PROB

INTEGRALS FOR MULTIVAR. NORMAL

```r
library(mvtnorm)

# Mean vector: (mu_1, ..., mu_n)
mean <- c(0, 1)
# Covariance matrix
cov <- matrix(c(1, 2,
                2, 4), nrow = 2, ncol = 2)

lower_bounds <- c(-Inf, -Inf)
# Upper bounds
upper_bounds <- c(2, 1)

# Compute the probability
probability <- pmvnorm(
    lower = lower_bounds,
    upper = upper_bounds,
    mean = mean,
    sigma = cov
)
# paste(): concatenate float and str
print(paste("Probability:", probability))
```

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- **MULTIVARIATE NORMAL / MULTINOMIAL:**
  - **MULTINOMIAL DISTRIBUTION:** ~> GENERALIZATION OF BINOMIAL EXPERIMENT: 1 CLASS ⟶ K+1 CLASSES

GIVEN $K+1$ CLASSES AS POSSIBLE OUTCOMES, IF WE MAKE n CHOICES INDEPENDENTLY WITH THE SAME CLASS PROBABILITIES $\pi_1, \dots, \pi_{K+1}$ / $\pi_{K+1} = 1 - \sum_{j=1}^{K} \pi_j$

$$\to P\left(\underbrace{X_1 = x_1, \dots, X_{K+1} = x_{K+1}}_{\text{MULTINOMIAL RANDOM VECTOR}}\right) = \underbrace{\frac{n!}{x_1! \cdot \dots \cdot x_{K+1}!}}_{\text{MULTINOMIAL COEFFICIENT}} \cdot \pi_1^{x_1} \cdot \dots \cdot \pi_{K+1}^{x_{K+1}} \Big/ n = \sum_{j=1}^{K} x_j + x_{K+1}$$

(R.V. ~> OUTCOME)

- $\begin{cases} K+1: & n^\circ \text{ OF MUTUALLY EXCLUSIVE EVENTS } (P(A \cap B) = 0) \\ n: & n^\circ \text{ OF TRIALS} \end{cases} \Big/ K: n^\circ \text{ R.V.s}$

(2D. BINOMIAL: $K=1$ ~> p, $(1-p)$)

- **SUMMARY:**
  - $\vec{X} \sim \text{MULTINOMIAL}(n, \tilde{\pi}_1, \dots, \tilde{\pi}_{K+1})$
  - $X_{K+1} = n - \sum_{i=1}^{k} X_i$
  - X_i: INTEGERS IN $0 \le X_i \le n$

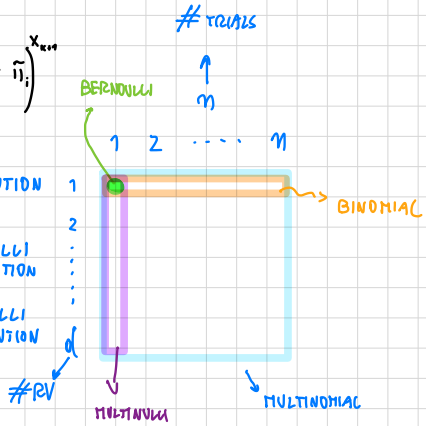
$$\to f(x_1, \dots, x_K) = \binom{n}{x_1, \dots, x_{K+1}} \tilde{\pi}_1^{x_1} \cdot \tilde{\pi}_2^{x_2} \cdot \dots \cdot \tilde{\pi}_K^{x_K} \cdot \left(1 - \sum_i \tilde{\pi}_i\right)^{x_{K+1}}$$

(FROM MULTINOMIAL)

$$E(\vec{X}) = \begin{pmatrix} \tilde{\pi}_1 \\ \vdots \\ \tilde{\pi}_m \end{pmatrix}$$

$$V(X_i) = \tilde{\pi}_i(1 - \tilde{\pi}_i)$$

- $K = 1$ (1 R.V.): $f(x_1) = \binom{n}{x_1} \cdot \tilde{\pi}_1^{x_1} \cdot (1 - \tilde{\pi}_1)^{x_{K+1}}$ ($m - x_1$) ⟶ BINOMIAL DISTRIBUTION
- $n = 1$ (# EXTRACTIONS IN EXPERIMENT = 1): $f(x_1, \dots, x_K) = \tilde{\pi}_1^{x_1} \cdot \dots \cdot \tilde{\pi}_K^{x_K} \cdot \left(1 - \sum_j \tilde{\pi}_j\right)^{1 - \sum x_j}$ ⟶ MULTINULLI DISTRIBUTION
- $K = 1, n = 1$: $f(x_1) = \begin{cases} 1, & p = \tilde{\pi}_1 \\ 0, & p = 1 - \tilde{\pi}_m \end{cases}$ ⟶ BERNOULLI DISTRIBUTION



- **PROPERTIES:**
  - IF $\vec{X} \sim \text{MULTINOMIAL}(n, \tilde{\pi}_1, \dots, \tilde{\pi}_{K+1}) \to \vec{X} = \sum_{i=1}^{K} n_i \Big/ n_i \sim \text{MULTINULLI}(\tilde{\pi}_1, \dots, \tilde{\pi}_K)$
  - $E(\vec{X}) = \sum_{i=1}^{K} E(n_i) = n \begin{pmatrix} \tilde{\pi}_1 \\ \vdots \\ \tilde{\pi}_K \end{pmatrix}$, $V(X_i) = n\tilde{\pi}_i(1 - \tilde{\pi}_i)$
  - $\text{CoV}(X_i, X_j) = -n\tilde{\pi}_i\tilde{\pi}_j$
  - $E[\tilde{a}\vec{X} + \tilde{b}] = \tilde{a}E[\vec{X}] + \tilde{b}$
  - $\text{Var CoV}(\underset{l \times K}{A} \cdot \underset{K+1}{\vec{X}}) = A \cdot \text{Var CoV}(\vec{X}) \cdot A^T$

- **MULTIVARIATE NORMAL:**

$$\text{Var CoV}(\vec{X}) = \Sigma_X$$

- $\vec{Z} \sim N_k(\vec{0}, I_k) \to X = \underset{n \times k}{A} \cdot \underset{k \times 1}{\vec{Z}} + \underset{n \times 1}{\vec{\mu}} \Big/ E(\vec{X}) = \vec{M}$, $\Sigma_X = A \cdot A^T$

POP (~> IN $\mathbb{R}^2$):

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right)$$

- CONDITIONAL NORMAL DISTRIBUTION: $X_2 | X_1 = x_1 \sim N_m\left(\vec{\mu}_2 + \Sigma_{21} \cdot \Sigma_{11}^{-1}(\vec{x}_1 - \vec{\mu}_1), \Sigma_{22} - \Sigma_{21} \cdot \Sigma_{11}^{-1}\Sigma_{12}\right)$ (IF $\exists \Sigma_{11}^{-1}$)

IF $n = 2$ (IN $\mathbb{R}^2$) $\Leftrightarrow \rho \neq \pm 1$

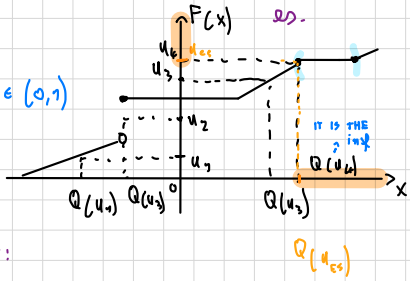
- $n_1 = 1, n_2 = 1$: $Y | X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right)$

- $P(\underbrace{X \le x_1 \wedge Y \le y_1}_{A}) = \int_{-\infty}^{x_1}\int_{-\infty}^{y_1} f(x,y)\,dx\,dy = \int_{-\infty}^{x_1} f_X(x) \underbrace{\left[\int_{-\infty}^{y_1} f_{Y|X}(y|x)\,dy\right]}_{\text{CDF OF A NORMAL}} dx$ ($P(x,y) = P(x)P(y|x)$)

• SIMULATION:

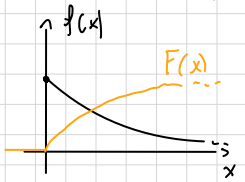
- IF X HAS CDF $F(x) = P(X \le x)$

→ QUANTILE FUNCTION: $Q(u) = \inf\{x : F(x) \ge u\}$ $\forall u \in (0,1)$



ex.

$X \sim Exp(\lambda) \rightarrow f_X(x) = \lambda e^{-\lambda x}, \; x>0$



- $F_X(x) = \int_0^x \lambda e^{-\lambda x} = -e^{\lambda t}\Big|_0^x = 1 - e^{-\lambda t}$

$\left(y = 1 - e^{-\lambda x} \Rightarrow 1 - y = e^{-\lambda x} \Rightarrow \ln(1-y) = -\lambda x \rightarrow x = -\frac{\ln(1-y)}{\lambda}\right)$

- $Q(u) = F_X^{-1}(x)$ ~> (NOT SURE AS NOTATION)

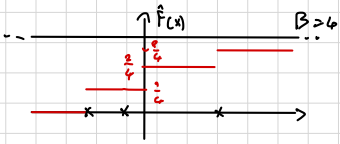
→ SIMULATION:

1. GET $U \sim$ UNIFORM $(0,1)$
2. RETURN $X = -\frac{\ln(1-U)}{\lambda}$

TO SIMULATE $X \sim Exp(\lambda)$

- HOW TO ESTIMATE $F(x_0)$:

1. GENERATE $x_1, \dots, x_B$ iid X
2. $\hat{F}(x_0) = \frac{1}{B}\sum_{i=1}^{B}(x_i \le x_0)$



- N.B.: IMPORTANT RULES FOR MATRIX DIFFERENTIATION
  - $\frac{\partial}{\partial x} A\bar{x} = A^T$
  - $\frac{\partial}{\partial x} x^T A x = 2Ax$

• LINEAR MODELS:

- β_0 is the average response when factor is set at the *reference* level.
- $\beta_i, \forall i \in [0, p-1]$ are the **differential mean** responses of other levels with respect to the reference level.

→ $Y_i = N(\beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_{p-1} \cdot x_{i,p-1}, \underbrace{\sigma^2}_{\text{ERROR VARIANCE}}) \rightarrow \bar{Y} \sim N(X\bar{\beta}, \sigma^2 I)$

→ $\hat{\beta} = (\bar{X}^T\bar{X})^{-1} \cdot \bar{X}^T \cdot \bar{Y}$ LS ESTIMATE OF β, $\xrightarrow[\text{REG.}]{\text{LINEAR}}$ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$, $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}$ ($\bar{x}$: MEAN)

→ $\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^T(Y - X\hat{\beta})}{n} = \frac{e^T \cdot e}{n}$ (e: RESIDUAL $\bar{e}$)

$e^T \cdot e$: SUM OF SQUARED RESIDUALS $= (Y - X\hat{\beta})^2 = (Y - \hat{Y})^2$

OR $S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$ OR $MSR = \frac{e^T \cdot e}{n-p}$ (p: # FEATURES)

→ A SINGLE COMPONENT: $\hat{\beta}_i \sim N(\beta_i, \underbrace{\sigma^2 (X^TX)^{-1}_{i+1,i+1}}_{\text{SQUARED OF STANDARD ERROR OF } \hat{\beta}_0}), \; i = 0, \dots, p-1$

IT WILL BE ESTIMATED WITH MSR

$SE(\hat{\beta}_i) = \sqrt{\sigma^2 (X^TX)^{-1}_{i+1,i+1}}$

- CI:

$$\frac{\hat{\beta}_i - \beta_i}{\underbrace{\sqrt{\sigma^2 (X^TX)^{-1}_{i+1,i+1}}}_{SE(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\frac{E^TE}{n-p}\cdot\frac{1}{\sigma^2}\,(X^TX)^{-1}_{i+1,i+1}}} \quad (N(0,1), \; \chi^2_{n-p}) = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} \overset{\text{THEOREM}}{\sim} t_{n-p}$$

($\frac{E^TE}{\sigma^2}$: σ^2 IS ESTIMATED WITH MSR)

- C.I. FOR $\hat{\sigma}^2$:

$C.I._{1-\alpha}$: $P\left[\chi^2_{1-\frac{\alpha}{2},n-p} < \frac{E^TE}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-p}\right]$

- C.I. FOR β:

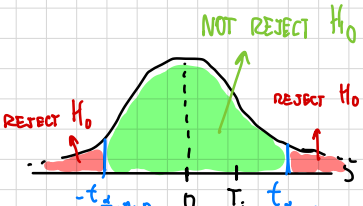
- t-STUDENT C.I. FROM ELEMENTARY STATISTICS: $SE(\hat{\beta}_i) = \sqrt{\sigma^2 (X^TX)^{-1}_{i+1,i+1}} = \sqrt{\frac{S^2}{n}}$

→ $C.I._{1-\alpha} = \left[\beta_i - t_{\frac{\alpha}{2},n-p} \cdot SE(\hat{\beta}_i) \; ; \; \beta_i + t_{\frac{\alpha}{2},n-p} \cdot SE(\hat{\beta}_i)\right]$

- SINCE: $P\left[\beta_i - t_{\frac{\alpha}{2},n-p} \cdot SE(\hat{\beta}_i) \le \hat{\beta}_i \le \beta_i + t_{\frac{\alpha}{2},n-p} \cdot SE(\hat{\beta}_i)\right] = 1 - \alpha$

- H_0 TESTING: ~> H_0: $\beta_1, \dots, \beta_{p-1} = 0$, $\beta_0 \neq 0$

- REJECT H_0 IF: $T_i = \left|\frac{\hat{\beta}_i - \overset{H_0}{\beta_i^0}}{SE(\hat{\beta}_i)}\right| > t_{\frac{\alpha}{2},n-p}$



→
- p-value ≤ α → H_0 REJECTED
- p-value > α → H_0 NOT REJECTED

↓ p-value
- ***: STRONGLY REJECT H_0
- ⋮
- 'NULL': NOT REJECT H_0

- **F TEST** ; ~> TEST SMALL VS BIG MODEL (IS SIGNIFICANT TO UPGRADE?)

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{p-1} \Big/ \frac{\|Y - \hat{Y}\|^2}{n-p} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{p-1} \Big/ \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p}$$

(numerator: EXPLAINED VARIANCE; denominator: UNEXPLAINED VARIANCE)

$\longrightarrow$ H_0 REJECTED $\Leftrightarrow$ $F > F_{\alpha, DOF_1, DOF_2, \text{UNDER } H_0}$

L> FROM COMPUTERS : REJECT IF p-VALUE $< \alpha$

- IN R:

$DOF_1 = n - p_1 - 1 = 200 - 2 - 1 = 197$ | $DOF_2 = n - p_2 - 1 = 200 - 3 - 1 = 196$

$RSS = \sum(Y_i - \hat{Y}_i)^2 = e^T e$ | $DOF = DOF_2 - DOF_1$ | SUM OF SQ $= RSS_2 - RSS_1$

$\longrightarrow F = 0.0036 \longrightarrow P[X > 0.036] \approx 0.95 =$ p-value $> \alpha \rightarrow H_0$ NOT REJ

```
# Perform the F-test
f_test <- anova(reduced_model, full_model)
```

```
Analysis of Variance Table

Model 1: Sales ~ TV + Radio
Model 2: Sales ~ TV + Radio + Newspaper
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 541.21
2    196 541.20  1 0.0092861 0.0034 0.9538
```

- **PREDICTIONS** :

- FITTING VALUE OF X_f IS : $\hat{Y}_f = X_f \hat{\beta} \longrightarrow \hat{Y}_f = X_f \hat{\beta} \sim N\left(X_f \beta, \sigma^2 \cdot X_f (X^T X)^{-1} X_f^T\right)$

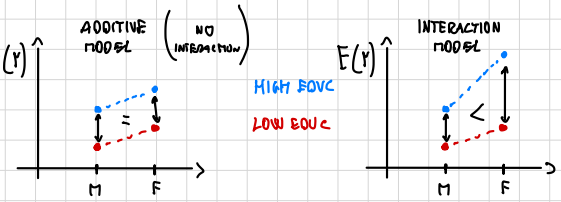
$\rightarrow$ C.I. : $X_f \cdot \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p} \cdot \sqrt{\frac{e^T e}{n-p} \cdot X_f (X^T X)^{-1} X_f^T}$

$\rightarrow$ PREDICTION INTERVAL : $X_f \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p} \cdot \sqrt{\frac{E^T E}{n-p}\left(1 + X_f (X^T X) X_f^T\right)}$

L> A BIT LARGER THEN C.I.

- **GENERAL INTERACTION MODEL** :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$



- ANOVA (SMALL, LARGE) GIVES A TEST FOR NULL H_0, H_0: ALL INTERACTIONS = 0

- **MEASURES OF FIT** :

$$R^2 = \frac{NUM}{DEN} = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}_0\|^2} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

(FRED. = $\hat{Y}$; GT = Y; MEAN OF DATA = $\bar{Y}$)

PYTHAGORA EXTENDED TO $\mathbb{R}^n$

- NUM: $\|\hat{Y} - \hat{Y}_0\|^2 = \|Y - \hat{Y}_0\|^2 - \|Y - \hat{Y}\|^2 = TSS - E^T E$ RSS

- DEN: $\|Y - \hat{Y}_0\|^2 = E_0^T E_0 = TSS$ TOTAL SUM OF SQUARES

$$R^2_{ADJ} = 1 - \frac{n-1}{n-p} \cdot \frac{RSS}{TSS}$$

- AIC $= 2k - 2 \cdot \ln(\hat{L})$
- BIC $= k \cdot \ln(n) - 2 \ln(\hat{L})$

k: # ESTIMATED PARAMETER IN THE MODEL ($\rightarrow p+1$ IN LINEAR MODELS ($\beta_0, \dots, \beta_{p-1}, \sigma^2$))

$\hat{L}$: MAXIMUM VALUE OF LIKELIHOOD FUNCTION

n: # DATA POINTS

$\longrightarrow$ $\downarrow$ AIC, BIC $\rightarrow$ BETTER MODEL

- METHODS FROM `lm()` CLASS: e.g. `summary(my_fit)`

- **GLM** ; `my_fit <- glm(y~predictors, family=binomial)`

Family	Default link function
binomial	(link = "logit") → logistic regression
gaussian	(link = "identity") → LINEAR MODELS
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log") → for count variables, like number of accidents
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

- ODDS : $\frac{p}{1-p}$ / $[0,1] \longmapsto [0, +\infty]$

- LOGIT (LOG-ODDS): $\log\left(\frac{p}{1-p}\right)$ / $[0,1] \longmapsto [-\infty, +\infty]$

- GLM FOR p: $\begin{pmatrix} \text{logit}(p_1) \\ \vdots \\ \text{logit}(p_n) \end{pmatrix} = X \cdot \beta$

$E(Y) = E(X\hat{\beta}) = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$ L> $Y_i \sim \text{Bernulli}(p_i)$

$\text{logit}(p) \longmapsto p$: $[-\infty; +\infty] \longmapsto [0,1]$

$p = \frac{e^l}{1+e^l} \rightarrow p_i = \frac{e^{X\beta_{(i)}}}{1+e^{X\beta_{(i)}}}$

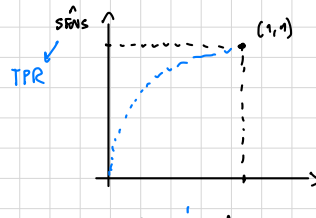
Function	Description
summary()	Displays detailed results for the fitted model
coefficients(), coef()	Lists the model parameters (intercept and slopes) for the fitted model
confint()	Provides confidence intervals for the model parameters (95 percent by default)
residuals()	Lists the residual values for a fitted model
anova()	Generates an ANOVA table comparing two fitted models
plot()	Generates diagnostic plots for evaluating the fit of a model
predict()	Uses a fitted model to predict response values for a new datase

- CONFUSION MATRIX:

TRUE \ PREDICTED	1	0
1	TP	FN
0	FP	TN

- SENSITIVITY = RECALL $(c=1)$ = $P(\hat{Y}=1 \mid Y=1)$
  - ↳ $\hat{\text{SENSITIVITY}} = \frac{TP}{TP+FN} = TPR$
- SPECIFICITY = RECALL $(c=0)$ = $P(\hat{Y}=0 \mid Y=0)$
  - ↳ $\hat{\text{SPECIFICITY}} = \frac{TN}{TN+FP} = 1-FPR$

- IF WE VARY t -> CHANGE OF SENSITIVITY AND SPECIFICITY
- ROC CURVE:



ROC CURVE: $\hat{SENS} = f(1-\hat{SPEC})$

$$f_{X|Y}(X=x \mid Y=y) = \frac{f(x,y)}{f_Y(y)} \quad \left[\int f(x,y)\,dx \right] \rightarrow \text{MARGINAL DENSITY OF } Y$$

$$f_{X|Y=y}(x|y) = \frac{f_X(x) \cdot f_{Y|X=x}(y|x)}{\int f_X(t) \cdot f_{Y|X=t}(y|t)\,dt}$$

- DAGs:

$$f(x_1,\dots,x_d) = \prod_i f_{x_i|\text{PARENTS}(x_i)}(x_i \mid \text{PARENTS}(x_i))$$

- MARKOVIAN CHAINS:

$X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_d$:

$$f(x_1,\dots,x_d) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1) \cdot f_{X_3|X_2}(x_3|x_2) \cdot \dots \cdot f_{X_d|X_{d-1}}(x_d|x_{d-1})$$

$X \rightarrow Y \leftarrow W \rightarrow Z$ $\quad Y = \text{PARENTS}(W,Z)$ $\quad X = \text{CHILDS}(Y)$

$$\rightarrow f(x,y,w,z) = f_{Z|Y}(z|y) \cdot f_{W|Y}(w|y) \cdot f_{Y|X}(y|x) \cdot f_X(x)$$

- INDEPENDENT RVs:

X_1 X_2 X_d : $f(x_1,\dots,x_d) = \prod_i f_{X_i}(x_i)$

- CONDITIONAL EXPECTATION:

$$E[Y|X=x] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x-\mu_X) \Big/ \rho \frac{\sigma_Y}{\sigma_X} = \frac{Cov(X,Y)}{V(X)}$$

$$V[Y|X=x] = V(Y) - \frac{Cov^2(X,Y)}{V(X)}$$

- $Cov(aX+b, Y) = a \cdot Cov(X,Y)$
- $Cov(aX+bY, cV+dW) =$
  $= ac\,Cov(X,V) + ad\,Cov(X,W) + bc\,Cov(Y,V) + bd\,Cov(Y,W)$

1. $E[E[Y|X]] = E[Y]$
2. $E(a|X) = a$ / a = const.
3. $E(aX+bY|Z) = a\,E(X|Z) + b\,E(Y|Z)$
4. $E(g(X)\cdot Y|X) = g(X)\cdot E(Y|X)$

$$V(Y|X) = E\left[(Y-E(Y|X))^2 \mid X\right]$$

5. $E[V(Y|X)] = V(Y) - V(E(Y|X))$
   ↳ $V(Y) = E[V(Y|X)] + V(E(Y|X))$

- BAYESIAN STATISTICS:

$$\underbrace{\Pi(\theta|\text{DATA})}_{\text{POSTERIOR}} = \frac{f(\text{DATA}|\theta) \cdot \overbrace{\Pi(\theta)}^{\text{PRIOR}}}{\int \Pi(t) \cdot f(\text{DATA}|t)\,dt} \propto f(\text{DATA}|\theta) \cdot \Pi(\theta)$$

$\tau = \frac{1}{\sigma^2}$: PRECISION

$\tau_0 = \frac{1}{\sigma_0^2}$: PRIOR PRECISION

	p.d.f.	$\mathbb{E}(x)$	$\text{Var}(x)$
$x \sim \text{Beta}(\alpha,\beta)$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$x \sim \text{Gamma}(\alpha,\beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

- $B(\alpha,\beta) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$ / $\Gamma(t) = \int_0^\infty t^{x-1}e^{-t}\,dt = (x-1)\,\Gamma(x-1)$

Examples

(1) $\theta \sim Beta(\alpha,\beta)$ is the conjugate prior to $X|\theta \sim Bin(n,\theta)$ with kernel:

$$q(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

(2) $\theta \sim N(\mu_0,\sigma_0^2)$ is the conjugate prior to $X|\theta \sim N(\theta,\sigma^2)$ with kernel:

$$q(\theta) = \exp\left(-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right)$$

(3) $\sigma^2 \sim InvGamma(\alpha,\beta)$ is the conjugate prior to $X|\sigma^2 \sim N(\theta,\sigma^2)$ with kernel:

$$q(\sigma^2) = (\sigma^2)^{-(\alpha+1)}\exp\left(-\frac{\beta}{\sigma^2}\right)$$

(4) $\lambda \sim Gamma(\alpha,\beta)$ is the conjugate prior to $X|\lambda \sim Poi(\lambda)$ with kernel:

$$q(\lambda) = \lambda^{\alpha-1}\exp(-\beta\lambda)$$

1. Let $X|\theta \sim Bin(n,\theta)$ with conjugate prior $\theta \sim Beta(\alpha,\beta)$, then

$$\theta|X \propto Beta(\alpha+X, \beta+n-X) \;:\; \mathbb{E}(\pi(\theta|X)) = \frac{\alpha+X}{\alpha+\beta+n}$$

2. Let $X|\theta \sim N(\theta,\sigma^2)$ with conjugate prior $\theta \sim N(\mu_0,\sigma_0^2)$, then

$$\theta|X \propto N(\mu_1,\sigma_1^2) \;:\; \mathbb{E}(\pi(\theta|X)) = \left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{X}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$$

3. Let $X|\sigma^2 \sim N(\theta,\sigma^2)$ with conjugate prior $\sigma^2 \sim InvGamma(\alpha,\beta)$, then

$$\sigma^2|X \propto InvGamma(\alpha+1/2, \beta+(X-\mu)^2/2) \;:\; \mathbb{E}(\pi(\sigma^2|X)) = \frac{\beta+(X-\mu)^2/2}{\alpha-1/2}$$

4. Let $X|\lambda \sim Poi(\lambda)$ with conjugate prior $\lambda \sim Gamma(\alpha,\beta)$, then

$$\lambda|X \propto Gamma(X+\alpha, 1+\beta) \;:\; \mathbb{E}(\pi(\lambda|X)) = \frac{X+\alpha}{1+\beta}$$

- $X_1, \dots, X_n \mid p \overset{iid}{\sim} \text{Bernoulli}(p)$ / $p \sim \text{Beta}(a, b)$

- CONDITIONAL DENSITY:

$$f(x_1, \dots, x_n \mid p) = \prod_{i=1}^{n} f_{X_i \mid p}(x_i \mid p) = p^{\sum x_i} \cdot (1-p)^{n - \sum x_i}$$

↳ $\mathcal{L}(p, x_1, \dots, x_n)$ ↳ LIKELIHOOD FUNCTION

- BAYES POSTERIOR:

$$\Pi(p \mid x_1, \dots, x_n) = \frac{\tilde{\Pi}(p) \cdot f(x_1, \dots, x_n \mid p)}{\int_0^1 \tilde{\Pi}(q) \cdot f(x_1, \dots, x_n \mid q)\, dq} \overset{\text{CONST. PROP.}}{=} \frac{p^{a-1}(1-p)^{b-1}}{\left[\int_0^1 t^{a-1}(1-t)^{b-1}\, dt\right]} \cdot p^{\sum x_i} \cdot (1-p)^{n - \sum x_i}$$

$$\propto p^{a-1}(1-p)^{b-1} \cdot p^{\sum x_i} (1-p)^{n - \sum x_i} \overset{\text{CONST. PROP.}}{\propto} \left[\int_0^1 \tilde{\Pi}(q) \cdot f(x_1, \dots, x_n \mid q)\, dq\right]$$

IT IS A BETA AGAIN

$$= p^{a-1}(1-p)^{b-1} \cdot p^{\sum x_i} (1-p)^{n - \sum x_i} = \underbrace{p^{a + \sum x_i - 1} (1-p)^{b + n - \sum x_i - 1}}$$

$$\rightarrow \Pi(p \mid x_1, \dots, x_n) = \frac{p^{a + \sum x_i - 1} (1-p)^{b + n - \sum x_i - 1}}{\int_0^1 t^{a + \sum x_i - 1} (1-t)^{b + n - \sum x_i - 1}\, dt} \rightarrow p \mid x_1, \dots, x_n \sim \text{Beta}\left(a + \sum_i x_i,\ b + n - \sum x_i\right)$$

- BAYESIAN POINT ESTIMATION:

  - NORMAL RANDOM SAMPLE:

$$E(\mu \mid x_1, \dots, x_n) = \frac{\tilde{\tau}_0 \mu_0 + n\tilde{\tau}\bar{X}}{\tilde{\tau}_0 + n\tilde{\tau}} = \frac{\tilde{\tau}_0}{\tilde{\tau}_0 + n\tilde{\tau}} \mu_0 + \frac{n\tilde{\tau}}{\tilde{\tau}_0 + n\tilde{\tau}} \bar{X} =$$

WEIGHTED AVERAGE OF: μ_0: PRIOR MEAN, $\bar{X}$: SAMPLE MEAN

$$= \left(1 - \frac{n\tilde{\tau}}{\tilde{\tau}_0 + n\tilde{\tau}}\right) \mu_0 + \frac{n\tilde{\tau}}{\tilde{\tau}_0 + n\tilde{\tau}} \bar{X} \quad / \quad W = \frac{n\tilde{\tau}}{\tilde{\tau}_0 + n\tilde{\tau}}$$

  - $\downarrow \tilde{\tau}_0$ ($\rightarrow$ HIGHER σ_0^2: HIGHER UNCERTAINTY) $\rightarrow \uparrow$ DATA IMPORTANCE FOR ESTIMATION

  - BINARY DATA CASE, $\text{Beta}(a, b)$:

$$E(p \mid x_1, \dots, x_n) = \frac{a + \sum x_i}{a + \sum x_i + b + n - \sum x_i} = \frac{a + \sum x_i}{a + b + n} = \frac{a + b}{a + b + n} \cdot \frac{a}{a + b} + \frac{n}{a + b + n} \cdot \frac{\sum x_i}{n} =$$

SAME PROPERTIES AS ABOVE

$$= \left(1 - \frac{n}{a + b + n}\right) \frac{a}{a + b} + \frac{n}{a + b + n} \bar{X}$$

↳ $\hat{p}$: MLE OF p

  - IF $a = b = 1$: $E(p \mid x_1, \dots, x_n) = \frac{1 + \sum x_i}{2 + n} = \tilde{p}$ BAYES ESTIMATOR

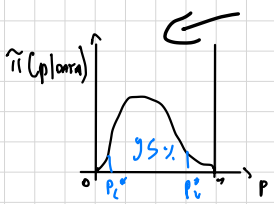
↳ IT GIVES UNIFORM ON p: GOOD PRIOR IGNORANCE

CREDIBILITY INTERVAL

$$\begin{cases} \dfrac{p_L^3 (1-p_L)^7}{\int t^3 (1-t)^7\, dt} = \dfrac{p_U^3 (1-p_U)^7}{\int t^3 (1-t)^7\, dt} \\ \displaystyle\int_{p_L}^{p_U} \frac{p^3 (1-p)^7}{\int_0^1 t^3 (1-t)^7\, dt}\, dp = 0.95 \end{cases}$$

$\rightarrow$ DISTRIBUTION OF A BETA

$$\begin{cases} d\text{Beta}(p_L, 4, 8) = d\text{Beta}(p_U, 4, 8) \\ p\text{beta}(p_U, 4, 8) - p\text{beta}(p_L, 4, 8) = 0.95 \end{cases}$$



ML

- IF A IS SYMMETRIC $\begin{cases} \longrightarrow \lambda_i \in \mathbb{R}, \text{ EIGENVECTORS ORTHOGONAL} \\ \longrightarrow x^T A x \geq 0 \end{cases}$, A IS DIAGONALIZABLE (SPECTRAL THEOREM)

MATRIX PROPERTIES:

- $AB \neq BA$
- $A(B+C) = AB + AC$
- $(A_1 \cdot \ldots \cdot A_m)^T = A_m^T \cdot \ldots \cdot A_1^T \rightarrow \begin{cases} (AB)^T = B^T A^T \\ (A^T B C)^T = C^T B^T A \end{cases}$
- $(A+B)^T = A^T + B^T$
- $Ax = B \rightarrow$ IF $\exists A^{-1}$: $x = A^{-1} B$

ORTHOGONAL MATRIX, DIAGONAL MATRIX

$$A = Q \begin{bmatrix} A & B \\ 0 & C \end{bmatrix} Q^T$$

- ORTHONORMAL BASIS :

$$\{\bar{v}_1, \ldots, \bar{v}_n\} \;/\; \langle \bar{v}_i, \bar{v}_j \rangle = 0 \quad i \neq j$$

PAC LEARNING:

VC DIM $= d < \infty \rightarrow$ CANNOT LEARN FROM $\frac{m}{2}$ SAMPLES

VC DIM $= \infty \rightarrow \mathcal{H}$ NOT PAC LEARNABLE

$1-\delta \rightarrow$ PROBABLY, $\leq \varepsilon \rightarrow$ APPROXIMATELY CORRECT LEARNING

$\begin{cases} \delta \rightarrow P \text{ OF SUCCESS} \\ 1-\delta \rightarrow P \text{ OF FAILURE} \end{cases}$

- THE LEARNER DO NOT KNOW D AND f
- THE LEARNER RECIEVES ε (ACCURACY PARAMETER) AND δ (CONFIDENCE PARAMETER)
- THE LEARNER CAN ASK FOR TRAINING DATA S, WITH m SAMPLES / m (.) ε, δ
- THE LEARNER SHOULD OUTPUT AN HYPOTHESIS h WITH PROBABILITY OF AT LEAST $1-\delta$ / IT HOLDS THAT $L_{D,f}(h) \leq \varepsilon \rightarrow P_{S \sim D^m}\left[L_{D,f}(h) \leq \varepsilon\right] \geq 1-\delta$

- THEOREM (NO FREE LUNCH):

FIX $\delta \in (0,1)$, $\varepsilon < 1/2$:

$\rightarrow \forall$ LEARNER A AND TRAINING SIZE m :

$\rightarrow \exists D, f$ / WITH $p \geq \delta$ OF m EXAMPLES, OVER TRAINING DATA S

IT HOLDS THAT $L_{D,f}(A(S)) \geq \varepsilon$

- $\forall$ LEARNER $\rightarrow \exists$ A FUNCTION WHICH DO NOT SUCCEEDS IN $L_{0,f} \leq \varepsilon$
- $\nexists$ A UNIQUE LEARNER : EVERY TIME A SPECIFIC ONE IS NEEDED
- THE AVG RESULT OF ALL CLASSIFIER IS WORST THEN RANDOM GUESSING
  - $\rightarrow L_{D,f}(\text{RANDOM GUESSING}) = 1/2$
- IF NO KNOWLEDGE IS INSERTED $\rightarrow \nexists$ SOMETHING BETTER THEN RANDOM GUESSING

- PRIOR KNOWLEDGE : IF LEARNER KNOWS $\mathcal{H} \subset Y^X$ ($\mathcal{H}$: HYPOTHESIS CLASS OF f) $\rightarrow$ PAC LEARNING IS POSSIBLE IF $\mathcal{H} \neq Y^X$ (NO FREE LUNCH THEOREM)

- PAC LEARNING OF FINITE HYPOTHESIS CLASS : ASSUME $\mathcal{H}$: FINITE HYPOTHESIS CLASS (ex. ALL $h : X \mapsto Y$)

- EMPIRICAL RISK MINIMIZATION (ERM): $\rightarrow h = ERM_{\mathcal{H}}(S)$

ERM LEARNING RULE:

- INPUT : TRAINING SET $S = (x_1, y_1), \ldots, (x_m, y_m)$
- EMPIRICAL RISK : $L_S(h) = \frac{1}{m} \left| \{ i : h(x_i) \neq y_i \} \right|$ (0-1 LOSS)
- OUTPUT : ANY $h \in \mathcal{H}$ / h MINIMIZES $L_S(h)$

- THEOREM:

FIX ε, δ, IF $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$: h / SATISFIES ERM OVER THE CHOICE OF S SIZE m

$\rightarrow \forall D, f$, WITH $p \geq 1-\delta$: $L_{D,f}(ERM_{\mathcal{H}}(S)) \leq \varepsilon$

PAC LEARNABILITY :

AN HYPOTHESIS CLASS $\mathcal{H}$ IS PAC LEARNABLE, IF $\exists$ FUNCTION $m_{\mathcal{H}} : (0,1)^2 \mapsto \mathbb{N}$ AND A LEARNING ALGORITHM /:

- $\forall \varepsilon, \delta \in (0,1)$ | - $\forall D$ OVER X, AND $\forall$ LABELING FUNCTION $f : X \mapsto \{0,1\}$,

WHEN RUNNING THE ALGORITHM ON $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid EXAMPLES GENERATED BY D AND LABELED BY f

$\rightarrow$ ALGORITHM RETURNS h / WH $p \geq 1-\delta$: $L_{D,f}(h) \leq \varepsilon$

$\rightarrow$ SAMPLE COMPLEXITY: min # INPUT DATA TO ASSURE

SAMPLE COMPLEXITY OF LEARNING $\mathcal{H}$ FOR ε, δ : $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \geq m_{\mathcal{H}}(\varepsilon, \delta)$

- COROLLARY :

LET $\mathcal{H}$ BE A FINITE HYPOTHESIS CLASS:

$\rightarrow \mathcal{H}$ IS PAC LEARNABLE WITH $m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$ / $m_{\mathcal{H}}$ OBTAINED USING $ERM_{\mathcal{H}}$ LEARNING RULE

VC DIMENSION:

- WE SAY THAT H "SHATTERS" C IF $|H_C| = 2^{|C|}$
  - → IF A SET OF FUNCTIONS $\in H_C$ IS ABLE TO PRODUCE EVERY COMBINATION OF $\{\pm 1\}^{|C|}$ OUTPUTS → H SHATTERS C
  - $|H_C| \leq 2^{|C|}$
- $\text{VC dim}(H) = d = \sup\{|C| : H \text{ shatters } C\}$ → IT IS THE HIGHEST CARDINALITY OF A SET OF C H SHATTERS THE SET C
  - → $\text{VC dim} \leq \log_2(|H|)$
  - H FINITE → PAC LEARNABLE
  - H INFINITE → $\nexists$ h PAC LEARNABLE (→ NEEDS OF VC DIM < ∞ FINITE)
- IN ORDER TO SHOW THAT $\text{VC dim}(H) = d$:
  1. SHOW THAT $\exists$ A SET C OF SIZE $|C| = d$ (→ JUST 1 SET IS NEEDED) / C IS SHATTERED BY H → $\text{VC dim}(H) \geq d$
  2. $\forall$ SET C / $|C| = d+1$, C IS NOT SHATTERED BY H → $\text{VC dim}(H) < d+1$

  ① + ② : $\text{VC dim}(H) = d$

THE FUNDAMENTAL THEOREM OF PAC LEARNING: TFAE: → THE FOLLOWING ARE EQUIVALENT

1. H HAS THE UNIFORM CONVERGENCE PROPERTY
2. ANY ERM RULE IS A SUCCESSFUL AGNOSTIC PAC LEARNER FOR H
3. H IS AGNOSTIC PAC LEARNABLE
4. H IS PAC LEARNABLE
5. ANY ERM RULE IS A SUCCESSFUL PAC LEARNER FOR H
6. H HAS A FINITE VC-DIM $= d < \infty$ → $\text{VC dim}(H) = d$ → H IS PAC LEARNABLE

- THEOREM:
  - LET: H BE A FINITE HYPOTHESIS CLASS OF FUNCTIONS $X \longmapsto \{0,1\}$ | LOSS FUNCTION: 0-1 LOSS | $\text{VC dim} = d < \infty$
  - → $\exists$ C_1, C_2 CONSTANTS, SUCH THAT: $C_1 \cdot \frac{d + \log(1/\delta)}{\varepsilon} \leq m_H(\varepsilon, \delta) \leq C_2 \frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$
  - AND $m_H(\varepsilon, \delta)$ IS ACHIEVED BY THE ERM LEARNING RULE
- LEMMA: LET H BE AN HYPOTHESIS CLASS / $\text{VC dim}(H) = d < \infty$ → $\forall C \subset X$ / $|C| = m > d+1$ → $|H_C| \leq \left(\frac{em}{d}\right)^d \approx K \cdot m^d$ (EULER'S NUMBER)

AGNOSTIC PAC LEARNING:

- D OVER $X \times Y$
- 0-1 LOSS: $\ell(h,(x,y)) = 1$ IF $h(x) \neq y$ | 0 IF $h(x) = y$
- SQUARED LOSS: $\ell(h,(x,y)) = (h(x) - y)^2$ → SOME LOSSES
- RISK: $L_D(h) \overset{\text{def}}{=} P_{(x,y)\sim D}[h(x) \neq y] \overset{\text{def}}{=} D(\{(x,y) : h(x) \neq y\})$ (BAYESIC LOSS OF CLASS)
- "APPROXIMATELY CORRECT" NOTION: $L_D(A(S)) \leq \min_{h\in H} L_D(h) + \varepsilon$ → MINIMUM LOSS AMONG ANY FUNCTION

	PAC	Agnostic PAC
Distribution	$\mathcal{D}$ over $\mathcal{X}$	$\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
Truth	$f \in \mathcal{H}$	not in class or doesn't exist
Risk	$L_{\mathcal{D},f}(h) = \mathcal{D}(\{x : h(x) \neq f(x)\})$	$L_{\mathcal{D}}(h) = \mathcal{D}(\{(x,y) : h(x) \neq y\})$
Training set	$(x_1,\dots,x_m) \sim \mathcal{D}^m$ $\forall i,\ y_i = f(x_i)$	$((x_1,y_1),\dots,(x_m,y_m)) \sim \mathcal{D}^m$
Goal	$L_{\mathcal{D},f}(A(S)) \leq \epsilon$	$L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

FORMAL DEF:

AN HYPOTHESIS CLASS H IS AGNOSTIC PAC LEARNABLE TO A SET $Z = X \times Y$ AND A LOSS FUNCTION $\ell : H \times Z \longmapsto \mathbb{R}^+$, IF $\exists$ A FUNCTION $m_H : (0,1)^2 \longmapsto \mathbb{N}$ AND A LEARNING ALGORITHM A WITH THE PROPERTY:

$\forall \varepsilon, \delta \in (0,1)$, $m \geq m_H(\varepsilon,\delta)$, AND A DISTRIBUTION D OVER Z (THEORETICAL TRUE DISTRIBUTION OF Z): $D^m\left(\{S \in Z^m : L_D(A(S)) \leq \min_{h\in H} L_D(h) + \varepsilon\}\right) \geq 1-\delta$

- DATASET $S \sim Z \times \dots \times Z$ (CARTESIAN PRODUCT / EACH $Z \sim D$ → $D \times \dots \times D$)
- LEARNER A ON S
- ALL POSSIBLE m-TUPLES IN Z
- LOSS ON BEST h
- EXPECTED LOSS IS AT MOST ε-LARGER THAN THE EXPECTED LOSS OF THE BEST h
- C.I.

- A TRAINING SET S IS ε-REPRESENTATIVE SAMPLE IF $\forall h \in H,\ |L_S(h) - L_D(h)| \leq \varepsilon$
- LEMMA: ASSUME THAT A TRAINING SET S IS $\frac{\varepsilon}{2}$-REPRESENTATIVE
  - → ANY OUTPUT OF $\text{ERM}_H(S)$: $h_S \in \arg\min_{h\in H} L_S(h)$
  - SATISFIES: $L_D(h_S) \leq \min_{h\in H} L_D(h) + \varepsilon$
  - → IF S IS $\frac{\varepsilon}{2}$-REPRESENTATIVE → $L_D(h_S) \leq \min_{h\in H} L_D(h) + \varepsilon$ (OUTPUT OF $\text{ERM}_H(S)$ → $h_S = \text{ERM}_H(S)$)

- UNIFORM CONVERGENCE: → IF UC → H CAN LEARN WELL FROM S
  - H HAS THE UNIF. CONV. PROPERTY IF $\exists$ FUNCTION $m_H^{UC} : (0,1)^2 \longmapsto \mathbb{N}$ / $\forall \varepsilon, \delta \in (0,1)$ AND $\forall D$: $D^m\left(\{S \in Z^m : S \text{ is } \varepsilon\text{-representative}\}\right) \geq 1-\delta$
  - → IF $\exists$ m_H^{UC} / $D^m\left(\{S \in Z^m : S \text{ is } \varepsilon\text{-representative}\}\right) \geq 1-\delta$ → UC
  - COROLLARY: → IF UC → $\exists$ BOUND (UPPER) FOR $m_H(\varepsilon,\delta)$ IF H HAS UC PROPERTY WITH A FUNCTION m_H^{UC}:
    - → H IS AGNOSTICALLY PAC LEARNABLE WITH $m_H(\varepsilon,\delta) \leq m_H^{UC}\left(\frac{\varepsilon}{2},\delta\right)$
    - IN THAT CASE, ERM_H PARADIGM IS A SUCCESSFUL AGNOSTIC PAC LEARNER FOR H

- **THEOREM** (FINITE CLASS → AGNOSTIC PAC LEARNABLE):

LET H BE FINITE AND $\ell \in [0,1]$

→ H IS AGNOSTICALLY PAC LEARNABLE WITH:

$$m_H(\varepsilon,\delta) \leq \left\lceil \frac{2\log(2|H|/\delta)}{\varepsilon^2} \right\rceil$$

$\lceil \cdot \rceil$: CEILING FUNCTION, e.g. $\lceil 2.4 \rceil = 3$

- **DISCRETIZATION TRICK:** SUPPOSE H PARAMETRIZED BY d NUMBERS, EACH NUMBER REPRESENTED BY b BITS

→ $|H| \leq 2^{d \cdot b}$

→ $m_H(\varepsilon,\delta) \leq \left\lceil \frac{2db + 2\log(2/\delta)}{\varepsilon^2} \right\rceil$

- **LEAST SQUARES:**

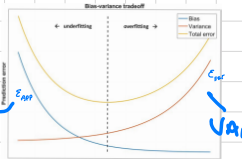
→ PROBLEM: $\min_{\bar{w} \in \mathbb{R}^d} \frac{1}{2} \| X^T \cdot \bar{w} - \bar{Y} \|^2 = \min_{\bar{w} \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \left(\langle \bar{w}, \bar{x}_i \rangle - y_i \right)^2$

→ SOL: $\bar{w} = \underbrace{(XX^T)^{-1}}_{\text{IF INVERTIBLE}} X Y$

$J_x(f \circ g) = J_{f(g(x))}(f) \cdot J_x(g)$, i.e. $\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}$

- **POLYNOMIAL FITTING:**

- DEFINE $\psi : \mathbb{R} \to \mathbb{R}^{n+1}$, $\psi(x) = (1, x_1, x^2, \dots, x^n)$
- DEFINE $\bar{a} = (a_0, a_1, \dots, a_n)$ AND OBSERVE: $p(x) = \sum_{i=0}^{n} a_i x^i = \langle \bar{a}, \psi(x) \rangle$ → HALF SPACE / LIN. REGRESSION
- → TO FIND $\bar{a}$ WE CAN SOLVE LS (LEAST SQUARES) W.R.T. $\left((\psi(x_1), y_1), \dots, (\psi(x_m), y_m) \right)$



BIAS → ε_{APP}, VAR → ε_{EST}

$L_D(h_S) = \varepsilon_{APP} + \varepsilon_{EST}$

- **APPROXIMATION ERROR** (c.) H) → $\varepsilon_{APP} = \min_{h \in H} L_D(h)$

→ IT IS THE RISK WE HAVE DUE TO RESTRICTING TO H

→ IT DOES NOT DEPEND ON S

→ ↑ (COMPLEXITY OF H / VC dim) → ↓ ε_{APP}

↑ m → ↓ ε_{APP} (e.g. CHANGING POLYNOMIAL DEGREE)

- **ESTIMATION ERROR** (c.) S) → $\varepsilon_{EST} = L_D(h_S) - \varepsilon_{APP}$

→ IT TELLS HOW GOOD AM I IN EXPLOITING H ON MY PROBLEM

→ IT IS THE RESULT OF L_S BEING ONLY AN ESTIMATE OF L_D

→ ↑ SIZE OF S → ↓ ε_{EST}

→ ↑ COMPLEXITY OF H → ↑ ε_{EST} ∵ ↑ COMPLEXITY → ↓ ε_{APP} → ↑ ε_{EST}

↓ IT CANNOT FIT WELL OTHER DATA NOT IN TRAINING SET

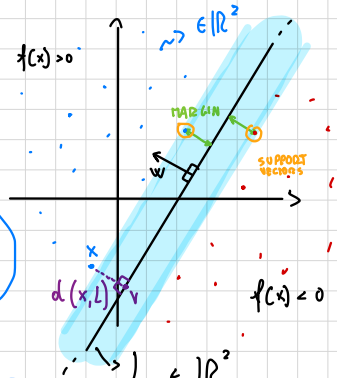
VALIDATION: HOEFFDING'S INEQUALITY, IF $\ell \in [0,1]$: $|L_V(h) - L_D(h)| \leq \sqrt{\frac{\log(2/\delta)}{2 \cdot m_V}}$

- LET V BE A NEW VAL. SET, CHOOSE $h^* \in ERM_H(V)$ → $L_D(h^*) \leq \min_{h \in H} L_D(h) + \sqrt{\frac{2\log(2|H|/\delta)}{|V|}}$

SVM: HALF-SPACE CLASSIFIER: $x \mapsto \text{sign}(\langle \bar{w}, \bar{x} \rangle + b) = \pm 1$, $\bar{w} \in \mathbb{R}^d$

$L = \{ v : \langle w, v \rangle + b = 0 \}$ (SET OF POINTS OF HYPERPLANE), $d(\bar{x}, L) = \min \{ \| \bar{x} - \bar{v} \| : \bar{v} \in L \}$

- MARGIN: $\min_i |\langle \bar{w}, \bar{x}_i \rangle + b|$ → CLOSEST DISTANCE FROM L TO $x_i, \forall i$ → $\text{MARGIN} = \frac{1}{\| \bar{w}_0 \|}$ (2 CLASSES SEPARATION)



HARD SVM:

FIND $\bar{w}, b$ TO MAXIMIZE MIN DISTANCE

$$\text{arg max}_{(\bar{w},b) : \|\bar{w}\| = 1} \left(\min_{i \in [m]} |\langle \bar{w}, \bar{x}_i \rangle + b| \right) \Big/ \forall i, \; y_i(\langle \bar{w}, x_i \rangle + b) > 0$$

(CLOSEST DISTANCE L TO x_i; ≥ 0)

- OR EQUIVALENTLY: $\text{arg max}_{(\bar{w},b) : \|\bar{w}\| = 1} \left(\min_{i \in [m]} y_i(\langle \bar{w}, \bar{x}_i \rangle + b) \right)$

BEST PLANE: $\langle \bar{w}_0, \bar{x} \rangle + b_0$

- OR EQUIVALENTLY: $(\bar{w}_0, b_0) = \text{arg min}_{\bar{w}, b} \left(\|\bar{w}\|^2 \right) \Big/ \forall i, \; y_i(\langle \bar{w}, x_i \rangle + b) \geq 1$

POINT (x_p, y_p, z_p), PLANE $\bar{n}$

$$d(P, \bar{n}) = \frac{|ax_p + by_p + cz_p + d|}{\sqrt{a^2 + b^2 + c^2}}$$

- MARGIN OF DISTRIBUTION:

D IS SEPARABLE WITH A (γ, S) - MARGIN, IF $\exists (\bar{w}^*, b^*) / \|\bar{w}^*\| = 1$ AND:

$$D\left(\{ (\bar{x}, y) : \|\bar{x}\| \leq S \wedge y(\langle \bar{w}^*, x \rangle + b^*) \geq 1 \} \right) = 1$$

(POINTS IN A FINITE SPACE; 2 CLASS SEPARATION)

→ THEOREM:

IF D SEPARABLE WITH (γ, S)-MARGIN → $m(\varepsilon, \delta) \leq \frac{S}{\varepsilon^2} \left(2\left(\frac{S}{\gamma}\right)^2 + \log\left(\frac{2}{\delta}\right) \right)$ (SAMPLE COMPLEXITY ↑)

↳ UNLIKE VC BOUNDS, m c.) $\frac{S}{\delta}$, NOT ON d

- **SOFT SVM**: $\underset{\bar{w}, b, \xi}{\operatorname{argmin}} \left(\lambda \cdot \|\bar{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right) \Big/ \forall i,\ y_i \left(\langle \bar{w}, x_i \rangle + b \right) \geq 1 - \xi_i,\ \xi_i \geq 0$ ($\frac{1}{m} \sum \xi_i = L_S^{\text{HINGE}}$)

↓ COST → ↑ ACC TEST
↑ # FEATURES ↛ ↑ ACC TEST

COST: $= \underset{\bar{w}, b}{\operatorname{argmin}} \left(\lambda \cdot \|\bar{w}\|^2 + L_S^{\text{HINGE}}(\bar{w}, b) \right)$ ↳ S: WHOLE DATASET

HINGE LOSS: $\ell^{\text{HINGE}}\left((\bar{w}, b), (\bar{x}, y)\right) = \max\{0, 1 - y(\langle \bar{w}, \bar{x} \rangle + b)\}$ ($1 - y(\langle \bar{w}, \bar{x} \rangle + b) = \xi$)

$= \max\{0, 1 - y \cdot \hat{y}\}$ ($\hat{y} \in \{-1, 1\}$)

$C \mapsto 0$: POINTS IN MARGIN

$C \mapsto \infty$: $\mapsto$ HARD SVM, $\xi_i \mapsto 0 \rightarrow 1 - \xi_i \mapsto 1$

- **ON-AVERAGE-REPLACE-ONE-STABLE** DEFINITION: $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m) \Big/ S = (z_1, \ldots, z_m)$ (REPLACED SAMPLE, POSITION i)

LET $\varepsilon : \mathbb{N} \mapsto \mathbb{R}$ BE A MONOTONICALLY DECREASING FUNCTION.

- ALGORITHM A IS ON-AVERAGE-REPLACE-ONE-STABLE WITH RATE $\varepsilon(m)$ IF $\forall$ DISTRIBUTION D:

$$\underset{\substack{(S, z') \sim D^{m+1} \\ i \sim U(m)}}{\mathbb{E}} \left[\ell\left(A\left(S^{(i)}\right), z_i\right) - \ell\left(A(S), z_i\right) \right] \leq \varepsilon(m)$$

- FRITZ-JOHNS; $\bar{w}_0 = \sum_i \alpha_i \bar{x}_i$
  ↳ OPTIMAL PLAIN WEIGHTS ARE A LIN. COMB. OF SUPPORT VECTORS

- **THEOREM**: IF A IS ON-AVERAGE-REPLACE-ONE-STABLE WITH RATE $\varepsilon(m)$: $\rightarrow \underset{S \sim D^m}{\mathbb{E}} \left[L_D\left(A(S)\right) - L_S\left(A(S)\right) \right] \leq \varepsilon(m)$ (LOSS BY TRUE DISTRIBUTION D OF DATA; LOSS BY TRAINING DATA)

- **THEOREM**.

TIKHONOV REGULARIZATION AS STABILIZER: (f PROPERTY.) $|f(x_1) - f(x_2)| \leq \rho \cdot |x_1 - x_2|$ (LIPSCHITZ)

ASSUME THAT THE LOSS FUNCTION IS CONVEX AND ρ-LIPSCHITZ.

$\rightarrow$ RLM RULE, WITH REGULARIZER $\lambda \|w\|^2$, IS ON-AVERAGE-REPLACE-ONE-STABLE WITH RATE $\varepsilon(m) = \frac{2\rho^2}{\lambda m}$, AND:

$$\underset{S \sim D^m}{\mathbb{E}} \left[L_D\left(A(S)\right) - L_S\left(A(S)\right) \right] \leq \frac{2\rho^2}{\lambda m} \quad (\text{RATE } \varepsilon(m))$$

- **FITTING-STABILITY TRADEOFF**: $\underset{S}{\mathbb{E}} \left[L_D(A(S)) \right] \leq L_D(\bar{w}^*) + \lambda \|\bar{w}^*\|^2 + \frac{2\rho^2}{\lambda m}$

- **THE REGULARIZATION PATH**:

  - RLM RULE: $\bar{w}(\lambda) = \underset{\bar{w}}{\operatorname{argmin}} \left(L_S(\bar{w}) + \lambda \|\bar{w}\|^2 \right)$
  - $\bar{w}(\lambda = 0)$: MINIMIZE JUST $L_S(\bar{w})$
  - $\bar{w}(\lambda = \infty)$: MINIMIZING $\|\bar{w}\|^2$ HAS A HIGHER IMPACT: $L_S(\bar{w}) \overset{\lambda \|\bar{w}\| \gg 0}{\longmapsto} 0$

- **SAMPLE COMPLEXITY OF SOFT-SVM**.

  - IF WE SET $\lambda = \frac{1}{B} \sqrt{\frac{2\rho^2}{m}}$ (CONST.) $\forall B > 0$, AND SINCE ℓ^{HINGE} IS AN UPPER BOUND FOR 0-1 LOSS:

$$\rightarrow \underset{S \sim D^m}{\mathbb{E}} \left[L_D^{0-1}\left(A(S)\right) \right] \leq \min_{\bar{w} : \|w\| \leq B} \left\{ L_D^{\text{HINGE}}(\bar{w}) \right\} + \sqrt{\frac{8\rho^2 B^2}{m}}$$

- **MARGIN / NORM VS DIMENSIONALITY**:

  - VC DIMENSION OF LEARNING HALFSPACES (.) DIMENSION OF
  - $\rightarrow \uparrow d \rightarrow \uparrow$ SAMPLE COMPLEXITY
  - FOR SVMs: SAMPLE COMPLEXITY (.) $\left(\frac{\rho}{\gamma}\right)^2 = \left(\rho^2 B^2\right)$

- **KERNELS**:

- POLYNOMIAL MAPPINGS: $p^{(k)}(x) = \sum_{j=0}^{k} w_j x^j = \langle \bar{w}, \psi(x) \rangle \Big/ \psi(x) = (1, x, x^2, \ldots, x^k)$ (DEGREE k)

- KERNEL TRICK: $K(\bar{x}_1, \bar{x}_2) = \langle \psi(\bar{x}_1), \psi(\bar{x}_2) \rangle$

- **THE REPRESENTER THEOREM**:

CONSIDER ANY LEARNING RULE, $f : \mathbb{R}^m \mapsto \mathbb{R}$, OF THE FORM:

$$\bar{w}^* = \underset{\bar{w}}{\operatorname{argmin}} \left\{ f\left(\langle \bar{w}, \psi(\bar{x}_1) \rangle, \ldots, \langle \bar{w}, \psi(\bar{x}_m) \rangle \right) + \lambda \|\bar{w}\|^2 \right\}$$

$\rightarrow \exists \bar{\alpha} \in \mathbb{R}^m \Big/ \bar{w}^* = \sum_{i=1}^{m} \alpha_i \cdot \psi(\bar{x}_i)$ ($\bar{w}^*$ IS A LINEAR COMBINATION OF $\psi(\bar{x}_i)$)

$\rightarrow$

- $\|\bar{w}\|^2 = \bar{\alpha}^T G \bar{\alpha}$ $\quad G_{ij} = K(\bar{x}_i, \bar{x}_j)$
- $\underset{\bar{\alpha} \in \mathbb{R}^m}{\operatorname{argmin}} \left\{ f(G \cdot \bar{\alpha}) + \lambda \bar{\alpha}^T G \bar{\alpha} \right\}$
- $\langle \bar{w}, \psi(\bar{x}) \rangle = \sum_j \alpha_j K(\bar{x}_j, \bar{x})$ (NEW INSTANCE)

$G = \begin{pmatrix} \langle \psi(\bar{x}_1), \psi(\bar{x}_1) \rangle & \langle \psi(\bar{x}_1), \psi(\bar{x}_2) \rangle \\ \langle \psi(\bar{x}_2), \psi(\bar{x}_1) \rangle & \langle \psi(\bar{x}_2), \psi(\bar{x}_2) \rangle \end{pmatrix}$ (POSITIVE SEMI-DEF, NOT ALWAYS $(\cdot)^{-1}$; $\|\psi(\bar{x}_i)\|^2 \rightarrow$ ANALOG FOR $i, j = 2$)

- **POLYNOMIAL KERNELS**: DEGREE OF POLYNOMIAL USED TO TRANSFORM $\bar{x}_1, \bar{x}_2$

$$K(\bar{x}_1, \bar{x}_2) = \left(1 + \langle \bar{x}_1, \bar{x}_2 \rangle\right)^K$$

- **GAUSSIAN KERNELS (RBF)**:

$$K(\bar{x}_1, \bar{x}_2) = e^{-\frac{\|\bar{x}_1 - \bar{x}_2\|}{2\sigma}}$$

→ CAN HAVE INFINITE FATURE MAPPING

→ IT CAN LEARN ANY POLYNOMIAL

- **LEMMA (MERCER'S CONDITION)**: A SYMMETRIC FUNCTION $K: X \times X \mapsto \mathbb{R}$ IMPLEMENTS AN INNER PRODUCT IN SOME HILBERT SPACE $\Leftrightarrow \forall \bar{x}_1, \dots, \bar{x}_m : G_{i,j} = K(\bar{x}_i, \bar{x}_j)$ IS A POSITIVE, SEMI-DEFINITE MATRIX

• CURSE OF DIMENSIONALITY

• DIMENSIONALITY REDUCTION:

λ OF A: $\det(A - \lambda I)$

$\bar{v}$ OF A: $(A - \lambda I) \cdot \bar{v} = 0$ ($\bar{v}$: CORRESPONDING EIGENVALUE OF $\bar{v}$)

$$d \left\{ \underbrace{\begin{bmatrix} \\ \bar{x} \\ \\ \end{bmatrix}}_{d \times 1} \longmapsto n \left\{ \underbrace{\begin{bmatrix} \\ W \\ \\ \end{bmatrix}}_{n \times d,\ n < d} \cdot \begin{bmatrix} \\ \bar{x} \\ \\ \end{bmatrix} = n \left\{ \underbrace{\begin{bmatrix} \\ \bar{y} \\ \\ \end{bmatrix}}_{n \times 1} \right. \right. \right.$$

- **PCA**: $\sim O(nd^2 + d^3)$ WE WANT TO BE ABLE TO APPROXIMATELY RECOVER $\bar{x}$ FROM $\bar{y} = W\bar{x}$

↳ IN PCA: $\tilde{x} = U\bar{y} = UW\bar{x}$ (U: $d \times m$, W: $m \times d$) — LINEAR RECOVERY

↳ MEASURES "APPROXIMATE RECOVERY", SOLVE:

$$\underset{W \in \mathbb{R}^{m,d},\ U \in \mathbb{R}^{d,m}}{\arg\min} \left\{ \sum_{i=1}^{n} \left\| \underbrace{\bar{x}_i}_{\text{ORIGINAL FEATURES}} - \underbrace{UW\bar{x}_i}_{\text{RECOVERED FEATURES}} \right\|_2^2 \right\} \quad \left\{ \sim n : \#\text{ SAMPLES} \right.$$

- IT IS A LOSSY PROCESS: $\tilde{x} \neq \bar{x}$
- $U \cdot W \approx I_d$
- $\bar{x}_{d \times 1} \xrightarrow{W} W\bar{x} = \bar{y}_{m \times 1} \xrightarrow{U} U \cdot \bar{y} = \tilde{x}_{d \times 1}$ (W: $m \times d$, $m < d$; U: $d \times m$)

ENCODING — DECODING

- **THEOREM**:

LET $A = \sum_{i=1}^{n} \underbrace{\bar{x}_i \bar{x}_i^T}_{d \times d}$, LET $\bar{u}_1, \dots, \bar{u}_m$ BE THE m LEADING EIGENVECTORS OF A:

→ SOLUTION OF PCA: COLUMNS OF U SET TO $\bar{u}_1, \dots, \bar{u}_m$ ($U = [\bar{u}_1 \ \bar{u}_2 \ \cdots \ \bar{u}_m]$); $W = U^T$

1. $\bar{\mu}_x$
2. $\frac{1}{n}[\ ][\ ]$
3. $\lambda, \bar{v}$

- **N.B.**: $\frac{1}{n-1}(\bar{x} - \mu)^T(\bar{x} - \mu) = \Sigma$ = COVARIANCE MATRIX

- **PCA AND SVD**:

$XX^T = [\|]$ $(X^TX)^T = [\|]$

$$SVD\underset{n \times d}{(X)} = \underset{n \times n}{U} \cdot \underset{n \times d}{S} \cdot \underset{d \times d}{V^T}$$

U: COLUMNS ARE EIGENVECTORS OF $X \cdot X^T$

S: DIAGONAL MATRIX, $s_{ii} \geq 0$, $= \sqrt{\lambda_i}$ OF X^TX

V: ROWS ARE EIGENVECTORS OF X^TX

- IN PCA: $\underset{d \times d}{\Sigma} = \frac{1}{n-1} X^T X$, ASSUMING DATA IN X ARE ZERO-CENTERED

- Σ IS SYMMETRIC → Σ CAN BE DIAGONALIZED. $\Sigma = \underset{d \times d}{V} \cdot \underset{d \times d}{L} \cdot \underset{d \times d}{V^T}$

, WHERE:

$$V = \left[\begin{matrix} | & & | & & | \\ \bar{v}_1 & \cdots & \bar{v}_i & \cdots & \bar{v}_d \\ | & & | & & | \end{matrix} \right] \ (\bar{v}_i: \text{EIGENVECTOR}), \quad L = \begin{bmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_i & \\ & & & \ddots \\ & & & & \lambda_d \end{bmatrix} \ (\lambda_i: \text{EIGENVALUE}) \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$$

DIAGONAL MATRIX

- $\bar{v}_i$: PRINCIPAL AXES / PROJECTION OF DATA ON $\bar{v}_i$ → PC_i

↳ $PC_j = [X] \cdot \begin{bmatrix} | \\ \bar{v}_j \\ | \end{bmatrix}$ ($n \times 1$)

↳ $x_{i,PCA} = [x_i] \begin{bmatrix} V \end{bmatrix}$ ($1 \times d$) ~ COORDINATES OF i-TH POINT ON NEW PC SPACE

- PROPORTION OF VARIANCE EXPLAINED: $PVE = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} = \frac{\lambda_i}{\text{trace}(\Sigma)}$

• RANDOM PROJECTIONS: $\sim O(nd_k)$

RANDOM PROJECTION — DATA PROJECTED IN k-DIMENSIONAL SPACE (k: TARGET DIMENSIONALITY)

- $\underset{d \times 1}{\bar{x}} \longmapsto \underset{k \times d}{W} \underset{d \times 1}{\bar{x}}$ / W IS A RANDOM MATRIX: $W_{i,j} \sim N\left(0, \frac{1}{n}\right)$

- LET $\bar{w}_i$ BE THE i^{TH} ROW OF W:

IN FACT:

→ $E\left[\|W\bar{x}\|^2\right] = \|\bar{x}\|^2$ — PROPERTY ② SATISFIED → $\|W\bar{x}\|^2 \sim \chi^2_n$

→ $P\left[\left|\frac{\|W\bar{x}\|^2}{\|\bar{x}\|^2} - 1\right| > \varepsilon\right] \leq 2e^{-\varepsilon^2 \frac{n}{6}}$ — HOEFFDING'S INEQUALITY

① DO NOT DISTORT DISTANCES:

$\|\bar{x}_i - \bar{x}_j\| \simeq \|\bar{y}_i - \bar{y}_j\|, \ \forall i,j$

→ $\frac{\|W\bar{x}_i - W\bar{x}_j\|}{\|\bar{x}_i - \bar{x}_j\|} \simeq 1, \ \forall i,j$

② $\forall \bar{x} \in Q$, $Q = \{\bar{x}_i - \bar{x}_j : i,j \in [m]\}$, WE HAVE $\frac{\|W\bar{x}\|}{\|\bar{x}\|} \simeq 1$

- **LEMMA (JOHNSON-LINDENSTRAUSS LEMMA)**: LET Q BE A FINITE SET OF VECTORS IN $\mathbb{R}^d$

↳ SETTING $\varepsilon = \sqrt{\frac{6 \cdot \log(2|Q|/\delta)}{n}} \leq 3$

→ PROPERTY ② SATISFIED, WITH C.I $1 - \delta$

COMPRESSED SENSING:

- PRIOR ASSUMPTION: $U^{-1} = U^T$, $\langle \bar{u}_i, \bar{u}_j \rangle = 0$, $\|\bar{u}_i\|_2^2 = 1$
  - $\bar{x} \approx U\bar{a}$ / U is ORTHONORMAL | $\|\bar{a}\|_0 := |\{i : a_i \neq 0\}| \leq S$, $S \ll d$
- ANY SPARSE SIGNAL CAN BE FULLY RECONSTRUCTED IF IT WAS COMPRESSED BY $\bar{x} \longmapsto W\bar{x}$ / W SATISFIES RESTRICTED ISOPERIMETRIC PROPERTY (RIP)
  - RESTRICTED ISOPERIMETRIC PROPERTY:

A MATRIX $W \in \mathbb{R}^{n,d}$ IS (ε, S)-RIP IF $\forall \bar{x} \neq 0 \,/\, \|\bar{x}\|_0 \leq S$: $\rightarrow$ $\left| \frac{\|W\bar{x}\|_2^2}{\|\bar{x}\|_2^2} - 1 \right| \leq \varepsilon \rightarrow 1-\varepsilon \leq \frac{\|W\bar{x}\|_2^2}{\|\bar{x}\|_2^2} \leq 1+\varepsilon$

PCA VS RANDOM PROJECTIONS:

- IF DATA IS $\bar{e}_1, \dots, \bar{e}_d$:
  - RANDOM PROJECTIONS $\rightarrow$ PERFECT
  - PCA $\rightarrow$ FAIL
- IF $d \gg 0$, DATA EXACTLY IN n-DIM SUBSPACE: (I ONLY NEED FIRST n PC)
  - RANDOM PROJECTIONS $\rightarrow$ MIGHT FAIL
  - PCA $\rightarrow$ PERFECT
    - $\hookrightarrow$ EG. DATA IN $\mathbb{R}^2$ AND I TAKE $n=2$ PC

$\bar{x} \longmapsto W\bar{x}$

- PCA: OPTIMAL IF LINEAR RECONSTRUCTION AND ERROR IS SQUARED DISTANCE (BUT WITH NON-LINEAR RECONSTRUCTION)
- RANDOM PROJECTIONS: PRESERVE DISTANCES, EXACT RECONSTRUCTION FOR SPARSE VECTORS

DECISION TREES:

- PREDICTION FUNCTION: $y(\bar{x}) = \sum_{w \in W} y^w(\bar{x}) \cdot \mathbb{1}_{[\bar{x} \in R_w]}$
- TRAINING LOSS:

GIVEN $y(\bar{x})$ AS ABOVE, TRAINING SET $\tilde{T} = \{(\bar{x}_i, y_i)\}_{i=1}^m$:

$$\ell_{\tilde{T}}(y) = \sum_{w \in W} \frac{1}{m} \sum_{i=1}^{m_w} \mathbb{1}_{[\bar{x}_i \in R_w]} \cdot \text{Loss}\left(y_i, y^w(\bar{x}_i)\right)$$

($\forall$ REGION; $\forall$ POINT in $\tilde{T}$; $\mathbb{1}$: 1 IF $\bar{x}_i \in R_w$, 0 OTHERWISE; CONTRIBUTION OF REGIONAL y^w TO OVERALL TRAINING LOSS)

- REGIONAL PREDICTION FUNCTION

GIVEN $p_z^w = \frac{1}{m_w} \sum_{\{(\bar{x},y) \in \tilde{T} : \bar{x} \in R_w\}} \mathbb{1}_{[y=z]}$

CLASS $y^w(\bar{x}) = \underset{z \in \{0,\dots,c-1\}}{\text{argmax}} \{p_z^w\}$

REG $y^w(\bar{x}) = \bar{y}_{R_w} := \frac{1}{m_w} \sum_{\{(\bar{x},y) \in \tilde{T} : \bar{x} \in R_w\}} y$

- OPTIMAL SPLITTING RULES:

REG $\frac{1}{m} \sum_{(\bar{x},y) \in \tilde{T} : x_j \leq \xi} (y - \bar{y}_T)^2 + \frac{1}{m} \sum_{(\bar{x},y) \in \tilde{T} : x_j > \xi} (y - \bar{y}_F)^2$

CLASS $\frac{1}{m} \sum_{(\bar{x},y) \in \tilde{T}} \mathbb{1}_{[y \neq y_T^*]} + \frac{1}{m} \sum_{(\bar{x},y) \in \tilde{T}_F} \mathbb{1}_{[y \neq y_F^*]}$ / y^*: MOST RELEVANT CLASS (MAJORITY VOTE)

- IMPURITIES:
  - ENTROPY: $-\sum_{z=0}^{c-1} p_z \cdot \log_2(p_z)$
  - GINI: $\frac{1}{2}\left(1 - \sum_{z=0}^{c-1} p_z^2\right)$
  - OTHER IMPURITY

... BUT: HIGH CORRELATED PREDICTION $\rightarrow \sigma^2$

BOOTSTRAPPING:

WE CAN SUBSTITUTE THEM WITH BOOTSTRAPPED ONES: WE CAN OBTAIN $\tilde{T}_1^*, \dots, \tilde{T}_B^*$ BY RESAMPLING FROM A SINGLE FIXED TRAINING SET $\tilde{T}$, AND USE THEM TO TRAIN B SEPARATE MODELS $\rightarrow$

WE OBTAIN A BAGGED ESTIMATOR OF THE FORM: (= BOOTSTRAPPED)

$$y_{bag}(\bar{x}) = \frac{1}{B} \sum_{b=1}^{B} y_{\tilde{T}_b^*}(\bar{x})$$

RANDOM FOREST $\rightarrow$ ONLY A SUBSET OF FEATURES CONSIDERED DURING TREE CONSTRUCTION

BOOSTING: $\rightarrow$ IF WEAK LEARNER $\rightarrow$ ↑ ACCURACY

2. "BOOST" y_0 TO A NEW LEARNER: $y_1 := y_0 + h_1$

$$h_1 = \underset{h \in H}{\text{argmin}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \text{Loss}\left(y_i, y_0(\bar{x}_i) + \gamma h(\bar{x}_i)\right) \right\}$$

GRADIENT BOOSTING:

$z = y_{b-1}(\bar{x})$

$\gamma = -\frac{\partial \text{Loss}(y_i, z)}{\partial z} = \frac{\partial (y_i - z)^2}{\partial z} = 2(y_i - y_{b-1}(\bar{x}_i)) = 2e_i^{(b)}$

ADA BOOST:

$$y_B(\bar{x}) = y_0(\bar{x}) + \sum_{b=1}^{B} h_b(\bar{x})$$

$h_b(\bar{x}) = \alpha_b c_b(\bar{x})$, $\alpha_b \in \mathbb{R}_+$, $c_b(\bar{x}) \in \{-1, 1\}$

- $\forall$ BOOSTING OPERATION:

$$(\alpha_b, c_b) = \underset{\alpha \geq 0,\, c \in C}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{Loss}\left(y_i, y_{b-1}(\bar{x}_i) + \alpha c(\bar{x}_i)\right) \right\}$$