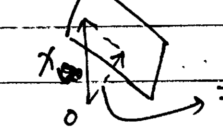
189 MT

Oct. 13

- Linear classifier:

$f(x) = \theta^T x + \theta_0$

$\hat{y} = \begin{cases} 1 & f(x) \geq 0 \\ -1 & f(x) < 0 \end{cases}$



$\langle x, \frac{\theta}{\|\theta\|} \rangle = \frac{-\theta_0}{\|\theta\|}$

normal vect. offset

Decision Boundary $H := \{x \in \mathbb{R}^d : \theta^T x + \theta_0 = 0\}$

$\tilde{x} := \begin{bmatrix} 1 \\ x \end{bmatrix}$ $\tilde{\theta} := \begin{bmatrix} \theta_0 \\ \theta \end{bmatrix}$ $\Rightarrow f(x) = \tilde{\theta}^T \tilde{x}$

- Perceptron Algorithm:

Initialize $\theta = 0$.

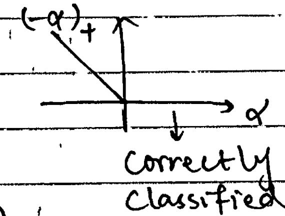
while $\exists\, y_i \neq \operatorname{sgn}(\theta \cdot x_i)$:

$\theta := \theta + y_i x_i$

return θ.

Note that $\theta^* = \sum \alpha_i y_i x_i$.

$\theta^* \cdot x = \sum \alpha_i y_i x_i \cdot x$



$J(\theta) := \sum_i (-y_i (\theta \cdot x_i))_+$

$J_i(\theta) := (-y_i (\theta \cdot x_i))_+$

$\Rightarrow J(\theta) = \sum J_i(\theta)$, $\nabla J(\theta) = \nabla J_i(\theta)$.

$\nabla J_i(\theta) = \begin{cases} -y_i x_i & \text{if } y_i (\theta \cdot x_i) < 0 \\ 0 & \text{o/w} \end{cases}$

SGD: $\theta \leftarrow \theta - \nabla J_i(\theta) = \theta + y_i x_i$

for misclassified pair (x_i, y_i)

$\simeq$ Perceptron Algorithm.

converges only if data are saparable.

- Support Vector Machine

For Linearly separable data:

$\exists \theta$ s.t. $\forall i$, $y_i(\theta \cdot x_i) > 0$ (strictly separable)

let $\delta = \min(y_i(\theta \cdot x_i))$

$\Rightarrow y_i(\frac{1}{\delta}\theta \cdot x_i) \geq 1$

- Projecting x to $\{a^T x + b = 0\}$:

$\hat{x} = x - \frac{|a^T x + b|}{\|a\|^2} a$

distance $|a^T x + b| / \|a\|$

margin $:= \min \frac{y_i(\theta \cdot x_i + \theta_0)}{\|\theta\|}$

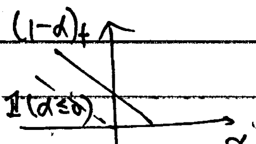
Maximize margin (QP)

$\max_{\theta, \theta_0} \min_i \frac{y_i(\theta \cdot x_i + \theta_0)}{\|\theta\|}$

$= \min_{\theta, \theta_0} \min_i y_i(\theta \cdot x_i + \theta_0)$

$= \min_{\theta, \theta_0} \|\theta\|^2$ s.t. $y_i(\theta \cdot x_i + \theta_0) \geq 1$

$\rightarrow$ Hard margin SVM



Relax by introducing slack var: $\xi_i \geq 0$

$y_i(\theta \cdot x_i + \theta_0) \geq 1 - \xi_i$

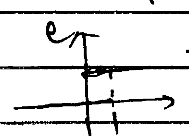
New cost function: $\min_{\theta, \theta_0} \|\theta\|^2 + C \sum_i \xi_i$

$\mathbb{I}(\alpha \leq 0) < (1-\alpha)_+$ s.t. $y_i(\theta \cdot x_i + \theta_0) \geq 1 - \xi_i$

convex relaxation $\xi_i \geq 0$

(upper bound) $\Leftrightarrow \min C \sum_i \max(1 - y_i(\theta \cdot x_i + \theta_0), 0) + \|\theta\|^2$

for misclassifications. or $\xi_i = (1 - y_i(\theta \cdot x_i + \theta_0))_+$



$\rightarrow$ Soft margin SVM.

With GD: $w^{(t+1)} \leftarrow (1-\alpha_t) w^{(t)} + \alpha C \sum e(y_i w^{(t)T} x_i) y_i x_i$

Equivalently $w^{(t+1)} = (1-\gamma) w^{(t)} + \eta \begin{cases} y_{ik} x_{ik}, & y_{ik} w^{(t)T} x_{ik} \leq 1 \\ 0 & o/w \end{cases}$

- Convexity. Equiv. def:

i) $f(tx + (1-t)x') \leq t f(x) + (1-t) f(x')$; $\forall t \in [0,1]$

ii) $f(x') \geq f(x) + \nabla f(x)^T (x' - x)$

iii) $(\nabla f(x') - \nabla f(x))^T (x' - x) \geq 0$

iv) $\nabla^2 f(x) \succeq 0$, $\forall x$

- Strong Convexity. m-strongly convex:

i) $f(tx + (1-t)x') \leq t f(x) - (1-t) f(x') - \frac{t(1-t)m}{2} \|x - x'\|^2$, $t \in [0,1]$

ii) $g(x) = f(x) - \frac{m}{2}\|x\|^2$ is convex

iii) $f(x') \geq f(x) + \nabla f(x)^T (x' - x) + \frac{m}{2}\|x' - x\|^2$

iv) $(\nabla f(x') - f(x)^T)(x' - x) \geq m \|x' - x\|^2$

v) $\nabla^2 f(x) \succeq mI$ (i.e. $\nabla^2 f(x) \succeq 0$, $m \leq \lambda_{min}(\nabla^2 f(x))$)

Strongly convex functions can be lower bounded by a quadratic function

Guarantees ≤ 1 global minimum

- Smoothness (upper bounded by a quadratic func.)

M-smooth: $\|\nabla f(x') - \nabla f(x)\| \leq M \|x' - x\|$

Equiv. def:

i) iii) iv) v) flipping ineq signs

- Gradient Descent

$D_u f(w^{(t)}) := \langle \nabla f(w^{(t)}), u \rangle \geq -\nabla f(w^{(t)})$

$u^* = -\nabla f(w^{(t)})$

$w^{(t+1)} \leftarrow w^{(t)} - \alpha_t \nabla f(w^{(t)})$ until $f(w^{(t)})$ cvg

- Taylor's Theorem:
  - $f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}(y-x)^T \nabla^2 f(\tilde{x})(y-x)$
    with $\tilde{x} = tx + (1-t)y$ for some $t \in [0,1]$
  - $F: \mathbb{R}^d \to \mathbb{R}^k$. $F(w+\delta) = F(w) + J_F(w)\delta + \mathcal{O}(\|\delta\|)$
  - $f: \mathbb{R}^d \to \mathbb{R}$, $f(x) := \frac{1}{2}x^T A x - x^T b$
    $\nabla f(x) = Ax - b$, $\nabla^2 f(x) = A$
    $f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{1}{2}(x-x^*)^T A (x-x^*)$
    $= f(x^*) + \frac{1}{2}(x-x^*)^T A (x-x^*)$
- Incremental Gradient Method
  $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
  $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$, $i_k \in [n]$
  If $g(x) := \nabla f_i(x)$, i randomly selected
  then $\mathbb{E}[g(x)] = \nabla f(x)$
  $\rightarrow$ Special case of SGD
- Ex: $f(x) = \frac{1}{2n}\sum_{i=1}^{n}(x-w_i)^2$
  $\nabla f_i(x) = x - w_i$, $\alpha_k := \frac{1}{k}$
  $\Rightarrow x_{k+1} = x_k - \frac{1}{k}(x_k - w_k) = \frac{1}{k}\sum w_i$
  Given a sequence $\{w_i\}$: SGD on $f(x) = \frac{1}{2}E[(x-w)^2]$
  where $w \sim (\mu, \sigma^2)$
  If run n steps with i.i.d. samples of w at each iteration, then $x_k = \frac{1}{n}\sum_{i=1}^{n} w_i$
  $f(x_k) = \frac{1}{2}E[(\frac{1}{n}\sum w_i - w)^2]$
  $= \frac{1}{2}E_w[E_{w_i}[(\frac{1}{n}\sum w_i - w)^2]] = \frac{1}{2n}\sigma^2 + \frac{1}{2}\sigma^2$
  If computing minimizer exactly:
  $f(x) = \frac{1}{2}E[x^2 - 2wx + w^2] = \frac{1}{2}x^2 - \mu x + \frac{1}{2}(\sigma^2 + \mu^2)$
  $x^* = \mu$, $f(x^*) = \frac{1}{2}\sigma^2$
  $\therefore$ After n iterations, $f(x) - f(x^*) = \frac{1}{2n}\sigma^2$.

- CCA
- Paired data matrices ($X_{n\times p}$, $Y_{n\times q}$), zero-meaned
  - Simultaneously find projection directions $u_{p\times 1}$, $v_{q\times 1}$ s.t. projected data have maximal correlation
    $\max_{u,v} \ \varepsilon(u,v; X,Y) = \rho(Xu, Yv)$
    $= \dfrac{Cov(Xu, Yv)}{\sqrt{V(Xu)\cdot V(Yv)}} = \dfrac{u^T X^T Y v}{\sqrt{u^T X^T X u \cdot v^T Y^T Y v}}$
  - Seeks k-dim basis, $k \le \min(p,q)$
    s.t. total projection correlations are maximized
  - $\max \varepsilon(u,v;X,Y)$ invariant to scaling & affine transform $X \leftarrow AXA$, $Y \leftarrow YB$
- Equivalently:
  $\max_{u,v} \ u^T C_{XY} v$ s.t. $u^T C_{XX} u = 1$, $v^T C_{YY} v = 1$

  Solved by Lagrangian:
  $L(u,v;a,b) = u^T C_{XY} v - a(u^T C_{XX} u - 1) + b(v^T C_{YY} v - 1)$
  $L_u = L_v = L_a = L_b = 0 \Rightarrow a,\ u^T C_{XY} v = \lambda$ (find max)
  $\left\{ \begin{array}{l} u^T C_{XX} u = 1 \\ v^T C_{YY} v = 1 \end{array} \right.$
  $\Rightarrow \underbrace{\begin{bmatrix} & C_{XY} \\ C_{YX} & \end{bmatrix}}_{M} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \underbrace{\begin{bmatrix} C_{XX} & \\ & C_{YY} \end{bmatrix}}_{D} \begin{bmatrix} u \\ v \end{bmatrix}$

  Rayleigh Quotient Optimization:
  $\max_{w} \ \varepsilon(w; M, D) = \dfrac{w^T M w}{w^T D w} = eig(M, D)$
  $Mw^* = \lambda_1 D w^*$
  - PCA/TLS: $M = C_{XX}$, $D = I$
  - CCA: $M = \begin{bmatrix} & C_{XY} \\ C_{YX} & \end{bmatrix}$ $D = \begin{bmatrix} C_{XY} & \\ & C_{YY} \end{bmatrix}$

• SVD: $X = USV^T \Rightarrow Xv_i = s_i u_i,\ X^T u_i = s_i v_i$

$$\begin{bmatrix} & X \\ X^T & \end{bmatrix}\begin{bmatrix} u_i \\ v_i\end{bmatrix} = s_i \begin{bmatrix} u_i\\ v_i\end{bmatrix}$$

$$\Rightarrow \left(\begin{bmatrix}u_i\\v_i\end{bmatrix}, s_i\right) = eig\left(\begin{bmatrix} & X\\ X^T & \end{bmatrix}\right)$$

$\max_{u,v} align(u,v;M) := u^T M v$
s.t. $\|u\|=1,\ \|v\|=1$

$M = USV^T,\ u = U\alpha,\ v = V\beta,\ \|\alpha\| = \|\beta\| = 1$

$\Rightarrow align(u,v;M) = \alpha^T S \beta \le s_1$

$\alpha^* = \beta^* = [1, 0, \dots, 0]^T$

$u^* = U\alpha^* = u_1,\ v^* = V\beta^* = v_1$

$$\left\{\begin{array}{l} \begin{bmatrix}\tilde u\\ \tilde v\end{bmatrix} = \begin{bmatrix}C_{xx} & \\ & C_{yy}\end{bmatrix}^{\frac{1}{2}}\begin{bmatrix}u\\ v\end{bmatrix} \Rightarrow CCA = \max \tilde u^T \tilde C_{xy} \tilde v \quad s.t.\ \|\tilde u\| = 1,\ \|\tilde v\| = 1 \\ \tilde C_{xy} := C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-\frac{1}{2}} \end{array}\right.$$

$\Rightarrow (\tilde u, \tilde v, \lambda) = svd(\tilde C_{xy})$

• CCA Steps:

i) Whitening

$$\left\{\begin{array}{l} \begin{bmatrix}\tilde u\\ \tilde v\end{bmatrix} = \begin{bmatrix}C_{xx} & \\ & C_{yy}\end{bmatrix}^{+\frac{1}{2}}\begin{bmatrix}u\\ v\end{bmatrix} \\ \begin{bmatrix}\tilde X & \tilde Y\end{bmatrix} := \begin{bmatrix}X & Y\end{bmatrix}\begin{bmatrix}C_{xx} & \\ & C_{yy}\end{bmatrix}^{-\frac{1}{2}} \end{array}\right.$$

$\tilde X, \tilde Y$ are decorrelated & isotropic along all directions in their own space

$\Rightarrow \rho(Xu, Yv) = \rho(\tilde X \tilde u, \tilde Y \tilde v)$

$C_{\tilde X \tilde X} = C_{\tilde Y \tilde Y} = I,\ C_{\tilde X \tilde Y} = \tilde C_{xy}$

ii) Align by correlation

$\max_{u,v} \rho(u,v; X,Y) \Leftrightarrow \max_{\tilde u, \tilde v} \rho(\tilde u, \tilde v; \tilde X, \tilde Y)$

$(\tilde u, \tilde v, \lambda) = svd(\tilde C_{xy})$

iii) Back-Projection $\begin{bmatrix}u\\ v\end{bmatrix} = \begin{bmatrix}C_{xx} & \\ & C_{yy}\end{bmatrix}^{-\frac{1}{2}}\begin{bmatrix}\tilde u\\ \tilde v\end{bmatrix}$

• k-CCA are orthogonal in whitened space, not in original

• CCA and Projection Regression:

• Training: given zero-meaned training data (X, Y)

i) $(U_{p\times k}, V_{q\times k}) = CCA(X, Y)$

ii) Compute CCA projections:

$X_c = XU,\ Y_c = YV$

Note that X_c, Y_c are zero-meaned.

iii) Fit a linear regressor A s.t. → projection

$Y_c = X_c A_{k\times k}$ $\quad A = U\underbrace{(U^TX^TXU)^{-1}}_{\text{decorrelation}}\underbrace{(U^TX^TYV)(V^TV)^{-1}V^T}_{\text{proj. back}}$

• Testing: given test data X

$\hat Y_c = XUA,\ \hat Y_c = \hat Y V$ whitening

$\Rightarrow \hat Y = \hat Y_c \underbrace{(V^TV)^{-1}V^T}_{k\times p}$

$k \le q$, $VV^T Y$ is singular.

• CCA simultaneously find projection directions in the two spaces such that the projected data have maximal correlation

PCA finds orthogonal projections that maximize variance in the single data space.

• To extract the relationship between vector-valued r.v. X and Y: $Z_X, Z_Y, Z_S \overset{iid}{\sim} N(0,1)$

$$\begin{bmatrix}X\\ Y\end{bmatrix} = \begin{bmatrix}A & B & 0\\ 0 & C & D\end{bmatrix}\begin{bmatrix}Z_X\\ Z_S\\ Z_Y\end{bmatrix}$$ characterizes randomness

· Singular value weight for OLS; PCA; Ridge:

☆ $w = \sum v_i \rho(s_i) u_i^T y = VDU^Ty$, D diagonal

Dual view: $w \in Span(v_i) = Range(X^T)$

☆ $w = X^T\alpha$.

OLS: $Xw = y \Rightarrow \alpha = U(SS^T)^\dagger U^T y$

Ridge: $(X^TX+\delta I)w = X^Ty \Rightarrow w = V(S^TS+\delta I)^{-1}S^TU^Ty$

~~(S invertible)~~ General solution $w + N(X)$

Kernel: $\alpha = (K+\delta I)^{-1} y + N(X^T)$

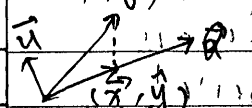
$\hat{z} = k(z, X)(ker(X,X)+\delta I)^{-1} y$

· LS: $Y + N_Y = Xw$, $N_Y \sim N(0, \sigma_Y^2)$

criterion: Minimize vertical projection distance

$(\vec{x}, y)$ $\min \varepsilon_{LS}(u) = \|y - Xu\|^2$

$= \|[y\ X]\begin{bmatrix}1\\-u\end{bmatrix}\|^2$



· $y = Xu \to u = X^\dagger y$

· $[y\ X]\begin{bmatrix}1\\u\end{bmatrix} = USV^T\begin{bmatrix}1\\u\end{bmatrix} = 0$

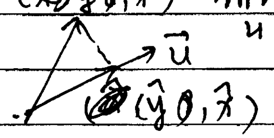
$\Rightarrow$ ~~SV^T~~ $\begin{bmatrix}1\\u\end{bmatrix} \propto V_{last}$

· u is the normal direction of the line

· TLS: $Y + N_Y = (X+N_X)w$, $N_X \sim N(0, \sigma_X^2)$, $N_Y \sim N(0,\sigma_Y^2)$

criterion: Minimize orthogonal projection distance

$(y, \vec{x})$ $\min_u \varepsilon_{TLS}(u) = \|[y, X] - [y, X]uu^T\|^2$



$= \|[y\ X](I - uu^T)\|^2$

$= \min_u \varepsilon_{PCA\_err}(u)$, $u^Tu=1$

$= C - \max_u \varepsilon_{PCA\_var}(u)$, $u^Tu = 1$

$= C - \max \|[y, X]u\|^2$, $u^Tu=1$

· TLS: PCA solution in joint (X,y) space

$[y\ X]u = USV^Tu \Rightarrow u = v_1$

· u is the direction of the line

· TLS: $y + \varepsilon_y = (X+\varepsilon_X) w$

$\min_{w,\varepsilon_X,\varepsilon_y} \|[\varepsilon_X\ \varepsilon_y]\|_F^2$ s.t. $(X+\varepsilon_X)w = y+\varepsilon_y$

i.e. $[X+\varepsilon_X\ \ y+\varepsilon_y]\begin{bmatrix}w\\-1\end{bmatrix} = 0$.

Compute SVD for $[X\ y]$ (assume full rank)

$[\varepsilon_X^*\ \varepsilon_y^*] = -\sigma_{d+1} u_{d+1} v_{d+1}^T$

$\begin{bmatrix}w^*\\-1\end{bmatrix} \propto v_{d+1}$ (last entry to -1, proved by contradiction given that $\sigma_{d+1}([X, y]) < \tilde\sigma_d(X)$)

$(v_{d+1}, \sigma_{d+1}^2) = eig([X\ y]^T[X\ y])$

$= eig(\begin{bmatrix}X^TX & X^Ty\\ y^TX & y^Ty\end{bmatrix})$

$\Rightarrow [X^TX\ \ X^Ty]\begin{bmatrix}w^*\\-1\end{bmatrix} = \sigma_{d+1}^2 w^*$

$\Rightarrow \hat w_{TLS} = (X^TX - \sigma_{d+1}^2 I)^{-1} X^T y$

(ridge regression with negative regularization)

· PCA: Maximum Variance View

$\max_{\|u\|=1} \varepsilon_{PCA\_var}(u) = \max_{\|u\|=1} u^TX^TXu$

K-PCA: $U = \arg\max \sum_{i=1}^k u_i^T X^TX u_i = \arg\max_{U^TU=I} tr(U^TX^TXU)$

Residue: $X(I - U_kU_k^T)$, $U^TU = I$

Rayleigh Quotient: $R(u; M) = \frac{u^TMu}{u^Tu}$

$\lambda_1 = R(v_1; M) \ge R(u; M) \ge R(v_d; M) = \lambda_d$

$\lambda_1 + \cdots + \lambda_k = R([v_1, \cdots, v_k]; M)$

$\ge \lambda_{d-k+1} + \cdots + \lambda_{d-1} + \lambda_d$

Covariance matrix View:

$\tilde X = XV$, $\tilde C = \frac1n \tilde X^T\tilde X = \frac1n \Sigma^2$ diagonal

$\Rightarrow$ features of $\tilde X$ are decorrelated

{ Gram $G = XX^T \rightsquigarrow$ sample similarity

{ Covariance $C = X^TX \rightsquigarrow$ feature similarity

orthogonal
$:= \|X\|_F^2 + \|XUU^T\|_F^2$

- $\Sigma_{PCA\_Err}(U) = \|X - \hat{X}\|_F^2 = \|X - XUU^T\|_F^2$
  $= tr(X(I-UU^T)^2 X^T)$
  $= tr(X(I-UU^T)X^T)$
  $= tr(XX^T) - tr(U^T X^T X U)$
  $= tr(XX^T) - U^T X^T X U$
  $= tr(XX^T) - \Sigma_{PCA\_Var}(U)$
- For Gaussian: Decorrelated = independent
- Joint Gaussian ~~zero-m~~
  - $Z \in \mathbb{R}^k$ is JG if: (equiv. def.)
    i) $\exists\, U \in \mathbb{R}^L,\ U_i \overset{i.i.d.}{\sim} N(0,1),\ R \in \mathbb{R}^{k \times L},\ \mu \in \mathbb{R}^k$
    s.t. $Z = RU + \mu$
    ii) $\forall\, \alpha \in \mathbb{R}^k$, $Z^T\alpha$ is normally distributed
    iii) (Non-degenerate case only)
    $f_U(u^T)$: $f_Z(z) = \frac{1}{\sqrt{|\det(\Sigma_Z)|}} \frac{1}{(\sqrt{2\pi})^k} e^{-\frac{1}{2}(z-\mu)^T \Sigma_Z^{-1}(z-\mu)}$
    $= |\det(R)| f_Z(Ru)$
    $\Sigma_Z = E[(Z-\mu)(Z-\mu)^T] = RR^T$ (covariance)
  - Level set: $f_Z(z) = k \rightsquigarrow x^T \Sigma^{-1} x = \alpha$
    $(\lambda_i, v_i) = eig(\Sigma_Z)$
    $x^T \Sigma^{-1} x$ is an ellipsoid with axes v_i, length $\sqrt{\lambda_i}$
  - $AZ \sim N(\mu_Z, \Sigma_Z)$
    $\Rightarrow AZ \sim N(A\mu_Z, A\Sigma_Z A^T)$

$f: \mathbb{R}^m \to \mathbb{R}^k$
$g: \mathbb{R}^n \to \mathbb{R}^m$
$J_{f\circ g}(x) = J_f(g(x))\, J_g(x)$
$k=1$: $\nabla f\circ g(x) = J_g(x)^T \nabla f(g(x))$

- $K(x_i, x_j)$ is a valid kernel if: (equiv. def)
  i) $\exists\, \phi(\cdot)$ s.t. $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$
  ii) $\forall\, D = \{x_1, \cdots, x_n\}$. Gram matrix $[K(D)]_{ij} = K(x_i, x_j)$ is PSD
- $k(x_i, x_j) = \alpha k_a(x_i, x_j) + \beta k_b(x_i, x_j)$, $\alpha, \beta \geq 0$
  $k(x_i, x_j) = \phi(x_i)^T \Sigma\, \phi(x_j)$, $\Sigma \succeq 0$ $(\tilde{\phi} = \Sigma^{\frac{1}{2}}\phi)$
  $k(x_i, x_j) = f(x_i) f(x_j) k_a(x_i, x_j)$ $(\tilde{\phi} = f \cdot \phi)$
  are valid kernels. $k(x_i, x_j) = k_1(x_i, x_j)\, k_2(x_i, x_j)$
- Moore-penrose pseudo inverse:
  $X^+ = \sum_{\sigma_i > 0} \sigma_i^{-1} v_i u_i^T \Rightarrow X^+ X = \sum_{\sigma_i > 0} v_i v_i^T$
  X^+X is an orthogonal projection onto the span of v_i i.e. Range(X^T).
  If rank$(X) = d$, then $X^+X = I$.
  If $d = n$, then $X^+ = X^{-1}$.
- $\frac{\partial(u \cdot v)}{\partial x} = \frac{\partial u^T v}{\partial x} = \frac{\partial u}{\partial x} v + \frac{\partial v}{\partial x} u$
  $\frac{\partial (Ax+b)^T C (Dx+e)}{\partial x} = D^T C^T (Ax+b) + A^T C (Dx+e)$
  $\frac{\partial (\|x-a\|)}{\partial x} = \frac{x-a}{\|x-a\|}$
  $\frac{\partial\, tr(AXBX^TC)}{\partial X} = A^T C^T X B^T + CAXB$
  $\frac{\partial \|XW - Y\|_F^2}{\partial W} = 2X^T(XW - Y)$

- $1 \le d < n$, $P \in \mathbb{R}^{n \times n}$ is a rank-d orthogonal projection matrix if (equiv. def):

  i) $\operatorname{rank}(P) = d$, $P = P^T$, $P^2 = P$

  ii) $\exists U \in \mathbb{R}^{n \times d}$ s.t. $P = UU^T$, $U^T U = I$

  $\circ$ $\forall v \in \mathbb{R}^n$, $Pv = \underset{w \in R(P)}{\operatorname{argmin}} \|v - w\|_2^2$

  ☆ $\operatorname{tr}(P) = d$

- Solve $\hat{H}|\psi\rangle = E|\psi\rangle$ by Differential Eqn / Alg. Method

- Angular Momentum: $\vec{L} = \vec{r} \times \vec{p}$

$$\begin{cases} L_x = y P_z - P_z y \\ L_y = z P_x - P_x z \\ L_z = x P_y - y P_x \end{cases}$$

$\rightarrow \hat{L} = -i\hbar (\vec{r} \times \vec{\nabla})$

$$\begin{cases} \hat{L_x} = -i\hbar\left(\hat{y} \frac{\partial}{\partial z} - z \frac{\partial}{\partial z} \hat{y}\right) \\ \hat{L_y} = -i\hbar\left(\hat{z} \frac{\partial}{\partial x} - \frac{\partial}{\partial x} \hat{z}\right) \\ \hat{L_z} = -i\hbar\left(\hat{x} \frac{\partial}{\partial y} - \hat{y} \frac{\partial}{\partial x}\right) \end{cases}$$

$[\hat{L_x}, \hat{L_y}]$

$= \cancel{-i\hbar} [yP_z, zP_x]$

$+ [zP_y, xP_z] - [yP_z, xP_z]$

$- [zP_y, zP_x]$

$= yP_x[P_z, z] + xP_y[z, P_z]$

$= i\hbar \hat{L_z}$

$$\begin{cases} [\hat{L_x}, \hat{L_y}] = i\hbar \hat{L_z} \\ [\hat{L_y}, \hat{L_z}] = i\hbar \hat{L_x} \\ [\hat{L_z}, \hat{L_x}] = i\hbar \hat{L_y} \end{cases}$$

$\Rightarrow$ Cannot measure more than one component of $\hat{L}$ simultaneously.

$L^2 := L_x^2 + L_y^2 + L_z^2$

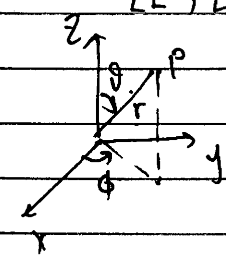
$[L^2, L_x] = [L_x^2, L_x] + [L_y^2, L_x] + [L_z^2, L_x]$

$= [L_x, L_x] L_x + L_x [L_x, L_x] + L_y [L_y, L_x] + [L_y, L_x] L_y$

$\quad + L_z [L_z, L_x] + [L_z, L_x] L_z$

$= -i\hbar (L_y L_z + L_z L_y) + i\hbar (L_z L_y + L_y L_z) = 0.$

$[L^2, L_y] = 0$, $[L^2, L_z] = 0$



$$\begin{cases} x = r\sin\theta\cos\varphi \\ y = r\sin\theta\sin\varphi \\ z = r\cos\theta \\ r \in [0, +\infty),\ \theta \in [0, \pi],\ \varphi \in [0, 2\pi] \end{cases}$$

$$\begin{cases} \hat{L_x} = -i\hbar\left(\sin\varphi \frac{\partial}{\partial \theta} - \cot\theta \cos\varphi \frac{\partial}{\partial \varphi}\right) \\ \hat{L_y} = -i\hbar\left(\cos\varphi \frac{\partial}{\partial \theta} - \cot^2\theta \sin\varphi \frac{\partial}{\partial \varphi}\right) \\ \hat{L_z} = -i\hbar \frac{\partial}{\partial \varphi} \end{cases}$$

$\hat{L}^2 = -\hbar^2 \left[\frac{1}{\sin\theta} \frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial}{\partial \theta}\right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \varphi^2}\right]$