# FREQUENTIST STATISTICS

- ESTIMATORS, MONTE-CARLO BOOTSTRAPPING FOR PARAMETERS, SIMILAR TO SAMPLING FROM BAYESIAN POSTERIOR WHEN PRIORS NOT VERY STRONG
- VARIANCE OF MLE, FISHER INFO MATRIX CURVATURE OF LIKELIHOOD SURFACE

**SCORE:** GRADIENT OF LL AT $\hat{\theta}$ $s(\hat{\theta}) = \nabla \log p(D|\theta) \, d\hat{\theta}$

**OBSERVED INFORMATION MATRIX:** $J(\hat{\theta}(D)) = -\nabla s(\hat{\theta}) = -\nabla^2 \log p(\theta)$
- HESSIAN OF NLL
- GRADIENT OF NEGATIVE SCORE FCN

- **FISHER INFO MATRIX** $I_N(\hat{\theta}|\theta^*) = E_\theta\left[J(\hat{\theta}|D)\right]$, $I_N(\hat{\theta}) = N \cdot I_1(\hat{\theta})$
- SAMPLING DISTRIBUTION OF MLE IS ASYMPTOTICALLY NORMAL
- **STD ERROR** $\hat{s}_e = I_N(\hat{\theta})^{-1/2}$

## FREQUENTIST DECISION THEORY

NO PRIOR, POSTERIOR, OR XPECTED LOSS → NO AUTOMATIC WAY FOR OPTIMAL ESTIMATOR

**RISK / XPECTED LOSS** $R(\theta^*, \delta) = E_{p(D|\theta)}\left[L(\theta^*, \delta(\tilde{D}))\right] = \int L(\theta^*, \delta(\tilde{D})) \, p(\tilde{D}|\theta^*) \, d\tilde{D}$

EXPECTATION WRT SAMPLING DISTRIBUTION
- FREQ. RISK AVERAGES OVER $D$ AND CONDITIONS ON $\theta^*$
- BAYESIAN XPECTED LOSS AVGS OVER $\theta$ AND CONDITIONS ON $D$

FREQUENTIST MAKES NO SENSE → $\theta^*$ IS UNKNOWN

## BAYES RISK

CONVERTS $R(\theta^*, \delta)$ INTO $R(\delta)$ → PRIOR ON $\theta^*$ $R_B = E_{p(\theta^*)}\left[R(\theta^*, \delta)\right] = \int R(\theta^*, \delta) \, p(\theta^*) \, d\theta^* \rightarrow \delta_B = \text{ARGMIN } R_B(\delta)$

ALSO INTEGRATED / PREPOSTERIOR RISK

**THEOREMS:**
- A BAYES ESTIMATOR CAN BE OBTAINED BY MINIMIZING POSTERIOR XPECTED LOSS FOR EACH $X$
- EVERY ADMISSABLE DECISION RULE IS A BAYES DECISION RULE WRT SOME, EV. IMPROPER, PRIOR.
- BEST WAY FOR MINIMIZING FREQUENTIST RISK IS TO BE BAYESIAN
- PICKING OPTIMAL ACTION CASE-BY-CASE (BAYESIAN) IS OPTIMAL ON AVERAGE (FREQUENTIST)

## MINIMAX RISK

$R_{MAX}(\delta) = \max_{\theta^*} R(\theta^*, \delta) \longrightarrow \delta_{MM} = \text{ARGMIN } R_{MAX}(\delta)$

- VERY PESSIMISTIC
- NOT USED XCEPT IN GAME-THEORETIC SCENARIOS B/C NATURE IS NOT ADVERSARIAL

**ADMISSIBLE ESTIMATORS:** IF $R(\theta, \delta_1) \leq R(\theta, \delta_2) \; \forall \theta \rightarrow \delta_1$ DOMINATES $\delta_2$, AN ESTIMATOR IS ADMISSIBLE IF NOT STRICTLY DOMINATED BY ANY OTHER

- SAMPLE MEDIAN IS USUALLY BETTER THAN SAMPLE MEAN
- **STEIN'S PARADOX** WHEN 3 OR MORE PARAMS ARE ESTIMATED SIMULTANEOUSLY, COMBINED ESTIMATORS ARE ON AVG MORE ACCURATE THAN SEPARATE ON AVG
- CONSTANTS CAN BE ~~IS~~ ADMISSIBLE ESTIMATORS, NOT ENOUGH THEN

**CONSISTENT ESTIMATORS** $\hat{\theta}(D) \rightarrow \theta^*, |D| \rightarrow \infty$

**UNBIASED ESTIMATORS:** $\text{BIAS}(\hat{\theta}(\cdot)) = E_{p(D|\theta_*)}\left[\hat{\theta}(D) - \theta_*\right] = 0$ SAMPLING DISTRIBUTION IS CENTERED ON $\theta_*$

$\theta_*$ TRUE PARAM

**MVUE** $\text{var}[\hat{\theta}] \geq \frac{1}{n I(\theta)}$ **CR BOUND**

**BIAS - VARIANCE TRADEOFF:** $MSE = \text{VARIANCE} + \text{BIAS}^2$, MIGHT MAKE SENSE TO USE BIASED EST. IF REDUCES VARIANCE

- IF 0-1 LOSS (CLASSIFICATION) BIAS AND VARIANCE COMBINE MULTIPLICATIVELY

IE, RIDGE REGRESSION, $\lambda > 0$ RESULTS IN BIAS

## EMPIRICAL RISK MINIMIZATION

RISK RELIES ON TRUE DATA DISTRIBUTION

FREQUENTIST RISK CAN BE COMPUTED WHEN TASK IS ESTIMATING OBSERVABLE QUANTITIES AND NOT HIDDEN VARS/PARAMS.

$L(\theta, \delta(D)) \longrightarrow L(y, \delta(x))$ $R(p_x, \delta) = E_{(x,y) \sim p_x}\left[L(y, \delta(x))\right] = \sum_x \sum_y L(y, \delta(x)) p_x(x,y)$

- $p_x$ IS NATURE, EMPIRICAL DISTRIBUTION

EMPIRICAL RISK: $R_{Emp}(D, \delta) = R(p_{Emp}, \delta) = \frac{1}{N} \sum^{N} L(y_i, \delta(x_i))$

- 0-1 LOSS → MISCLASSIFICATION RATE
- $L_2$ LOSS → MSE

$\delta_{ERM}(D) = \text{ARGMIN } R_{Emp}(D, \delta)$

- IN UNSUPERVISED LEARNING $L(y, \delta(x)) \to L(x, \delta(x))$ RECONSTRUCTION ERROR $\delta(x) = \text{DECODE}(\text{ENCODE}(x))$

- IF PRIOR IS 'NATURE' IS EQUAL TO EMPIRICAL → OVERFITTING

COMPLEXITY PENALTY: $R(D, \delta) = R_{Emp}(D, \delta) + \lambda C(\delta)$ ↳ FCN OF COMPLEXITY

- PICK $C$ AS DOF OR VC DIMENSION
- PICK $\lambda$ → $\hat{\lambda} = \text{ARGMIN } \hat{R}(\hat{\delta}_\lambda)$
  - → VIA CROSS-VALIDATION (AVG RISK OVER FOLDS AND TRY DIFFERENT $\lambda$ VALUES)
  - → VIA THEORETICAL UPPER BOUND

- VIA CV NOT FEASIBLE WITH > 2 PARAMS, EMPIRICAL BAYES W/ GRADIENT BASED OPTIMIZERS, NOT FEASIBLE FOR UNSUPERVISED LEARNING
- ONE-STANDARD ERROR RULES: PICK SIMPLEST MODEL W/ RISK NO MORE THAN > 1STD ERR ABOVE RISK OF BEST MODEL

- UPPER BOUND VIA STATISTICAL LEARNING THEORY (SLT)

USING CV IS INEFFICIENT. LET'S BOUND THE RISK FOR ANY DISTRIBUTION AND HYPOTHESIS SPACE WITH $R_{Emp}$, SAMPLE SIZE, AND $|H|$

- $|H|$ IS FINITE $P\left(\max |R_{Emp}(D, h) - R(p_x, h)| > \epsilon\right) \leq 2|H| e^{-2N\epsilon^2}$
- $|H|$ IS INFINITE (REAL VALUES) USE VC DIMENSION, NOT ALWAYS EASY TO COMPUTE, BOUNDS ARE LOOSE

- COLT COMPUTATIONAL LEARNING THEORY, TAKES COMPUTATIONAL COMPLEXITY OF LEARNER INTO ACCOUNT
- A $H$ IS PAC LEARNABLE IF THERE'S A P-TIME ALGORITHM IDENTIFYING A FUNCTION PAC (PROBLY APPROXIMATELY CORRECT)

## SURROGATE LOSS FUNCTIONS

FOR OPTIMIZING 0-1 LOSS, OR OTHER NONSMOOTH METRICS (AUC, F1). USE MLE. CONSTRUCTS A DECISION FUNCTION AND USES LOSS ON THAT

LOG-LOSS: $L_{NLL}(y, \eta) = -\log p(y|x, w) = \lg(1 + e^{-y\eta})$, $\eta$ IS LOG ODDS RATIO, DECISION FCN | MINIMIZING LOGLOSS = MAXIMIZING LIKELIHOOD

HINGE LOSS: $L_{HINGE}(y, \eta) = \max(0, 1 - y\eta)$

# FREQUENTIST PATHOLOGIES

**CONFIDENCE INTERVALS** CONDITION ON UNKNOWN $\theta$ AND AVG ON FUTURE DATA $D$

**P-VALUES** ONLY USEFUL TO REJECT THE NULL HYPOTHESIS → NEVER GATHER EVIDENCE IN FAVOR OF IT, SENSITIVE TO STOPPING RULE LEAD TO NON EARLY TERMINATION, DIFFERENT P-VALUES FOR SAME SITUATION ON BINOMIAL AND NEG-BINOMIAL!

**LIKELIHOOD PRINCIPLE IS VIOLATED**: INFERENCE SHOULD BE BASED ON LIKELIHOOD OF OBSERVED DATA, AND ON EVENTS WHICH ACTUALLY HAPPENED, NOT WHICH MIGHT HAVE HAPPENED