Pattern Recognition

Review of Prerequisites in Math and Statistics

Prepared by
Li Yang

Based on
Appendix chapters of
Pattern Recognition, 4th Ed.
by S. Theodoridis and K. Koutroumbas
and figures from
Wikipedia.org

Probability and Statistics

- **Probability** P(*A*) of an event *A* : a real number between 0 to 1.
- **Joint probability** P(A ∩ B) : probability that both A and B occurs in a single experiment.

  P(A ∩ B) = P(A)P(B) if A and B and **independent**.
- Probability P(A ∪ B) of union of A and B: either A or B occurs in a single experiment.

  P(A ∪ B) = P(A) + P(B) if A and B are mutually exclusive.

- **Conditional probability**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Therefore, the **Bayes rule**:

$$P(A|B)P(B) = P(B|A)P(A) \text{ and } P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- **Total probability:** let $A_1, \ldots, A_m$ such that $\sum_{i=1}^{m} P(A_i) = 1$ then

$$P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i)$$

- **Probability density function (pdf)**: p(x) for a continuous random variable x

$$P(a \le x \le b) = \int_a^b p(x)dx$$

Total and conditional probabilities can also be extended to pdf's.

- **Mean** and **Variance**: let p(x) be the pdf of a random variable x

$$E[x] = \int_{-\infty}^{+\infty} xp(x)dx, \text{ and } \sigma^2 = \int_{-\infty}^{+\infty} (x - E[x])^2 p(x)dx$$

- **Statistical independence**:

$$p(x, y) = p_x(x)p_y(y)$$

- **Kullback-Leibler divergence (Distance?) of pdf's**

$$L(p(\mathbf{x}), p'(\mathbf{x})) = -\int p(\mathbf{x}) \ln \frac{p'(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

Pay attention that $L(p(\mathbf{x}), p'(\mathbf{x})) \ne L(p'(\mathbf{x}), p(\mathbf{x}))$

- **Characteristic function** of a pdf:

$$\Phi(\mathbf{\Omega}) = \int_{-\infty}^{+\infty} p(\mathbf{x}) \exp(j\mathbf{\Omega}^T \mathbf{x}) d\mathbf{x} = E[\exp(j\mathbf{\Omega}^T \mathbf{x})]$$

$$\Phi(s) = \int_{-\infty}^{+\infty} p(x) \exp(sx) dx = E[\exp(sx)]$$

- **2nd Characteristic function:** $\Psi(s) = \ln \Phi(s)$

- **n-th order moment**: $\dfrac{d^n \Phi(0)}{ds^n} = E[x^n]$

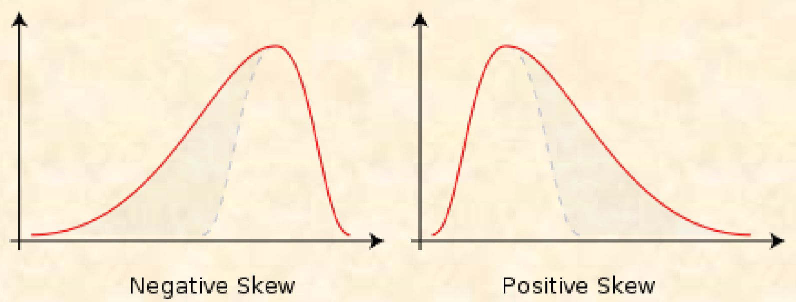
- **Cumulants**: $\kappa_n = \dfrac{d^n \Psi(0)}{ds^n}$

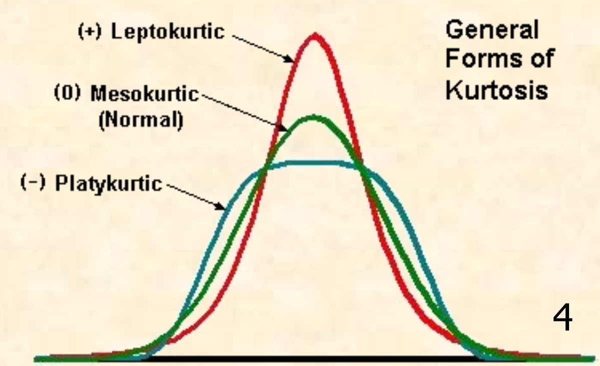
When $E[x] = 0$, then

$\kappa_0 = 0, \kappa_1 = E[x] = 0,$

$\kappa_2 = E[x^2] = \sigma^2, \kappa_3 = E[x^3]$ (Skewness)

$\kappa_4 = E[x^4] - 3\sigma^4$ (Kurtosis)





Discrete Distributions

- **Binomial distribution B(n,p):**

  Repeatedly grab n balls, each with a probability p of getting a black ball. The probability of getting k black balls:

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

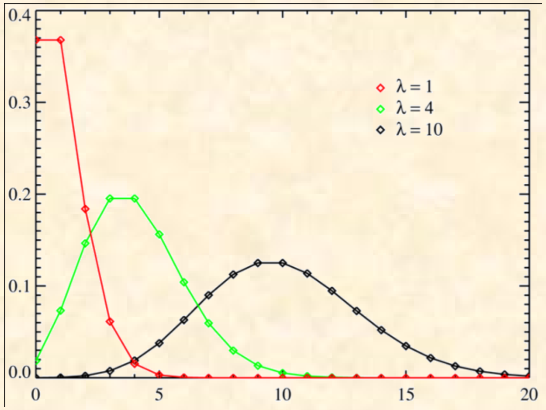
- **Poisson distribution**

  probability of # of events occurring in a fixed period of time if these events occur with a known average.

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

When $n \to \infty$ and np remains constant,

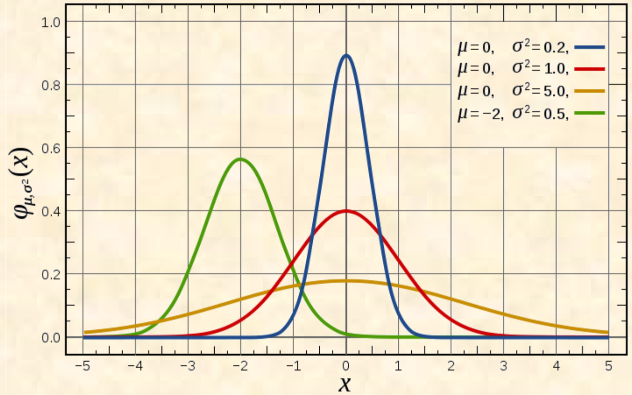
$B(n, p) \to Poisson(np)$



Normal (Gaussian) Distribution

- **Univariate N(μ, σ²):**

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



- **Multivariate N(μ, Σ):**

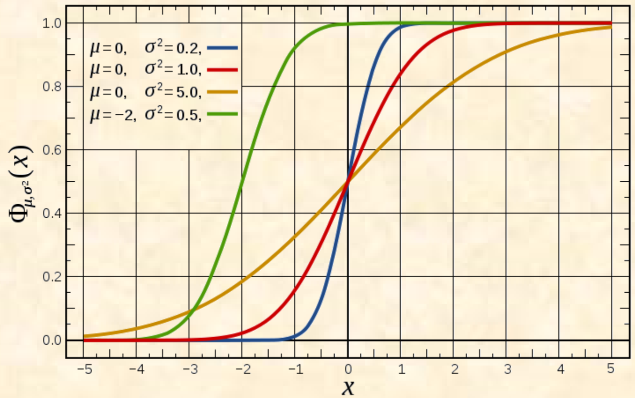
$$p(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

with the mean μ and the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1l} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{l1} & \sigma_{l2} & \cdots & \sigma_l^2 \end{bmatrix}$$

where $\sigma_i^2 = E[(x_i - \mu_i)^2]$ and

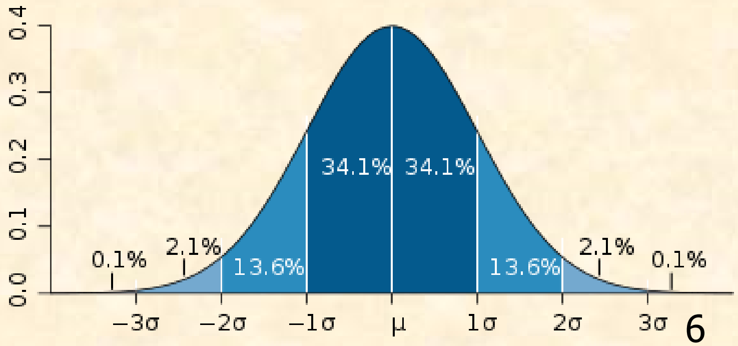
$\sigma_{ij} = \sigma_{ji} = E[(x_i - \mu_i)(x_j - \mu_j)]$



- **Central limit theorem:**

Let $z = \sum_{i=1}^{n} x_i$, then $\frac{z-\mu}{\sigma} \sim N(0,1)$ when $n \to \infty$

irrespective of the pdf's of x_i's.



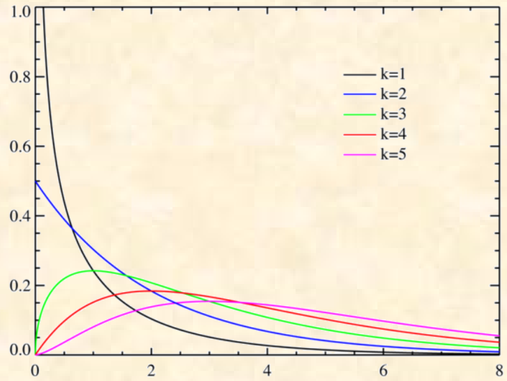
Other Continuous Distributions

- **Chi-square (X^2) distribution** of k degrees of freedom:

  distribution of a sum of squares of *k* independent standard normal random variables, that is, $\chi^2 = x_1^2 + x_2^2 + \cdots + x_k^2$ where $x_i \approx N(0,1)$

$$p(y) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} e^{-y/2} \text{step}(y),$$

$$\text{where } \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$



- Mean: k, Variance: 2k

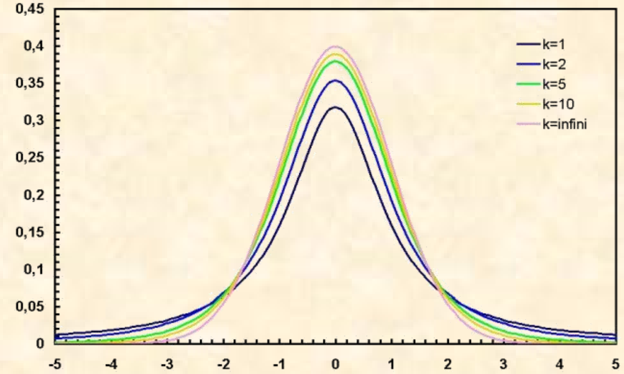
- Assume $x \sim \chi^2(k)$
  - Then $(x-k)/\sqrt{2k} \sim N(0,1)$ as $k \to \infty$ by central limit theorem.
  - Also $\sqrt{2x}$ is approximately normally distributed with mean $\sqrt{2k-1}$ and **unit variance.**

Other Continuous Distributions

- **t-distribution:** estimating mean of a normal distribution when sample size is small.

  A t-distributed variable $q = x/\sqrt{z/k}$ where $x \approx N(0,1)$ and $z \approx \chi^2(k)$

$$p(q) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\,\Gamma(k/2)}\left(1+\frac{q^2}{k}\right)^{-(k+1)/2}$$



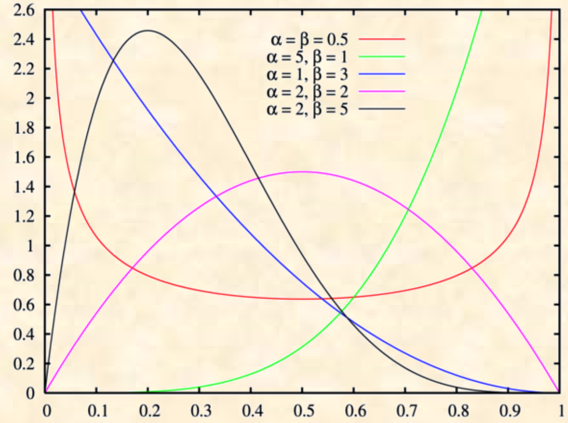
  Mean: 0 for k > 1,

  variance: k/(k-2) for k > 2

- **β-distribution:** Beta(α,β): the posterior distribution of *p* of a binomial distribution after α−1 events with *p* and β − 1 with 1 − *p*.

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$



Linear Algebra

- Eigenvalues and eigenvectors:

  there exists λ and v such that $Av = \lambda v$
- Real matrix A is called *positive semidefinite* if $x^T A x \geq 0$ for **every** nonzero vector x;

  A is called *positive definite* if $x^T A x > 0$.
- Positive definite matrixes act as positive numbers.

  All positive eigenvalues
- If A is symmetric, $A^T = A$,

  then its eigenvectors are orthogonal, $v_i^T v_j = 0$.
- Therefore, a symmetric A can be diagonalized as

$$A = \Phi \Lambda \Phi^T \text{ and } \Phi^T A \Phi = \Lambda$$

$$\text{where } \Phi = [v_1, v_2, \ldots v_l] \text{ and } \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots \lambda_l)$$

Correlation Matrix and Inner Product Matrix

Principal component analysis (PCA)

- Let x be a random variable in R^l, its correlation matrix $\Sigma=E[xx^T]$ is positive semidefinite and thus can be diagonalized as

$$\Sigma = \Phi \Lambda \Phi^T$$

- Assign $x' = \Phi^T x$, then $\Sigma' = E(x' x'^T) = \Phi^T \Sigma \Phi = \Lambda$
- Further assign $x'' = \Lambda^{-1/2} \Phi^T x$, then $\Sigma'' = E(x'' x''^T) = I$

Classical multidimensional scaling (classical MDS)

- Given a distance matrix $D=\{d_{ij}\}$, the inner product matrix $G=\{x_i^T x_j\}$ can be computed by a bidirectional centering process

$$G = -\frac{1}{2}(I - \frac{1}{n} ee^T) D (I - \frac{1}{n} ee^T) \text{ where } e = [1,1,\ldots,1]^T$$

- G can be diagnolized as $G = \Psi \Lambda' \Psi^T$
- Actually, $n\Lambda$ and Λ' share the same set of eigenvalues, and

$$\Phi = X^T \Psi \text{ where } X = [x_1,\ldots,x_n]^T$$

$$\text{Because } G = XX^T, X \text{ can then be receovered as } X = \Psi \Lambda'^{1/2}$$

Cost Function Optimization

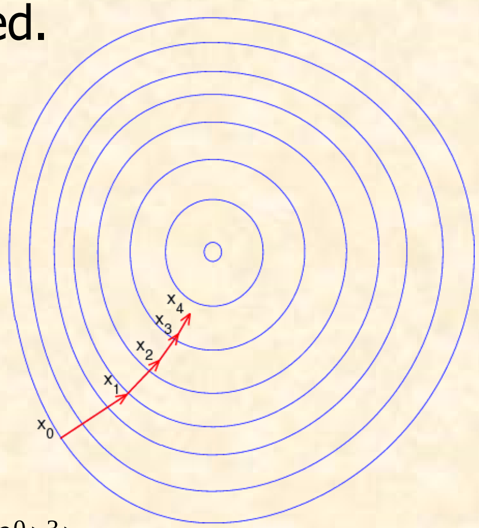
❖ Find θ so that a differentiable function J(θ) is minimized.

❖ **Gradient descent method**

- Starts with an initial estimate θ(0)
- Adjust θ iteratively by

$$\theta_{new} = \theta_{old} + \Delta\theta$$

$$\Delta\theta = -\mu \frac{\partial J(\theta)}{\partial \theta}\Big|_{\theta=\theta_{old}}, \text{ where } \mu > 0$$



- Taylor expansion of J(θ) at a stationary point θ^0

$$J(\theta) = J(\theta^0) + (\theta-\theta^0)^T \mathbf{g} + \frac{1}{2}(\theta-\theta^0)^T \mathbf{H}(\theta-\theta^0) + O((\theta-\theta^0)^3)$$

$$\text{where } \mathbf{g} = \frac{\partial J(\theta)}{\partial \theta}\Big|_{\theta=\theta^0} \text{ and } \mathbf{H}(i,j) = \frac{\partial^2 J(\theta)}{\partial\theta_i \partial\theta_j}\Big|_{\theta=\theta^0}$$

Ignore higher order terms within a neighborhood of θ^0

$$\theta_{new} - \theta^0 = (I - \mu\mathbf{H})(\theta_{old} - \theta^0)$$

$\mathbf{H}$ is positive semidefinite, then $\mathbf{H} = \Phi\Lambda\Phi^T$, we get

$$\Phi^T(\theta_{new} - \theta^0) = (I - \mu\Lambda)\Phi^T(\theta_{old} - \theta^0)$$

which will converge if every $|1-\mu\lambda_i| < 1$, i.e., $\mu < 2/\lambda_{\max}$.

Therefore, the convergence speed is decided by $\lambda_{\min}/\lambda_{\max}$.

❖ Newton's method

- Adjust θ iteratively by

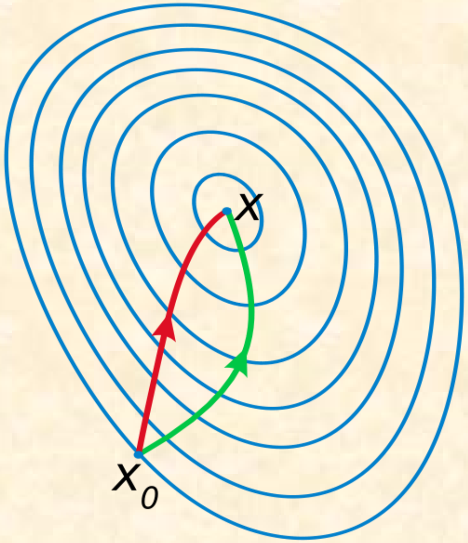
$$\Delta\theta = -\mathbf{H}_{old}^{-1}\frac{\partial J(\theta)}{\partial\theta}\Big|_{\theta=\theta_{old}}$$

- Converges much faster that gradient descent. In fact, from the Taylor expansion, we have

$$\frac{\partial J(\theta)}{\partial\theta} = \mathbf{H}(\theta-\theta^0)$$

$$\theta_{\mathrm{new}} = \theta_{old} - \mathbf{H}^{-1}(\mathbf{H}(\theta_{old}-\theta^0)) = \theta^0$$

- The minimum is found in one iteration.

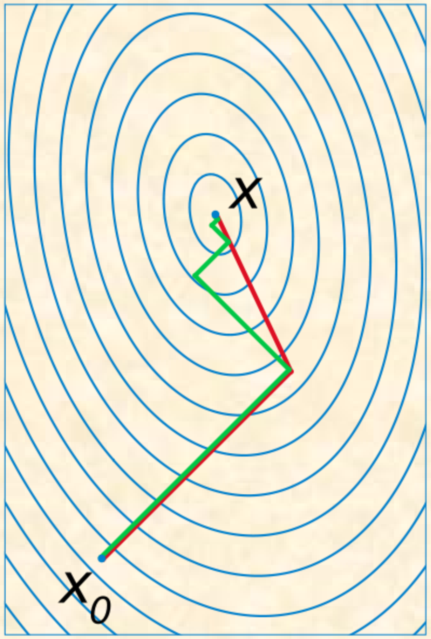


❖ Conjugate gradient method

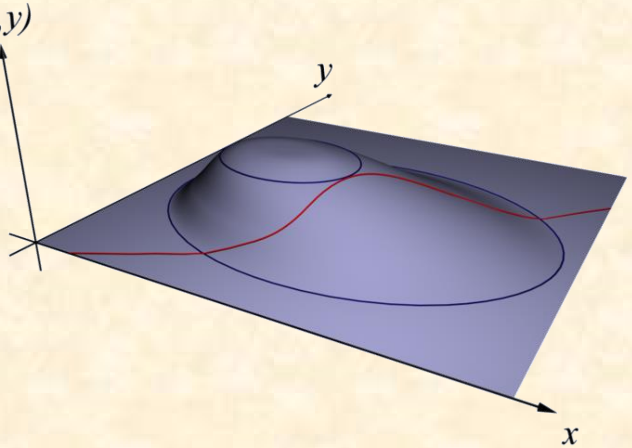
$$\Delta\theta_t = g_t - \beta_t\Delta\theta_{t-1}$$

$$\text{where } g_t = \frac{\partial J(\theta)}{\partial\theta}\Big|_{\theta=\theta_t}$$

$$\text{and } \beta_t = \frac{g_t^T g_t}{g_{t-1}^T g_{t-1}} \text{ or } \beta_t = \frac{g_t^T(g_t - g_{t-1})}{g_{t-1}^T g_{t-1}}$$



Constrained Optimization with Equality Constraints



Minimize J(θ)
subject to $f_i(\theta)=0$ for i=1, 2, …, m

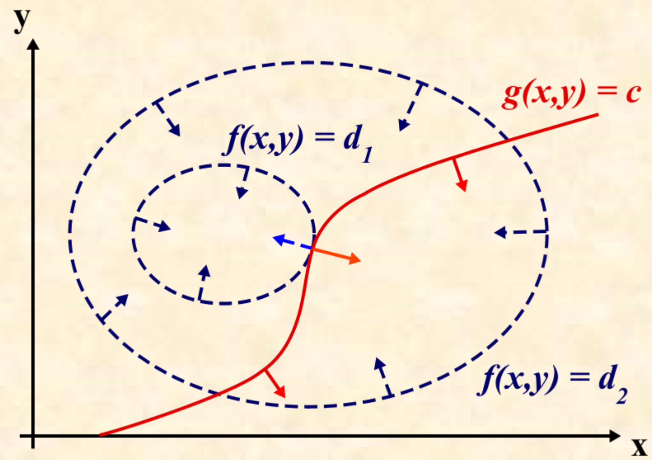
- Minimization happens at

$$\frac{\partial J(\theta)}{\partial \theta} = \lambda \frac{\partial f_i(\theta)}{\partial \theta}$$

- **Lagrange multipliers:** construct

$$L(\theta,\lambda) = J(\theta) - \sum_{i=1}^{m} \lambda_i f_i(\theta)$$

and solve $\dfrac{\partial L(\theta,\lambda)}{\partial \theta} = \dfrac{\partial L(\theta,\lambda)}{\partial \lambda} = 0$



Constrained Optimization with Inequality Constraints

Minimize $J(\theta)$ subject to $f_i(\theta) \geq 0$ for $i=1, 2, \ldots, m$

- $f_i(\theta) \geq 0$ $i=1, 2, \ldots, m$ defines a feasible region in which the answer lies.
- **Karush–Kuhn–Tucker (KKT) conditions:**

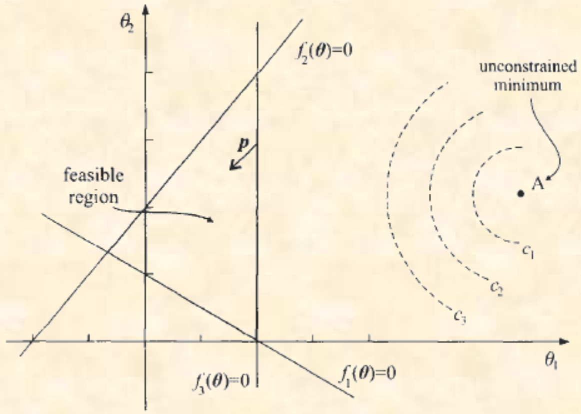
  A set of necessary conditions, which a local optimizer θ_* has to satisfy.

  There exists a vector λ of Lagrange multipliers such that

$$(1)\quad \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) = 0$$

$$(2)\quad \lambda_i \geq 0 \text{ for } i = 1,2,\ldots,m$$

$$(3)\quad \lambda_i f_i(\boldsymbol{\theta}_*) = 0 \text{ for } i = 1,2,\ldots,m$$



(1) Most natural condition.
(2) $f_i(\theta_*)$ is inactive if $\lambda_i = 0$.
(3) $\lambda_i \geq 0$ if the minimum is on $f_i(\theta_*)$.
(4) The (unconstrained) minimum in the interior region if all $\lambda_i = 0$.
(5) For convex $J(\theta)$ and the region, local minimum is global minimum.
(6) Still difficult to compute. Assume some $f_i(\theta_*)$'s active, check $\lambda_i \geq 0$.

- **Convex function:**

$$f(\theta): S \subseteq \Re^l \to \Re \text{ is convax if } \forall \theta, \theta' \in S, \lambda \in [0,1]$$

$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda) f(\theta')$$

- **Concave function:**

$$f(\lambda\theta + (1-\lambda)\theta') \geq \lambda f(\theta) + (1-\lambda) f(\theta')$$

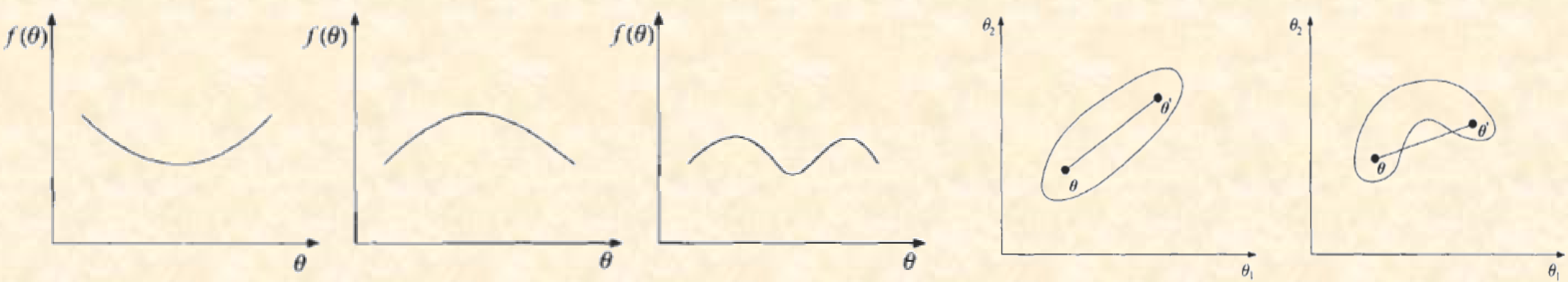
- **Convex set:**

$$S \subseteq \Re^l \text{ is a convax set if } \forall \theta, \theta' \in S, \lambda \in [0,1]$$

$$\lambda\theta + (1-\lambda)\theta') \in S$$

Local minimum of a convex function is also global minimum.

If $f(\theta)$ is concave, then $X = \{\theta \mid f(\theta) \geq b\}$ is a convex set.



❖ Min-Max duality

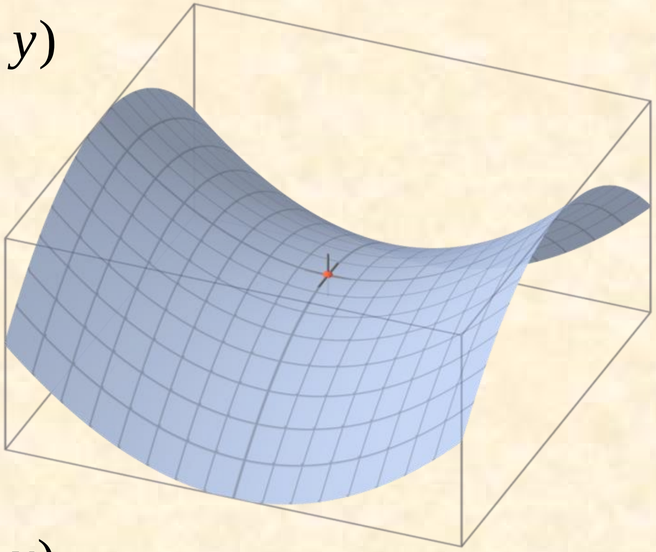
Game : A pays $F(x, y)$ \$ to B while A chooses x and B chooses y

A's goal : $\min_x \max_y F(x, y)$, B's goal : $\max_y \min_x F(x, y)$

The two problems are dual to each other.

In general : $\min_x F(x, y) \le F(x, y) \le \max_y F(x, y)$

Therefore, $\max_y \min_x F(x, y) \le \min_x \max_y F(x, y)$



Saddle point condition :

If there exists (x_*, y_*) such that

$F(x_*, y) \le F(x_*, y_*) \le F(x, y_*)$

or equivalent ly :

$F(x_*, y_*) = \max_y \min_x F(x, y) = \min_x \max_y F(x, y)$

❖ Lagrange duality

- Recall the optimization problem:

$$\text{Minimize} \quad J(\theta) \qquad \text{s.t. } f_i(\theta) \ge 0 \text{ for } i = 1,2,\dots,m$$

$$\text{Lagrange function } : L(\theta,\lambda) = J(\theta) - \sum_{i=1}^{m} \lambda_i f_i(\theta)$$

Because $\max_{\lambda \ge 0} L(\theta,\lambda) = J(\theta)$, we have

$$\min_{\theta} J(\theta) = \min_{\theta} \max_{\lambda \ge 0} L(\theta,\lambda)$$

- Convex Programming

For a large class of applicatio ns, $J(\theta)$ is convex, $f_i(\theta)$'s are concave then, the minimizati on solution (θ_*, λ_*) is a saddle point of $L(\theta,\lambda)$

$$L(\theta_*,\lambda) \le L(\theta_*,\lambda_*) \le L(\theta,\lambda_*)$$

$$L(\theta_*,\lambda_*) = \min_{\theta} \max_{\lambda \ge 0} L(\theta,\lambda) = \max_{\lambda \ge 0} \min_{\theta} L(\theta,\lambda)$$

Therefore, the optimizati on problem becomes $\max_{\lambda \ge 0} \min_{\theta} L(\theta,\lambda)$, or

$$\max_{\lambda \ge 0} L(\theta,\lambda) \qquad \text{subject to } \frac{\partial}{\partial \theta} L(\theta,\lambda) = 0$$

MUCH SIMPLER!

Mercer's Theorem and the Kernel Method

- Mercer's theorem:

  Let $x \in \Re^l$ and given a mapping $\phi(x) \in H$,

  (H denotes Hilbert space, i.e. finite or infinite Euclidean space)

  the inner product $\langle \phi(x), \phi(y) \rangle$ can be expressed as a **kernel function**

  $$\langle \phi(x), \phi(y) \rangle = K(x, y)$$

  where $K(x, y)$ is symmetric, continuous, and positive semi-definite.

  The opposite is also true.

The kernel method can transform any algorithm that solely depends on the dot product between two vectors to a kernelized vesion, by replacing dot product with the kernel function. The kernelized version is equivalent to the algorithm operating in the range space of φ. Because kernels are used, however, φ is never explicitly computed.