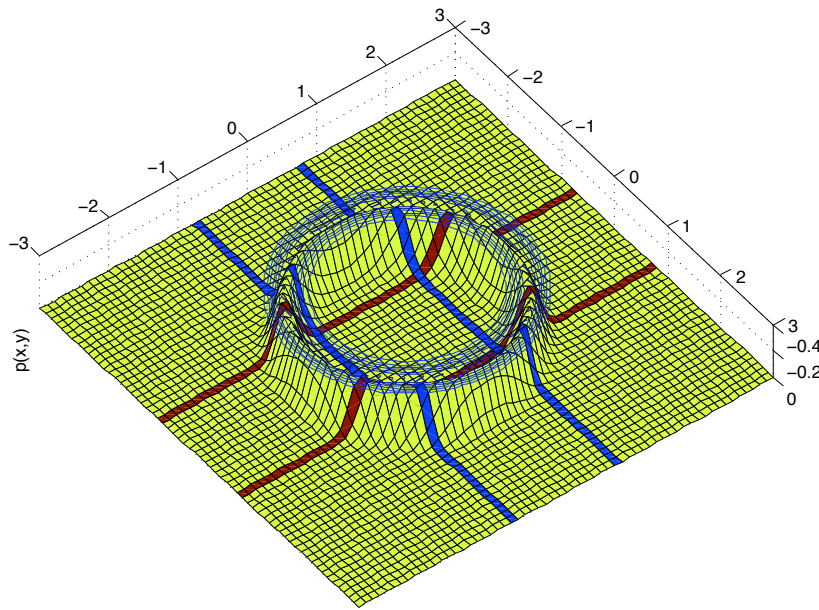
Over-fitting



Occam Rule

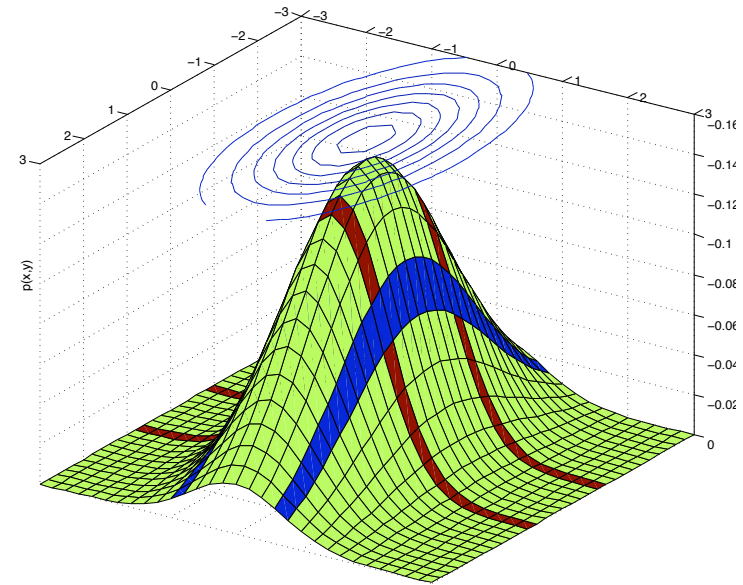


Dependence and correlation



- Uncorrelated
- Dependent

$$\exp\left(\frac{\left(\sqrt{x^2 + y^2} - r\right)^2}{\sigma^2}\right) \neq p(x)q(y)$$



- Uncorrelated
- Independent

$$\exp\left(-\frac{x^2}{\sigma_1^2} - \frac{y^2}{\sigma_2^2}\right) = \exp\left(-\frac{x^2}{\sigma_1^2}\right) \exp\left(-\frac{y^2}{\sigma_2^2}\right)$$