# 行列式

$C(x_i \cdots x_j)$ 逆序对个数 $\begin{vmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{vmatrix} = \sum (-1)^{\tau(j_1 \cdots j_n)} a_{1j_1} \cdots a_{nj_n}$

$$\begin{vmatrix} a_1+b_1 & \cdots & a_n+b_n \\ c_1 & \cdots & c_n \\ d_1 & & d_n \end{vmatrix} = \begin{vmatrix} a_1 & & a_n \\ c_1 & \cdots & c_n \\ d_1 & & d_n \end{vmatrix} + \begin{vmatrix} b_1 & & b_n \\ c_1 & \cdots & c_n \\ d_1 & & d_n \end{vmatrix}$$

放缩2次

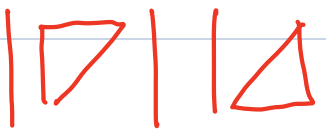
$\det(AB) = \det(A)\det(B)$

把某行k倍加到另一行，行列式不变 ⇐ 体积里一条边放缩

如果2列成比例，那么行列式为0

余子式 $\begin{vmatrix} \cdots & \vdots & \cdots \\ \cdots & i & \cdots \\ & j & \end{vmatrix}$ $M_{ij}$：去掉i行j列的det

$C_{ij}$ 代数余子式为 $(-1)^{i+j} M_{ij}$ $|A|$ 为一行/一列 $a_{ij} C_{ij}$

 对角线的积  $(-1)^{\frac{n(n-1)}{2}}$ 对角线（交换）

$$\begin{vmatrix} A & 0 \\ * & B \end{vmatrix} = \begin{vmatrix} A & * \\ 0 & B \end{vmatrix} = \det(A)\det(B) \qquad \begin{vmatrix} 0 & A \\ B & * \end{vmatrix} = \begin{vmatrix} * & A \\ B & 0 \end{vmatrix} = (-1)^{mn}|A||B|$$

# 矩阵

$A_{m\times n}$ $B_{n\times s}$ 若 $AB=0$ $r(A)+r(B) \le n$

Sylvester：$\text{rank}(AB) \ge \text{rank}(A) + \text{rank}(B) - n$

$\text{Null}(AB) \le (n-r(A)) + (n-r(B))$

把2次型写成矩阵乘法：$x_1^2 + 5x_2^2 - 4x_1x_2 = (x_1\ x_2)\begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

伴随矩阵 $A^*A=|A|E$ $A^*=\begin{vmatrix} A_{11} & & \\ & \ddots & \\ & & A_{nn} \end{vmatrix}$ 代数余子式

$(A^{-1})^*=\frac{1}{|A|}A$ $(kA)^*=k^{n-1}A^*$ $(AB)^{-1}=B^{-1}A^{-1}$

初等矩阵 $\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$ → 行列×k / 行列倍加 / 行列对调 } P

PA 即为对A相同行变换

AP则为列变换

经过初等变换 rank 不变

$rank(A)=rank(B) \Leftrightarrow$ A、B可经过多次初等变换互换

所有可逆矩阵都可以看成 $E_nE_{n-1}\cdots E_1I$，其中E为初等矩阵

矩阵的秩

矩阵乘法其实在压扁，压扁2次(AB)

$rank(AB)\le \min(rank(A), rank(B))$

可以认为乘法在互相组合行，列，组合出新的不会 rank 更大

如果可以用 $\beta_1-\beta_n$ 表示 $\alpha_1-\alpha_m$ $rank(\beta)>rank(\alpha)$

如果A的k阶行列式不为0，那么 $rank\ge k$ $\begin{bmatrix} \square & \end{bmatrix}\Rightarrow$ 扣去 $\begin{bmatrix} & \\ & \square \end{bmatrix}$

$rank(A+B)\le rank(A)+rank(B)$ $\Leftarrow$ $\dim(U+V)=\dim(U)+\dim(V)-\dim(U\cap V)$

$rank\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}=rank(A)+rank(B)$

特征值与相似对角化

特征向量 — 特征值

实对称矩阵一定可以正交对角化 $Q^T\Lambda Q$，特征向量相互正交

$A+kI\Rightarrow\lambda+k$ $kA\Rightarrow k\lambda$ $A^n\Rightarrow\lambda^n$ $A^{-1}\Rightarrow\frac{1}{\lambda}$

$A\sim B$, $\det(\lambda I-A)=\det(\lambda I-B)$ $rank(A)=rank(B)$

$trace(A)=\sum\lambda_i=trace(B)$, $A\sim\Lambda\sim B$, $A^n\sim B^n$, $A^{-1}\sim B^{-1}$

相似本质上是更换坐标系后的观测

如果一个矩阵 $A = uv^T$（2个向量张成的）

1. rank=1（都是u的倍数） $rank(AB) \le \min(rank(A), rank(B))$

2. $\lambda = u^T v$，其余都是0。 3. 代表把空间映射到一个直线上

如果矩阵A对称，那么 $\lambda_{max} = \frac{x^T A x}{\|x\|_2} = \|Ax\|_2$ (max)

任意情况下 $\sigma_{max} = \max \|Ax\|_2$ $\max \|Ax\|_2 = \lambda_{max}\left(\frac{A+A^T}{2}\right)$ } $\|x\|_2 = 1$

对称部分

正定矩阵

正定矩阵一定对称，顺序主子式一定大于0

A正定 $\Rightarrow$ $A^{-1}$ 正定 { 对称：$(A^{-1})^T = (A^T)^{-1} = A^{-1}$ ; $\lambda > 0$, $\frac{1}{\lambda} > 0$

PCA

对于数据 X $\Rightarrow$ $Cov = \frac{1}{n-1}(X-\mu)^T(X-\mu)$

$$X-\mu = \begin{bmatrix} x_{11}-\mu_1 & \cdots & x_{1n}-\mu_n \\ \vdots & & \vdots \\ x_{n1}-\mu_1 & \cdots & x_{nn}-\mu_n \end{bmatrix}$$

找到几个让 Varience 尽量大的方向

$\Leftrightarrow$ 特征值最大的几个特征向量

Cov矩阵对称，不同特征值对应的特征向量正交

相同特征值的也可以彼此正交化

$$V = \left[\frac{v_1}{\|v_1\|_2}, \frac{v_2}{\|v_2\|_2} \cdots \frac{v_n}{\|v_n\|_2}\right]$$

$$X_{PCA} = (X-\mu)\cdot V$$

SVD $AV = \Sigma U$，其中 $V$ 为 $A^TA$ 的特征向量

$A \Rightarrow$ 旋转($U$) $\Rightarrow$ 放缩($\Sigma$) $\Rightarrow V^T$(旋转)

$A = U\Sigma V^T \quad A^TA = V\Sigma^2V^T \quad AA^T = U\Sigma^2U^T$

$U, V$ 都为正交矩阵(旋转) $\Sigma$ 是 $A^TA$ | $AA^T$ 特征值的 sqrt

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_m & \end{array}\right] \quad A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$$

$$\sigma_{max} = \sup_x \frac{\|Ax\|_2}{\|x\|_2} \quad \|A\|_F = \sqrt{\sum \sigma_i^2}$$

$$\|x\|_1 = \sum |x_i|$$

$$\|x\|_2 = \sqrt{\sum x_i^2}$$

## Example 1

$$\|x\|_\infty = \max |x_i|$$

Example: The following functions are matrix norms:

- $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$ (maximum of the $\ell_1$ norms of columns). Moreover, this is an induced matrix norm:

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1.$$

列和

- $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$. Moreover, this is an induced matrix norm:

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

最大拉伸长 $\sigma_{max}$

这个向量 $x$ 就是 $A^TA$ 对应 $\lambda_{max}$ 的特征向量

- $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$ (maximum of the $\ell_1$ norms of rows). Moreover, this is an induced matrix norm:

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty.$$

行和

- $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2}$. This is known as the Fröbenius norm.

$$= \sqrt{\sum \sigma_i^2}$$

$$\min_X \|A - X\|_F \text{ s.t. } \text{rank}(X) \le k \quad X = A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$$

Linear Regression $y = \theta_x^T x + \theta_0 = \theta^T \bar{x}$ $\theta^T = [\theta_x \ \theta_0]$ $\bar{x} = [x, 1]^T$

$$\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum (y_i - \theta^T \bar{x}_i)^2 = \arg\min_\theta \frac{1}{n} \|Y - X\theta\|^2$$

$$= \arg\min (y - X\theta)^T (y - X\theta)$$

导数为 $-2X^T y + 2X^T X\theta = 0$ $X^T X\theta = X^T y$

invertible $\theta = X^{-1} y$

U, V是保留奇异值那么多个

Not invertible ⇒ pseudo-inverse ⇒ $X = U\Sigma V^T$ ⇒ $X^\dagger = U\Sigma^{-1} V^T$

## Derivative of A Matrix

Scalar by vector: $\frac{\partial 1\times 1}{\partial n\times 1} = n\times 1$

$\frac{\partial a^T x}{\partial x} = a$ $\frac{\partial x^T a}{\partial x} = a$ $\frac{\partial x^T A x}{\partial x} = (A + A^T) x$ $\frac{\partial \|x\|_2^2}{\partial x} = 2x$

Scalar by Marix $\frac{\partial 1\times 1}{\partial n\times m} = n\times m$

$\frac{\partial \text{trace}(AB)}{\partial A} = B^T$ $\frac{\partial \text{tr}(AB)}{\partial B} = A^T$

$\frac{\partial x^T A x}{\partial A} = x x^T$ $\frac{\partial \text{tr}(A^T A B)}{\partial A} = A(B + B^T)$ $\frac{\partial \det(A)}{\partial A} = |A| \cdot (A^{-1})^T$

vector by vector

$\frac{\partial A x}{\partial x} = A$ $\frac{\partial x^T A}{\partial x} = A^T$

# Probability

如果 $x>0$ $\int_0^{\infty} P(X>t)\,dt$

$$EX=\int_{-\infty}^{+\infty} x\,dF(x) \quad \text{if } \int_{-\infty}^{+\infty}|x|\,dF(x) \text{ 存在}$$

$$=\int_{-\infty}^{+\infty} xf(x)\,dx \quad \text{如果 } f(x) \text{ 存在}$$

$$Var(X)=E(X-EX)^2=EX^2-(EX)^2 \quad Cov(X,Y)=E(X-E(X))(Y-E(Y))$$

$$=E(XY)-EXEY$$

Moments: $E(X^k)=\int_{-\infty}^{+\infty} x^k\,dF(x)$

Normal: $f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

$$(\frac{1}{\sqrt{2\pi}})^n\cdot\frac{1}{|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

在独立情况下是 $\sigma_1\sigma_2\cdots\sigma_n$ $\begin{vmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2\end{vmatrix}$

$y\sim N(\mu,\Sigma)$ $a^Ty\sim N(a^T\mu,a^T\Sigma a)$ $Ay+d\sim N(A\mu+d,A\Sigma A^T)$

$$y=\begin{pmatrix}y_1\\y_2\end{pmatrix}\quad \mu=\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\quad \Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$$

$y_1\sim N(\mu_1,\Sigma_{11})$

$y_2\sim N(\mu_2,\Sigma_{22})$

$X\sim N(\mu_x,\Sigma_x)$ $y\sim N(\mu_y,\Sigma_y)$

$x+y\sim N(\mu_x+\mu_y,\Sigma_x+\Sigma_y)$ $\begin{pmatrix}x\\y\end{pmatrix}\sim N\left(\begin{pmatrix}\mu_x\\\mu_y\end{pmatrix},\begin{pmatrix}\Sigma_x&\\&\Sigma_y\end{pmatrix}\right)$ 左乘 $\begin{bmatrix}1\\1\end{bmatrix}$

归一化: $(\Sigma^{\frac{1}{2}})^{-1}(y-\mu)$ $\Sigma=\sum\lambda_i e_i e_i^T$ $\Sigma^{\frac{1}{2}}=\sum\sqrt{\lambda_i}e_ie_i^T$

MGF: $E(e^{tX})$ $\frac{d^n}{dt}\Big|_{t=0}=E(X^n)$ PGF:

泰勒展开: $f(x+d)=f(x)+f'(x)d+\frac{1}{2}f''(x)d^2+\cdots+\frac{1}{n}f^{(n)}(x+\theta d)d^n$

$$f(x)+\nabla f(x)^Td+\frac{1}{2}d^T\nabla^2 f(x)d\cdots\cdots$$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2 \quad L\text{-Smooth}$$

$$+ \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) + O(\|y-x\|^2)$$

$$\text{or } + \frac{1}{2}(y-x)^T \nabla f(\xi)(y-x) \quad x < \xi < y$$

# Optimization

目标：做 constrained optimization

Local minimum $x^*$ 下 $\nabla f(x^*)^T d \geq 0$ feasible d: $x + \alpha d \in \Omega$

Let $f \in C^1$, if $x^*$ is a local minimum and interior point $\nabla f(x^*) = 0$

**Convex 问题** → 只有1个 local min 好解

Convex Set: $\alpha \in [0,1]$

$\forall x, y$, $\alpha x + (1-\alpha) y \in \Omega$ （连线都在）

Convex function:

$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$ ← 决定是否 strict

① $f+g$ convex ② $cf$, $c>0$

③ $\{x : x \in \Omega, f(x) \leq c\}$ convex Set. 小于的都"连续区间"

等价于 $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

$\nabla^2 f(x) \geq 0$

**Constrained** ← → **unconstrained**

**在KKT这个人/墙条件下才会有（KKT必要）**

$\min f(x)$ 势能 Active Set.

s.t. $C_E(x) = 0$ $i \in E$ 墙上 $A(x) = \{i \mid i \in I \mid c_i(x) = 0\}$

$C_I(x) \geq 0$, $i \in I$ 墙内 这些在边界上，要警用了

LICQ：$\{\nabla c_i(x), i \in A(x)\}$，线性无关 [边界上各个方向力无关]

如果最优解 $x^*$ 处 LICQ，于 $\lambda^*$ ⇐ 力 KKT 理想情况

① $\nabla f(x^*) - \sum \lambda_i^* \nabla c_i(x) = 0$ $\lambda_i^* \geq 0$ $\lambda_i^* c_i(x^*) = 0$

人对墙力 墙反作用力 不靠墙就没力

对偶问题

我们把墙换成风，$\sum \lambda_i c_i(x)$ 动能，$f(x)$ 是势能

$L(x, \lambda) = f(x) - \lambda^T c(x)$ 能量为0平衡了态

$q(\lambda) = \min_x L(x, \lambda) \leq p^*$（有墙） $c(x)$ 为距离 墙的深度

$\max_{\lambda} q(\lambda)$，调整风力，模拟有墙的情况

这个 $\lambda$ 是不同约束的力量，$\lambda$ 越大越难动摇（$\lambda$ 为打破这个约束可以收益多少）

Slater 条件：如果有一个可行点 $\hat{x}$ 满足所有 $C_i(x) > 0$ $i \in I$，$C_i(x) = 0$ $i \in E$

⇒ 强对偶 + 凸问题 KKT 充要，非凸必要 墙内有空间，没有定死

普通问题在 $x^*$ 如果有 CQ（LICQ ⇒ MFCQ），KKT 必要

凸问题下，KKT 的解就是全局最优 ⇒

↳ 再记个 Slater 可得 KKT 必要 ⇐

MFCQ：对于 $i \in E$ $\nabla C_i(x)^T d = 0$ $i \in I$ $\nabla C_i(x)^T d > 0$

有一个方向 d 可以走 5方止 被挤在这里

LICQ：  线性无关

凸共轭 $f^*(y) = \sup_x (\langle y, x \rangle - f(x))$

**在LP问题中强对偶，KKT充要，可用来验证答案**

Fenchel 对偶 $f^*(k) = \sup_x (x^T k - f(x))$ 给定斜率 k，kx 到 $f(x)$ max d

如果 f 是 convex + closed，$f^{**}(x) = f(x)$ $f(x) + f^*(y) \geq \langle x, y \rangle$

对于 $\min_x f(x) + g(Ax)$ 对偶问题是 $\sup_y -f^*(-A^T y) - g^*(y)$

一般都是对 $\|x\|_1$ 这种不光滑不可求导的

# Linear Programming

$$\min c^T x \quad Ax = b \quad x \geq 0 \Rightarrow$$

- 求max的话是负的min
- $\sum a_{ij} x_j \leq b_i \Rightarrow \sum a_{ij} x_{ij} + x_{n+1} = b_i \quad x_{n+1} \geq 0$
- (没有说 $x_i \geq 0$)
- 有 $x_i$ 无限制的话，$x_i' \geq 0 \quad x_i'' \geq 0 \quad x_i = x_i' - x_i''$

# Gradient descent

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla L(\theta^{(t)})$$

条件数 $K = \frac{L}{\mu}$，$L$ 是最大曲，$\mu$ 是最小曲

往垂直等高线方向走会对 $-2\eta g$

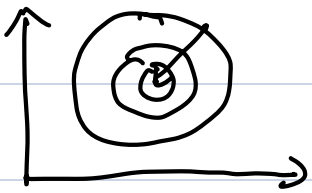
最优 $\frac{2}{L+\mu}$ 下

$$\theta^t - \theta^* \le \left(\frac{1-K}{1+K}\right)^t (\theta^0 - \theta^*)$$

收缩：$L(\theta^{(t)}) - L(\theta^*) \le \left(1-\frac{1}{K}\right)^t \left(L(\theta^{(0)}) - L(\theta^*)\right)$

保守 $\frac{1}{L}$ 下

K越大，收缩越慢



optimal step size $\frac{2}{L+\mu}$

$O(K \log \frac{1}{\varepsilon})$ iterations

在 $x$ 很小的时候 $1-x \approx e^{-x}$

$\left(1-\frac{1}{K}\right)^t \approx e^{-\frac{t}{K}}$ 给定误差 $\varepsilon$，$e^{-\frac{t}{K}} = \varepsilon$ $t = -K \log \varepsilon$

误差随时间指数↓ $= K \log \frac{1}{\varepsilon}$

# SGD:

SGD在使用Batch下会有

$$E(L\theta^{t+1}) \le L\theta^t - (\eta - \frac{L\eta^2}{2}) \|\nabla L\theta^t\|_2^2 + \frac{L\eta^2\sigma^2}{2}$$

$\eta \to 0$ Common choice for SGD Variance Term

are $O(\frac{1}{t})$ or $O(\frac{1}{\sqrt{t}})$

误差随时间

$\frac{K-1}{K+1}$ $\frac{\mu-L}{\mu+L}$

对于 convergence rate，去算 $\theta_{t+1} - \theta_t$。

$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$ $\nabla f(\theta_t) \approx \nabla f(\theta^*) + (\theta_t - \theta^*) \nabla^2 f(\theta^*)$

$= \theta_t - \eta(\theta_t - \theta^*) \nabla^2 f(\theta^*)$ $e_{t+1} = (1 - \eta \nabla^2 f(\theta^*)) e_t$

$e_{t+1} = (1-\eta\mu) e_t$ $\eta = \frac{1}{L}, (1-\frac{1}{K})^t$, $\eta = \frac{2}{L+\mu}, (\frac{K-1}{K+1})^t$

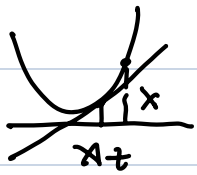
对于 max $\eta$，算 $f(\theta_{t+1}) = f(\theta_t) + \nabla f(\theta_t)^T(\theta_{t+1}-\theta_t) + \frac{L}{2}\|d\|^2$

$= f(\theta_t) - \nabla f^T(\theta_t)\cdot\eta\nabla f(\theta_t) + \frac{L}{2}\eta^2\|g_t\|^2$

$\eta - \frac{L}{2}\eta^2 > 0$，$\eta < \frac{2}{L}$

## 牛顿法

求零点：$g(x) \approx g(x_t) + \nabla g(x_t)(x - x_t)$




$g(x) = 0 \qquad x = x_t - \frac{g(x_t)}{\nabla g(x_t)}$

如果要求极值，相当于是 $\nabla g(x) = 0$

$x = x_t - \frac{\nabla g(x_t)}{\nabla^2 g(x_t)}$ ⇐ 不适用于多元情况（无除法）

$L(\theta) = L(\theta_t) + \nabla L(\theta_t)(\theta-\theta_t) + (\theta-\theta_t)^T\nabla^2 L(\theta_t)(\theta-\theta_t)$

$\nabla L(\theta) = \nabla L(\theta_t) + \nabla^2 L(\theta_t)(\theta-\theta_t) = 0$

$\theta = \theta_t - (\nabla^2 L(\theta_t))^{-1}\nabla L(\theta_t)$

$= \theta_t - [H(\theta_t)]^{-1}\nabla L(\theta_t)$ ← 需要严格凸才有逆

这是一种收敛极快的方法，$\varepsilon_{t+1} \le C\cdot\varepsilon_t^2$

但是要在离的很近的时候才收敛快，可以 GD + Newton

算 $\nabla^2 L(\theta)$ 比较费时间，也经常不用

$\nabla f(x_k) = \underbrace{\nabla f(x^*)}_{0} + H(x^*)(x_k - x^*) + O(\|x_k - x^*\|^2)$

$x_{k+1} - x^* = x_k - H^{-1}(x_k)\underbrace{(H(x^*)(x_k - x^*) + O)}_{\nabla f(x_k)} - x^*$

$e_{k+1} = \underbrace{[I - H^{-1}(x_k)H(x^*)]e_k}_{0} - \underbrace{H^{-1}(x_k)O}_{C\cdot e_k^2}$

# Logistic Regression

$\sigma(x) = \frac{1}{1+e^{-x}}$ $\sigma'(x) = \sigma(x)(1-\sigma(x))$

$$P(y=1|x) = \hat{y} = \sigma(Xw) = \frac{1}{1+e^{-xw}}$$ $xw$

$$L(w) = -\sum[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$$

$$= -y^T \log \hat{y} + (1-y)\log(1-\hat{y}_i)$$

⇑ 定义 $P(y|w,x) = (\hat{y}_i)^{y_i}(1-\hat{y}_i)^{(1-y_i)}$ $\begin{cases} y_i=0 & 1-\hat{y}_i \\ y_i=1 & \hat{y}_i \end{cases}$

$$\frac{\partial L}{\partial w} = X^T(\hat{y}-y) = X^T(\sigma(Xw)-y)$$

$$\nabla^2 L = X^T S X \quad S = \mathrm{diag}(\hat{y}_i(1-\hat{y}_i))$$

$u^T X^T S X u = v^T S v$ $S_{ii} = \hat{y}_i(1-\hat{y}_i) > 0$ $v^T S v > 0$

$\nabla^2 L > 0$ convex

另一种情况 $y = \pm 1$

那么可以写成 $\sigma(YXW)$, $Y = \mathrm{diag}(y_1 \sim y_n)$

$L(w) = 1^T \log(1+\exp(-YXw))$

$$\nabla L(w) = \frac{1^T e(-YXw)}{1+e(-YXw)}$$

# Generalization Error

$\hat{R}(\theta) = \frac{1}{n}\sum l_i(x_i, y_i, \theta)$ 是你能算出的 Empirical

$R(\theta) := E[l(x, y, \theta)]$ population risk, generalization loss

$R(\hat{\theta}) = \underbrace{\hat{R}(\hat{\theta})}_{\text{能有的}} + \underbrace{(R(\hat{\theta}) - \hat{R}(\hat{\theta}))}_{\text{Gap}}$

# Bias - Variance Trade-off

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

想要 Low variance & Low Bias

# Uniform Convergence

$n \to \infty$ $\hat{R}(\theta) \to R(\theta)$ $\sup_{\theta \in H} |\hat{R}(\theta) - R(\theta)| < \varepsilon$