Stat 991 Lecture 3 - Prediction regions & Conformal inference

(given $Z_1, Z_2, Z_3, \ldots, Z_n \in \mathcal{Z}$, predict Z_{n+1} (or some part of it))

- e.g. supervised setting: $Z_i = (X_i, Y_i)$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$Z_1, \ldots, Z_n$: training data. Z_{n+1}: test datapoint Z_t

- prediction region: map $T : \mathcal{Z}^n \to \mathcal{B}$ ("subsets of $\mathcal{Z}$", sigma-algebra over $\mathcal{Z}$)

$[Z_t \in T(Z)]$ {

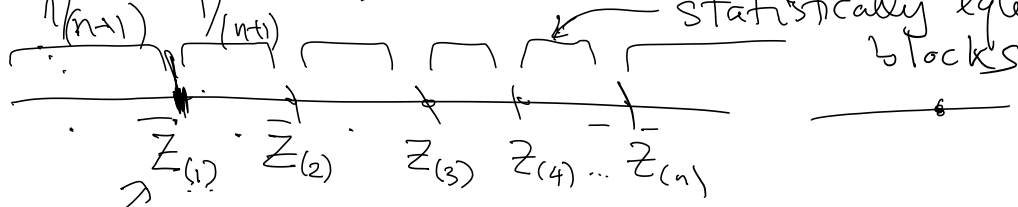
$Z_{n+1} \in T(Z_1, Z_2, \ldots, Z_n)$

- marginal coverage: $P_{Z, Z_t}\left(Z_t \in T(Z)\right) \geq 1 - \alpha$

(Z: train, Z_t: test)

- contrast with point prediction: $f : \mathcal{Z}^n \to \mathcal{Z}$

- intuition 1 ->

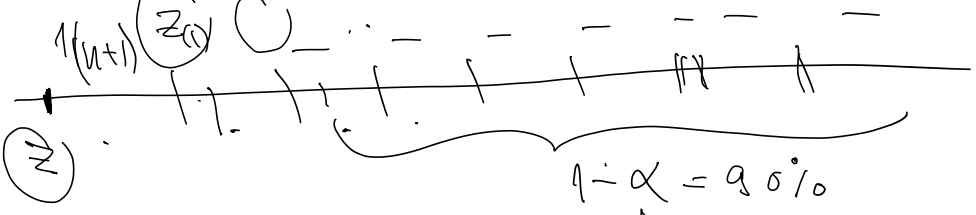


$Z_{(i)}$ - ith order statistic of sample Z

$$P\left(Z_t \in [Z_{(j)}, Z_{(j+1)}]\right) \geq \frac{1}{n+1}$$

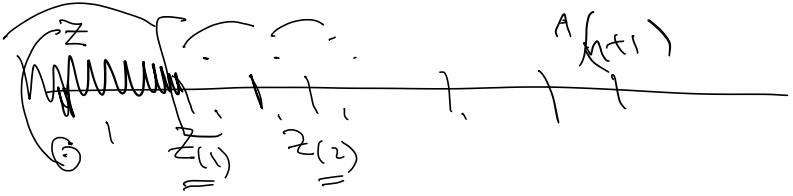
(assuming ties occur w/p = 0)

$Z_t, Z_1, \ldots Z_n \overset{iid}{\sim} f$ ← density on $\mathbb{R}$



- ranks: $\text{rank}(x, S) = \#\{s \in S : s \leq x\}$

$\text{rank}(1, \{2, 0\}) = 1$, b/c $0 \leq 1 \to 2 \neq 1$



$\left(\frac{1}{n+1}, \frac{2}{n+1}, \ldots\right)$

stat. equiv blocks

$\text{rank}(z; \{z, z_1, z_2, \ldots z_n\}) = 1$

$z_1 \rightarrow n+1 \rightarrow \frac{n+1}{n+1} = 1$

$z_2 \rightarrow n \rightarrow \frac{n}{n+1}$

$z_{k+1} \rightarrow n+1-k \rightarrow \frac{n+1-k}{n+1} = 1 - \frac{k}{n+1}$

want $\geq 1-\alpha$: $1 - \frac{k}{n+1} \geq 1-\alpha$

$\alpha \geq \frac{k}{n+1}$

$\alpha(n+1) \geq k \rightarrow k = \lfloor \alpha(n+1) \rfloor$ (floor)

floor of 0.5 = 0

$\lfloor 0.5 \rfloor = 0$

$\lfloor 1 \rfloor = 1$

Conclusion : $T(z) = \{ z : \text{rank}(z, \{z, z_1 \ldots z_n\}) \geq \lfloor \alpha(n+1) \rfloor + 1 \}$

has $1-\alpha$ marginal coverage

$T(z) = \{ z : \text{rank}(z, \{z_1, \ldots z_n\}) \geq \lfloor \alpha(n+1) \rfloor \}$

Example : $n=5$

$1-\alpha = 0.9$ 90%

$\lfloor 0.1 \cdot 6 \rfloor = \lfloor 0.6 \rfloor = 0$

$\Rightarrow T(z) = \mathbb{R} = (-\infty, \infty)$

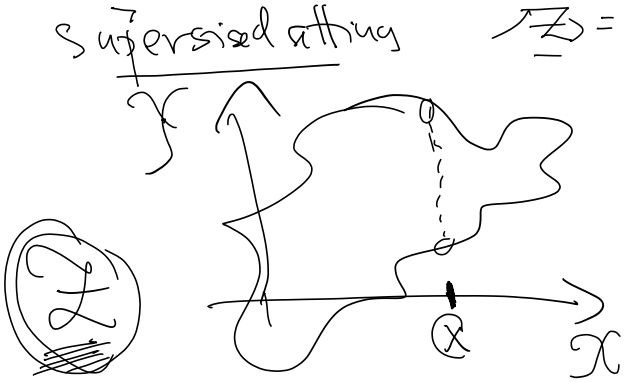
80%

$\lfloor 0.2 \cdot 6 \rfloor = \lfloor 1.2 \rfloor = 1$

$T(z) = [z_{(1)}, \infty)$

Thm Fraser '50s all "non-parametric" prediction regions need to be based on order statistics.

Supervised setting $Z_i = (X_i, Y_i)$



$$T(x; \underbrace{z_1 \dots}_{z}) = \{ y : (x, y) \in T(z) \}$$

Claim · 4:18

$$P\Big(\underbrace{(y_t) \in T(x_t, z)}_{(x_t, y_t)}\Big) = \underbrace{P\big(Z_t \in T(z)\big)}_{\geq 1-\alpha}$$

$(x_1, y_1) \dots (x_n, y_n)$

Principle: given "exchangeable" data $Z_1 \dots Z_n$

consider "symmetric" fns $\begin{bmatrix} m(z_1, z_1) \in \mathbb{R} \\ m(z_2, z) \\ \vdots \\ m(z_n, z) \in \mathbb{R} \end{bmatrix}$ preserve "exch."

& then use 1-> construction.

$Z_1 \dots Z_n \to \dfrac{\sum_{i=1}^n Z_i}{n}$

$m(z_1 \dots z_n) = m(z_{\pi_1}, z_{\pi_2} \dots z_{\pi_n})$ for any permutation π on $\{1 \dots n\}$ S_n

Def: $m(z) = m(z_\pi) \quad \forall \pi \in S_n$

symm function

$Z_1 \dots Z_n$

$m(a, b_1 \dots b_n) - \left| a - \dfrac{b_1 + \dots + b_n}{n} \right|$

$\left| z_1 - \dfrac{\sum z_i}{n} \right|$

$m(z_1, z) \quad m(z_2, z) \quad \dots \quad m(z_n, z)$

$m(a$

If $m(a,\cdot)$ is symmetric (in the last n variables)
& $z_1 \dots z_n$ are exchangeable

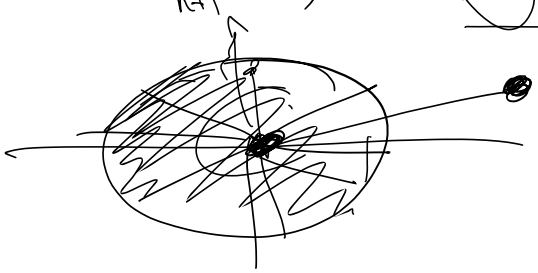
then $m(z_1, z), \dots, m(z_n, z)$ are exchangeable

z

$\downarrow$ $m(a, b_1 \dots b_n) = \left\| a - \frac{\sum b_i}{n} \right\|$ $\|\cdot\|$ norm on $\mathcal{Z}$

$\mathbb{R}$ $\|z_1 - \bar{z}\|, \|z_2 - \bar{z}\|, \dots \|z_n - \bar{z}\|$ $\leftarrow$

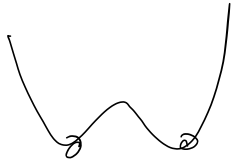
$T(z_{n+1}, z) = y_{n+1}$ tank $\left(\|z_{n+1} - \bar{z}^t\|, \{\|z_j - \bar{z}^t\|\}_{j=1:n} \right)$



$z^t = \{z\} \cup \{z_{n+1}\}$
$= \{z_1 \dots z_{n+1}\}$

$\geq \{L((n+1)\alpha)$
$\leq \lceil (n+1)(1-\alpha) \rceil\}$

E.g: $Z_i = (x_i, y_i)$



$$m(z, Z^+) = |y - \hat{f}(x, Z^+)|$$

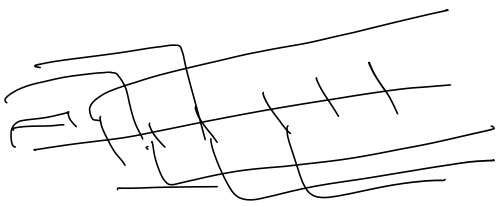
$\hat{f}(x, Z^+)$ is symmetric in the 2^{nd} variable Z^+

(prediction fn. does not depend on order)

OLS: $\hat{f}(x, Z^+) = x^T (X^T X)^{-1} X^T y$ ← OLS. ← $\min_\beta \|y - X\beta\|_2^2$

$X : (n+1) \times \dim(x)$

$y : (n+1) \times 1$



$$|y_{n+1} - \hat{f}(x_{n+1}, Z^+)| \leq q_\alpha \left\{ |y_i - \hat{f}(x_i, Z^+)| \right\}_{i=1\dots n}$$

← quantiles.

y_{n+1}

$$y_i - x_i^T (X^T X)^{-1} X^T \begin{pmatrix} y_1 \\ \vdots \\ y_{n+1} \end{pmatrix}$$

u_i

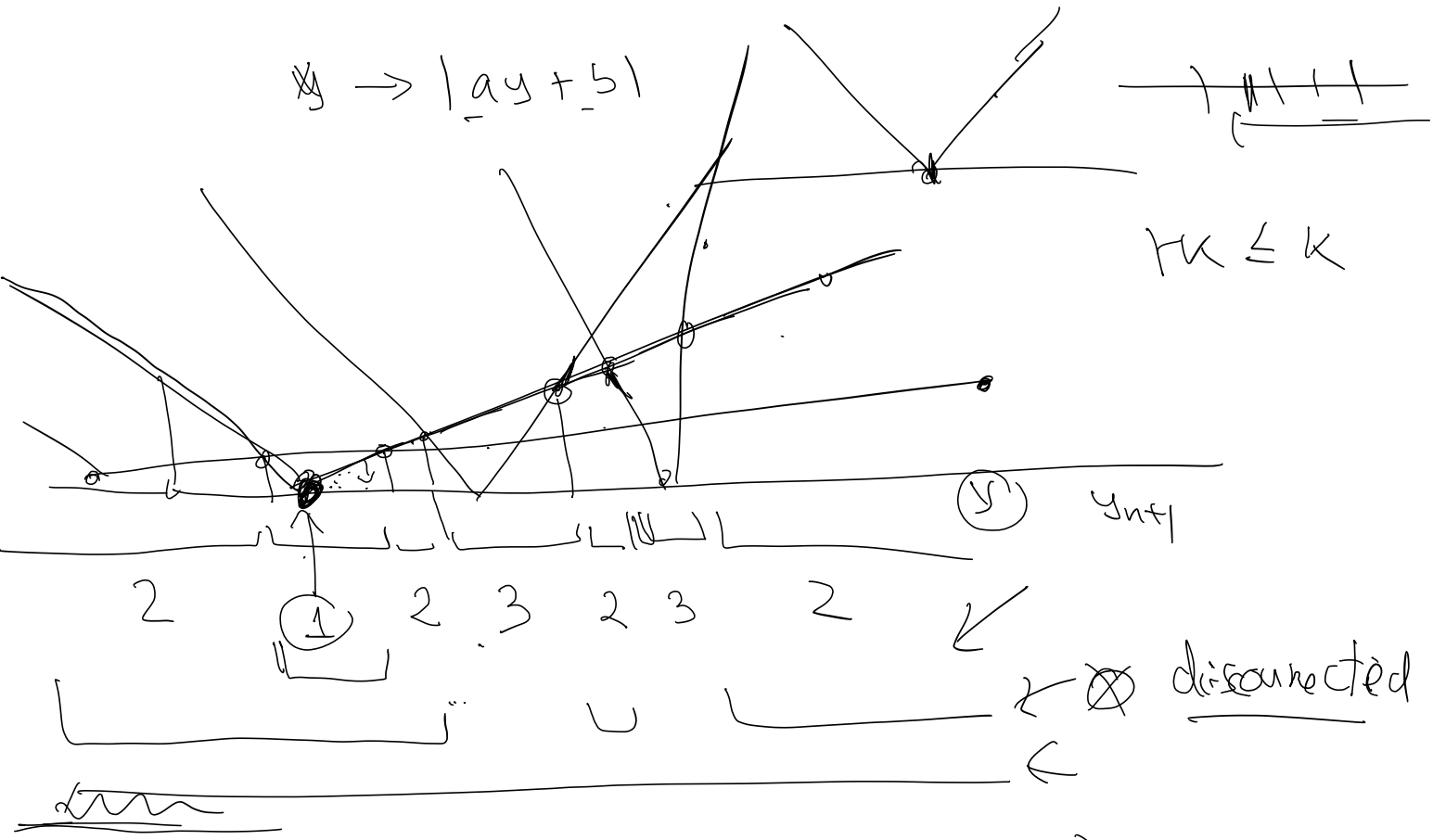
$$= y_i - u_i^T y = y_i - \sum_{j=1}^{n+1} u_{ij} y_j$$

$$= v_i^T y = a_i y_{n+1} + b_i$$

need to have sufficiently many inequalities of the form

$$|\bar{a}_{n+1} y_{n+1} + \bar{b}_{n+1}| \leq |\bar{a}_i y_{n+1} + \bar{b}_i|, \quad i = 1 \dots n$$

$y \longrightarrow |a_i \cdot y + b_i|$



Exchangeability : $(Z_1, Z_2, Z_3 \dots, Z_n \dots)$

each sequence
if for any $n \ge 1$ permutation $\pi \in S_n$

$(Z_1, Z_2 \dots Z_n \dots)$ has the same distribution as

$(Z_{\pi_1}, Z_{\pi_2} \dots Z_{\pi_n}, Z_{n+1} \dots)$

e.g. : if $(Z_i)_{i \ge 1}$ are iid. $\Rightarrow$ exch.

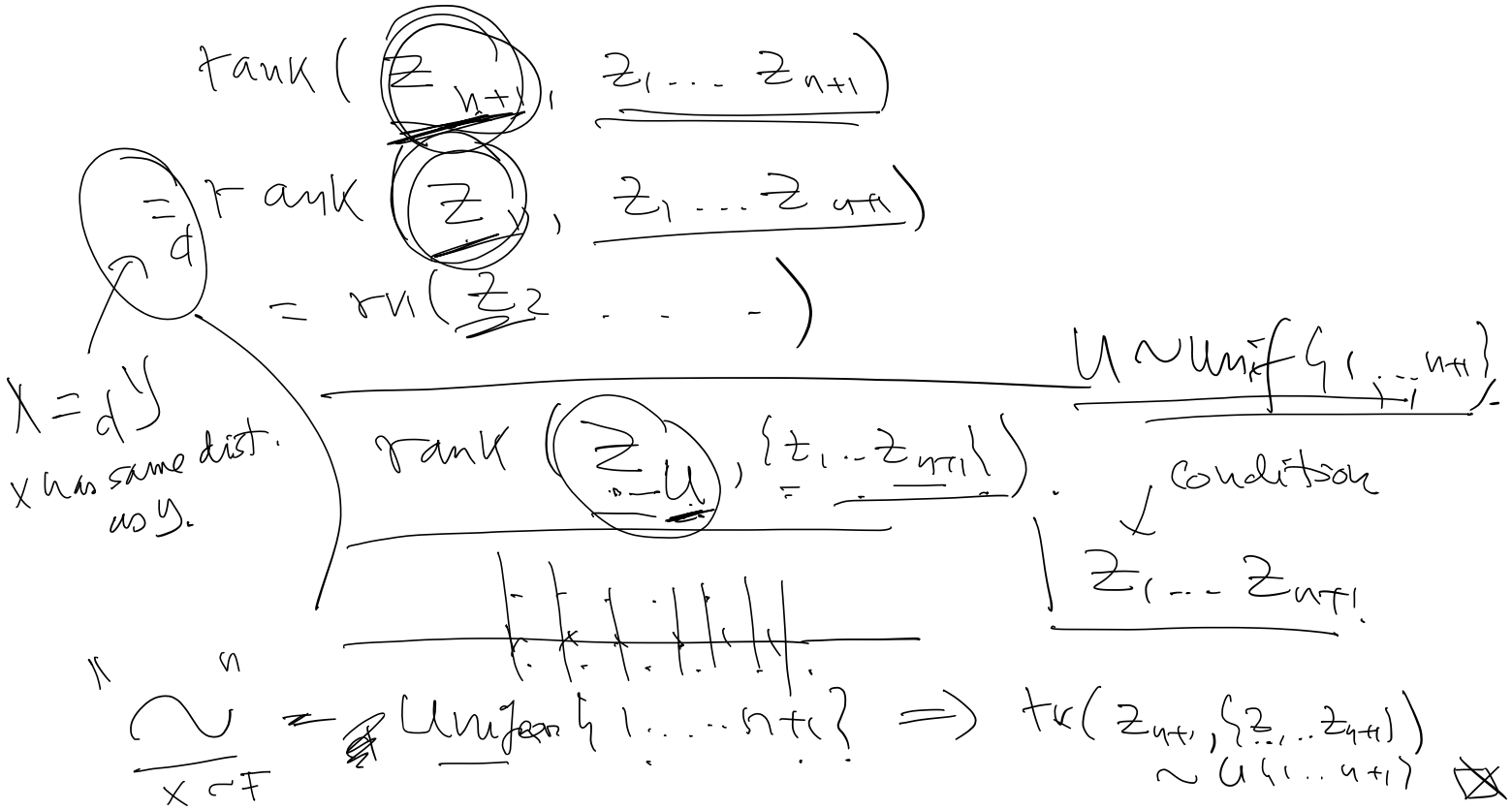
de Finetti's thm '937 : exch $\Rightarrow$ mixture of iids.

$(\exists)$ latent param Γ
s.t. cond. on Γ, $(Z_i)_{i \ge 1}$ are iid

Pf of rank construction : $\forall k \; (\underbrace{Z_{n+1}}, Z_1 \dots Z_n) \ge \dots$

claim: $(Z_1, Z_2 \dots Z_{n+1})$ are exch, then π is uniformly dist. $\{1 \dots n+1\}$

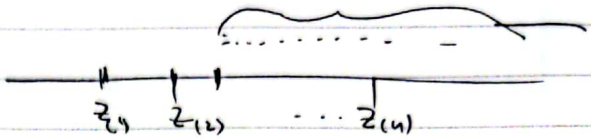
& this occur w/p $= 0$


rank(Z_{n+1}, $z_1 \dots z_{n+1}$)
$=_d$ rank(Z_j, $z_1 \dots z_{n+1}$)
$= rv(z_2 \dots)$
$U \sim Unif\{1 \dots n+1\}$
$X =_d Y$
X has same dist. as Y.
rank(Z_U, $\{z_1 \dots z_{n+1}\}$)
conditioning
$z_1 \dots z_{n+1}$
$\sim F$
$Unif\{1 \dots n+1\}$
$\Rightarrow$ rk(Z_{n+1}, $\{z_1 \dots z_{n+1}\}$) $\sim U\{1 \dots n+1\}$

STAT 991

Lec 4:

Recap. $Z = (Z_1 \dots Z_n)$ $\quad$ (1-D)



$$T(Z) = \{ z : \text{rank}(z; Z) \geq \lfloor \alpha(n+1) \rfloor \}$$

Claim: if $(Z_1 \dots Z_n, Z_{n+1})$ are exchangeable, then

$$\underset{Z, Z_{n+1}}{P}\left(Z_{n+1} \in T(Z) \right) \geq 1 - \alpha.$$ ← marginal coverage (validity)

How to use it? - reductions

(1). Functions preserving exchangeability.

(2). Supervised fitting: fiber / cross-sectional.

$$T(x, Z) = \{ y : (x, y) \in T(Z) \}$$

~~1-D const.~~

→ Eg consider maps $Z_1 \longrightarrow Z'$

$$\begin{matrix} z_1 \\ \vdots \\ z_n \end{matrix} \longrightarrow \begin{matrix} m_1(z) = z_1' \\ \vdots \\ m_k(z) = z_k' \end{matrix} = \text{call it } Z' = m(Z)$$

~~when~~ ar. Suppose $(Z_1 \dots Z_n)$ are exch. [

When are $Z_1' \dots Z_k'$ exch?

Claim: (Dean/Verducci) If for each permutation $\pi \in S_n$, there is a permutation $\pi' \in S_k$ such that

$$m(Z_\pi) = [m(Z)]_{\pi'}$$

(permute Z by π $\Rightarrow$ permute $m(Z)$ by π')

Then $Z' = m(Z)$ is exch. ~~pf~~ coordinates.

$z_1 \quad z_2 \quad z_3$

Example: $z_1 \cdot \frac{z_2+z_3}{2} \quad z_2 \cdot \frac{z_1+z_3}{2} \quad z_3 \cdot \frac{z_1+z_2}{2}$

More generally.

$$m_i(z_1 \dots z_n) = \tilde{m}(z_i; \{z\})$$

permutation inv. wrt 2nd argument.

$\{z\}, \{z_i\}$

or $m_i(z_1 \dots z_n) = \tilde{m}(z_i, \{z_{-i}\})$

used to check how well z_i "conforms" to z.

Remark: $m(z_\pi) = [m(z)]_\pi$, ~~m is a mean~~ : ~~m is~~ m is permutation equivariant. [or co-variant]

- ~~this~~ is an instance of more general group equivariance

$\forall g \in G, \exists g' \in G' : m(gz) = g' m(z)$.

- studied in
  - representation theory of finite groups [Serre '77]
  - modern perspective: Deep Network architecture design. Deep Sets [Zaheer et al '17] PointNet.

Research questions:

- ~~How~~ what are the implications? "deep conformal prediction" in theory

Key point: preserved under compositions

$$z \to z' \to z''$$

pres. equi. pres. equi.

full algo

input:

given $= (x_i, y_i)$ $i = 1 \dots n$

Or $(x_{n+1}, \cdot)$

Output:
- Prediction set for y_{n+1}.

$$\Gamma(x_{n+1}, Z)$$

- for each y_{n+1} in [how discretize?].
- calculate $W_i = \Delta(y_i, f(x_i, Z^+))$, $i = 1 \dots n+1$.
  - preserve exch. $Z^+ \to Z^+_\pi$: f. be perm. inv.
  - then, $w(Z^+_\pi) = [w(Z^+)]_\pi$ s.o. ov.
  - what happens if you permute simply
- if $\text{rank}(W_{n+1}, W) \le \lceil (n+1)(1-\alpha) \rceil$, include y_{n+1} in $\Gamma(x_{n+1}, Z)$

- How obtain f? → $f(\cdot; x)$ can be any function of z^+, if it depends symmetrically (alg a set)
- e.g. ERM: $\min_{f \in H} \frac{1}{n+1} \sum_{i=1}^{n+1} L(y_i, f(z_i))$.

  ↳ run any alg after randomly permuting data.

  $\#$ 1 alg. in ML

What we really need is $W =_d W_\pi$. $\forall \pi$. (exch. of w)

enough of: $Z^+ =_d Z^+_\pi$ $\forall \pi$.

$$W_i = \Delta(y_i, f(x_i, Z^+_{\tilde{\pi}})), \quad \tilde{\pi} \sim \text{Unif}(S_{n+1})$$

more generally

$z \to [z_\pi]_{\tilde\pi}$

$$w(z^\dagger_\pi)_{\pi^{-1}(i)} = \Delta\Big(y_{\pi\cdot\tilde\pi^{-1}(i)}, f\big(x_{\pi\cdot\tilde\pi^{-1}(i)}, z^\dagger_{\tilde\pi\circ\pi}\big)\Big)$$

$$= \Delta\Big(y_i, f\big(x_i, \underbrace{z^\dagger_{\tilde\pi\circ\pi}}_{=d\, z^\dagger_{\tilde\pi}}\big)\Big) = d.$$

$$w(z)_i = \big(y_i, f(x_i, z^\dagger_{\tilde\pi})\big)$$

So, we find $w(z^\dagger_\pi)_{\pi^{-1}(i)} =_d w(z^\dagger)_{\tilde\pi}$ $\forall i$ $\forall \pi$.

moreover/similarly: $w(z^\dagger_\pi)_{\pi^{-1}} =_d w(z^\dagger)$

$$w(z^\dagger_\pi) =_d [w(z^\dagger)]_\pi .$$

Now, if $z^\dagger_\pi =_d z^\dagger$ i.e. z is exchangeable

$\Rightarrow$ $w =_d w_\pi$, so w is exch.

ℓ $z^\dagger$ are exch.

n.b. need to use symmetric

Conclusion: if $\tilde\pi \sim \mathrm{Unif}(S_{n+1})$, & $f(\cdot, z^*)$ is any fn.

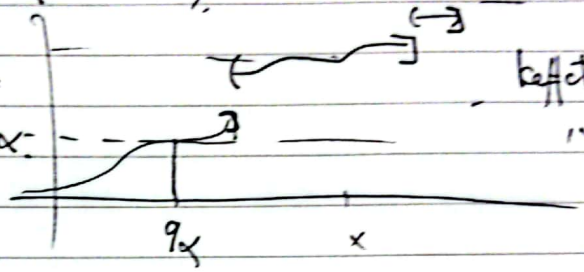
then $\Delta(y_i, f(x_i, z^*))$, $i \in [n+1]$ are exch.

Alternative presentation: in terms of quantiles.

Quantile: Let F be a cumulative distribution function (cdf) on $\mathbb{R}$.

$$F(x) = \mathbb{P}(X \le x) \quad \text{for some } X.$$

α-th quantile: $q_\alpha =$



- left-cts with limit from the right. "cagl\`ad"

$$q_\alpha = F^{-1}(\alpha) = \inf\{x : F(x) \ge \alpha\}.$$

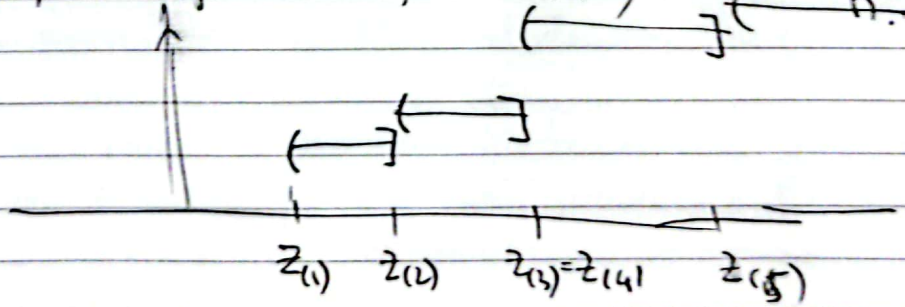
two cases:- if there is a value $\tilde{q}_\alpha$ s.t. $F(\tilde{q}_\alpha) = \alpha$.
then q_α is the smallest such value.

or else $F(q_\alpha) > \alpha$ but $\lim_{p \nearrow q_\alpha} F(p) < \alpha$

iff F has a point mass at q_α, & $\alpha \neq$ act in range of F.

~~Specific example?~~

$Z_1, \dots Z_n \to$ empirical cdf: $F_n(x) = \dfrac{|\{i : Z_i \le x\}|}{n}$



turns out: $q_\alpha = Z_{(\lceil n\alpha \rceil)}$

How remember this?

$\alpha \in (0, 1/4]$ $\to$ $F_n(x) \ge \alpha \iff \hat{F}_n(x) \ge \frac{1}{4} \iff$

$x \ge Z_{(1)} \Rightarrow q_\alpha = x = Z_{(1)}$

back to CI: $\text{rank}(z, \{z_1, \dots, z_n\}) \geq \lfloor (n+1)\alpha \rfloor$

$\Leftrightarrow z \geq q_\beta(\{z\})$, $\beta = \frac{\lfloor (n+1)\alpha \rfloor}{n}$

slight abuse of notation.

$\left[\Leftrightarrow \text{rk}(z, z^+) > \lfloor (n+1)\alpha \rfloor + 1 \Leftrightarrow z \geq q_{\beta'}(\{z, z\}), \beta' = \frac{\lfloor (n+1)\alpha \rfloor + 1}{n+1}\right.$

- Computational burden: need to re-~~call~~ run Alg 1 for all $y \in \mathcal{Y}$. [ε-train f]
  - big drawback.
  - efficient for certain special ML models
    - ridge & lasso regularization

Examples: ← due to comput. burden, ~~not~~ not easy to find good example [later]

Alternative perspective via Statistical decision theory. (Takeuchi '75+)

- Decision theory: general unifying framework for statistics
  - usually for estimation problems (i.e. fixed param)
  - loss, risk → admissible, minimax optimal etc.
- Extension to prediction [of a random quantity] by Takeuchi & collaborators.

$\mathbb{E}_{(x,y)\sim P_\theta} \sigma(x,y)$

$P_{(x,y)\sim P_\theta, u}\left(y \in \overline{T}_u(x)\right) \geq 1-\alpha \quad \forall \theta \in (H)$

- $(X,Y) \sim P_\theta, \quad \theta \in (H)$

randomized pred. region

- $\overline{T}_u(x) = \{ y \mid f(x,y) \geq u \}, \quad f : \mathcal{X} \times \mathcal{Y} \to [0,1]$
  $u \sim U[0,1].$

- marginal cov: $\forall \theta : \; P_{(x,y)\sim P_\theta, u}$

- $\overline{T}_u$ is a locally best prediction region at some $\theta_0 \in (H)$,
  for some loss $L : \mathcal{Y} \to [0,\infty)$.
  & Lebesgue measure μ over $\mathcal{Y}$. (if $\overline{T}_u$ minimizes → marginally valid) $\forall \theta \in (H')$

$$\mathbb{E}_{X \sim P_{\theta_0}, u} \int_y \mathbb{I}\left(y \in \overline{T}_u(x)\right) L(y) \, d\mu(y)$$

$$= \int_y \mathbb{E}_{X \sim P_{\theta_0}} \sigma(X, y) \, L(y) \, d\mu(y)$$

- ~~Strictly~~ Exact coverage: $P_\theta\left(y \in \overline{T}_u(x)\right) = 1-\alpha \quad \forall \theta \in (H)$

- Suppose (T) complete sufficient statistic $W = w(X,Y)$

  ↪ $\mathbb{E}_\theta W = 0, \; \forall \theta \in (H) \Rightarrow W = 0 \; P_\theta\text{-a.s.} \; \forall \theta$

  ↪ $\mathcal{L}((X,Y) \mid W)$ does not depend on θ.
  (↑ law of)

example: $Z_1 \dots Z_n$ iid $\sim f$ - density on $\mathbb{R}$.

$W = \{Z_1 \dots Z_n\}$ = order statistics - complete. suff.

$Z \mid W$ = ranks - dist unif over S_n.

then: Exact. cov $\Leftrightarrow P_{\cdot}\left(y \in \overline{T}_u(x) \mid W = w\right) = 1-\alpha \quad \forall w$

(↑ doesn't depend on θ)

~~map~~ ~~$\phi = \phi(x,$~~

Suppose: $\exists$ 1-1 map $(x, y) \longleftrightarrow (x, v, w)$ (comp. suff.)

s.t. $y = \phi(x, v, w)$ has jacobian $|\exists| = |\exists|(x, v, w)$

Theorem. (Takeuchi) The locally best prediction region is.

$$f\left(\underset{\phi(x,v,w)}{x, y}\right) = \begin{cases} 1 : P_\theta(x, v \mid w) > c_w \cdot P_{\theta_0}(x) \cdot L\left(\overset{\phi(x,v,w)}{y}\right)|\exists| \\ \cdots \quad = \quad \leftarrow \text{fulfill constraint.} \\ 0 \quad < \end{cases}$$

pf: Follows from Neyman-Pearson lemma. w/ given W

"null density", $\sim P_{\theta_0}(x)\, L(y) \sim P_{\theta_0}(x)\, L\left(\phi(x,v,w)\right)|\exists|$

"alternative": $\sim$ density of (x, y) give $W = w$.

$P_\theta(x, v \mid w)$.

reject null $\Leftrightarrow$ $y \in$ Set.

Example: $Z_1 .. Z_{n+1} \overset{iid}{\sim} f$: &

Exact. coverage $P\left(Z_{n+1} \in \Gamma(Z) \mid \{Z_1 .. Z_{n+1}\}\right) = 1-\alpha$

$\forall\, z_1 .. z_{n+1}$

$$\frac{1}{n+1} \sum_{i=1}^{n+1} f\left(Z_i \mid \{z_1, .. z_{n+1}\}\right) = 1-\alpha.$$

Lec 5

$Z \in \mathcal{Z}$

* $Z \sim P_\theta$, $\theta \in \Theta$ - statistical model

* a statistic: $W : \mathcal{Z} \to \mathcal{Z}'$ is complete (for this model) if for all $\psi : \mathcal{Z}' \to \mathbb{R}$

$E_\theta \psi(W) = 0 \ \forall \theta \in \Theta \Rightarrow \psi(W) = 0 \ P_\theta$ a.s. $\forall \theta \in \Theta$

next, introduce suffi. (p. 3) ← if

See "Theory of Point Estimation" book, p. 42

* — ~~bug~~ - examples & non-examples (?)

(A) $Z_1, Z_2 \sim N(\theta, 1)$, $\theta \in \mathbb{R}$.

suffi:

(1) $W = Z_1 - Z_2$ → not complete — X

$E_\theta W = 0 \ \forall \theta$, but $W \neq 0$

(2) $W = Z_1$ → X

$E_\theta f(W) = 0 \ \forall \theta$

~~$E_{Z \sim N(\theta,1)} f(W + \theta) = 0 \ \forall \theta$~~

$\int f(w) \varphi(w+\theta) \, dw = 0 \ \ \cdot d/d\theta$

$\int f(w) \varphi'(w+\theta) \, dw = 0 \ \ \forall \theta$

$\int f(w) \varphi^{(k)}(w+\theta) \, dw = 0 \ \ \forall \theta$.

$\{\varphi^{(k)}\}_{k \geq 0}$ form basis for $L^2(\mathbb{R})$ → $f = 0$ in $L^2(\mathbb{R})$

$\Rightarrow f = 0$ wrt. Lebesgue measure

(3) $W = \frac{Z_1 + Z_2}{2}$ suff.

$W \sim N(\theta, \frac{1}{2})$. → follows from (2). ✓

(B) $Z_1, Z_2, \ldots Z_n \sim U[\theta, \theta+1]$, $\theta \in \mathbb{R}$. ✓

$W = (Z_{(1)}, Z_{(2)})$ → not complete,

$W = Z_{(2)} - Z_{(1)}$

$\therefore E_\theta(W - \frac{1}{3}) = 0. \quad \forall \theta.$

Here,

(c) $Z_1 \ldots Z_n \overset{iid}{\sim} f$, where f any density on $\mathbb{R}$. ✓

$W = \{Z_1, \ldots, Z_n\} = \{\}$

complete

[construct rich parametric subfamily.

$\propto \exp\left(-\sum_{j=1}^{n} \theta_j \left(\sum_i Z_i^j\right) - \sum Z_i^{2n}\right)$]

* A statistic $W : \mathcal{Z} \to \mathcal{Z}'$ is sufficient. (for Z in this model)

if the distribution of $Z \mid W = w$ does not depend on θ, ~~w a.s.~~ P_θ a.s.

"all info about θ is in W"

- Sufficient & complete

"has info about θ, & no ancillary info"

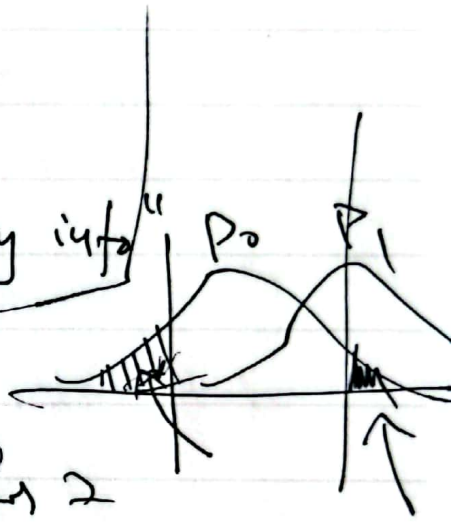


fig 2

Neyman-Pearson Lemma (30's).

- Given two proba densities P_0, P_1,

~~fixed~~

an optimal test : $\phi : \mathcal{Z} \to [0,1]$.

- interpreted as, given Z, reject H_0 w/ p $\phi(z)$

of the null hypothesis

$H_0 : Z \sim P_0$.

of level α :

$\mathbb{E}_{Z \sim P_0} \phi \leq \alpha$

so $P_{H_0}(\text{reject}) \leq \alpha$

Maximizing the power

$\mathbb{E}_{Z \sim P_1} \phi$ $\quad P_{H_1}(\text{reject})$

has the form.

$$\Phi(z) = \begin{cases} 1 & p_1(z) > c \cdot p_0(z) \\ \tau & = \\ 0 & p_1(z) < c\, p_0(z) \end{cases}$$

where τ is the unique constant

s.t. $\exists \tau$ s.t. $\mathbb{E}_{p_0} \Phi = \alpha$.

Details of Takeuchi's Theorem:

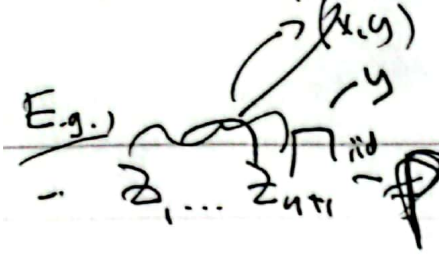
constraint: $\int_{x,y} f(x,y)\, \overbrace{p(x,y|w)}^{\text{null}} = 1-\alpha$.

max: $\int_{x,y} f(x,y)\, p(x|w)\, L(y)\, d\mu(y)$.

alternative.

$f = 1$ if alt density $>$ c·null density. D.

& change variables



E.g.) $Z_1, \dots Z_{n+1} \sim$ density on $\mathbb{R}^d$

- $W(Z) = \{Z_1, \dots Z_{n+1}\}$ - complete; suff.
- $Z^\dagger \longleftrightarrow (W, R)$ one-to-one. [a.s. distinct]. assume in everything below.
  - R: ranks.
- $Z^\dagger \mid W$ ~ uniform over Π.
- exact. coverage: $P\left(y \in T(x) \mid W = w\right) = 1-\alpha$

$$P\left(Z_{n+1} \in T(Z) \mid W = w\right) = 1-\alpha.$$

$$= \frac{1}{(n+1)!} \cdot \{Z_i\}$$

$$T(Z) = \left\{ z_{n+1} \mid t(\dots, z_{n+1}) \geq t(z_{n+1}, Z) \geq u \right\}$$

$$\frac{1}{(n+1)!} \sum_{\pi \in S_{n+1}} P_u\left(Z_{\pi(n+1)} \in T\left(Z^\dagger_{-\pi(n+1)}\right)\right)$$

$$= \frac{1}{(n+1)!} \sum_{\pi \in S_{n+1}} f\left(Z_{\pi(n+1)}, Z^\dagger_{-\pi(n+1)}\right)$$

Suppose f is symmetric in last n vars

[or, average it above]

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} f\left(Z_i, \{Z^\dagger_{-i}\}\right) = 1-\alpha.$$

can make it depend on full z^+

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} f(z_i, z^+) = 1-\alpha. \quad (*)$$

- optimal pred region w/ exact $1-\alpha$ coverage, for some $z_1 \dots z_{n+1}$ iid P_0:

given $w = \{z^+\}$

P_0 & for some $f(\cdot, z^+)$ satisfying $(*)$.

$f = 1$ iff.

$$P_\theta(x, v \mid w) > C_w \, P_{\theta_0}(x) \, L(y) \cdot |\mathcal{J}|$$

($|\mathcal{J}|$: = const.; $L(y)$: choose const)

$(x, v \mid w) \to (z, R, \{z\})$

$P_{\theta_0}(x) \to P_0(z_1) \dots P_0(z_n)$

Unif. over all ranks, does not dep. on $\{z\}$

$= C$ >

$\Leftarrow$ $P_0(z_1) \dots P_0(z_n) < C \cdot (\{z\})$

$\Rightarrow$ $P_0(z_{n+1}) > C \cdot (\{z\})$

Thus: optimal pred region:

~~$T(z_{n+1}) = \{ P_0(z_{n+1}) > C(\{z\}) \}$~~

~~s.t.~~ $+ \{ P_0(z_{n+1}) = C(\{z\}) \}$

$$f(z_{n+1}, \{z_i\}) = \begin{cases} 1, & P_0(z_{n+1}) > c(\{z_i\}) \\ \gamma, & = \\ 0 & < \end{cases}$$

s.t. $\frac{1}{n+1} \sum_{i=1}^{n+1} f(z_{n+1}, \{z_i\}) = 1-\alpha.$

$\Rightarrow$ Sort $P_0(z_{i_1}) \le P_0(z_{i_2}) \le \dots \le P_0(z_{i_{n+1}})$

$0 \dots 0 + \underbrace{\gamma + \dots + \gamma}_{\text{if. equal}} + 1 + \dots + 1 = \le (n+1)(1-\alpha)$

$\Leftarrow$ Conformal inference. w.l. Conformity measure P_0.

Conclusion : ~~CI~~ is. opti

when $z_i \sim$ iid P. P density on $\mathbb{R}$.

- among all exactly valid pred. sets.

$$P(z_{n+1} \in T_\alpha(z)) = 1-\alpha \quad \forall P.$$

$$T_\alpha(z) = \{ z_{n+1} : f(z_{n+1}; z) > \alpha \}.$$

- CI minimizes average ~~length~~ mis-coverage prob.

$$\int P(z_{n+1} \notin T_\alpha(z)) \, d z_{n+1}, \quad \text{under } P = P_0$$

with conformity measure $P_0(z)$

A lot to explore

- other loss fns L ?
- ~~multip~~ predicting multiple outputs.
  e.g., z_{n+1}, z_{n+2} ? shrinkage
- supervised setting ?
- optimality against more than a fixed density ?
- implicit optimality ~~for~~ in ML ? existing methods ?
- minimax optimality?

Split (inductive) CI — [Vovk '02, Papadopoulos et al '02]

$$Z = (\underbrace{Z_1, \dots, Z_{n_1}}_{Z_T}, \underbrace{Z_{n_1+1}, \dots, Z_{n_0}}_{Z_C})$$

"training set" "Calibration set"

-- "proper training set" → indep. of Z, use it to train $fm : \mathcal{Z}^{n_1+1} \to \mathbb{R}$

- construct

$$M_\cdot = \Big(m(Z_{n_1+1}, Z_T),\ m(Z_{n_1+2}, Z_T), \dots,\ m(Z_{n_0}, Z_T),\ m(Z_t, Z_T) \Big)$$

Claim conditional on Z_T.

if (Z_C, Z_t) exch,

then, entries of $M_\cdot$ are exch.

- so this also holds unconditionally over Z_T if $Z_T \perp\!\!\!\perp Z_C$ (doesn't have to be iid)
- so, can use 1-D rank. construction for $M_\cdot$: $\text{rank}(m(Z_t, Z_T); M_{\cdot t})$
- no need for retraining

$$\geq \lfloor \{(n_0 - n) + 1\}\alpha \rfloor$$

$$(\text{or. } \leq \lceil n(1-\alpha) \rceil$$

good if small → ^ <11

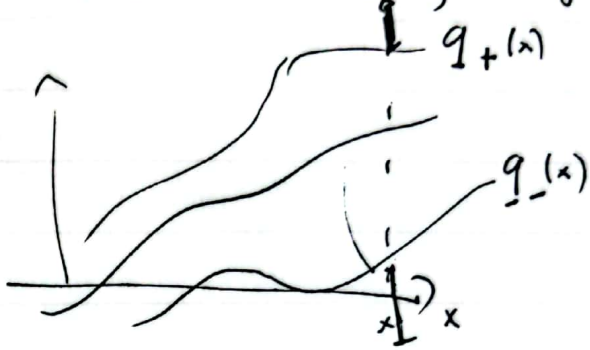
e.g. $z=(x,y)$, $M(z,z_T) = |y - f(x,z_T)|$

any predictor of y trained on z_T, n. sym. model

Romano, Patterson, Candès, 2019

- examples: quantile reg.

e.g. $z=(x,y)$



$$M(z,z_T) = \max\{q_-(x,z_T) - y,\ y - q_+(x,z_T)\}$$

$\longrightarrow$

- $q_- = q_+ = q \rightarrow |q(x,z_T) - y|$
- $q_- \le q_+ \rightarrow$ "distance" to region $[q_-, q_+]$
  - $\rightarrow$ negative inside the region
  - minimized at $y = \frac{q_- + q_+}{2}$

pred region: $M(z_{t+1}, z_T) \le q_\beta\left(M(z_i, z_T)\right)$

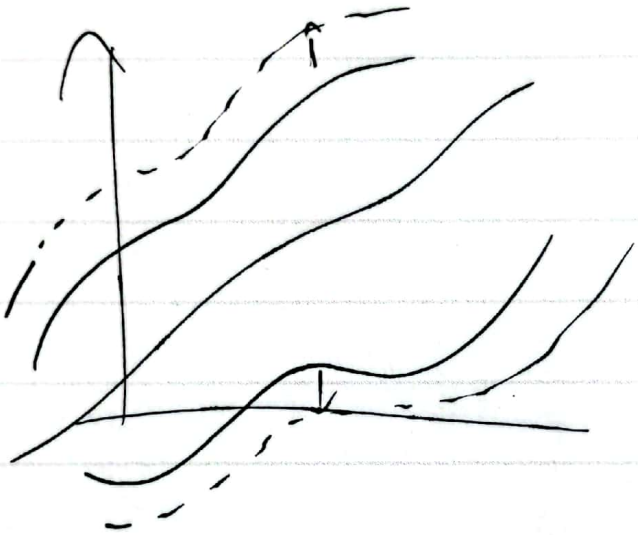
$r_K \le \lceil (n+1)(1-\alpha) \rceil$ $\quad (n+1)(1-\alpha) = n\beta$

$q_\beta = z_{(\lceil n\beta \rceil)}$ $\quad \beta = \left(1 + \frac{1}{n}\right)(1-\alpha)$

$$q_-(-x) - y \leq q_\beta$$
$$y - q_+(x) \leq q_\beta$$

$$y_t \in [\, q_-(x_t) - q_\beta,\ q_+(x_t) + q_\beta \,].$$

$$y_t \in [\, q_-(x_t) - q_\beta,\ q_+(x_t) + q_\beta \,]$$



⇒ look at their paper for experiments.

STAT 991. Lec 6

- Recall inductive Conformal Prediction
  - $Z = \{\underbrace{z_1, \ldots, z_{n_1}}_{Z_T}, \underbrace{z_{n_1+1}, \ldots, z_{n_0}}_{Z_C}\}$ → n.
  - $M = (m(z_{n_1+1}, Z_T), \ldots, m(z_{n_0}, Z_T))$.
  - $T(Z) = \{\text{rank}(z, M) \geq \lfloor (n+1)(\alpha) \rfloor, \lceil (n+1)(1-\alpha) \rceil\}$.

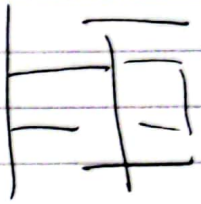
Claim: • Condit for any fixed $Z_T = z_T$.

• if (Z_C, Z_t) excl., then $P(Z_t \in T(Z)) \geq 1-\alpha$.

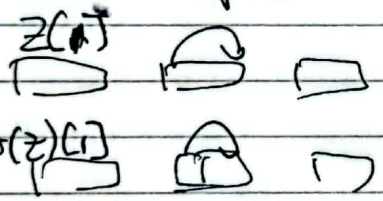
- example: quantile reg $m(z, Z_T) = \max\{q_-(x, Z_T) - y, y - q_+(x, Z_T)\}$ [Romano et al'19]
  → see paper for details
-

- Beyond marginal validity : - standard guarantee. $P(y \in C(x)) \geq 1-\alpha$
  - conditional guarantee $P(y \in C(x) \mid A(x)=a) \geq 1-\alpha \quad \forall a.$
  → almost always better.

- "Cluster-conditional" / Mondrian CP [Vovk et al '03]



  - K pre-defined clusters,
  - $Z[k]$, $k \in [K] := \{1 \dots K\}$ - data in clus. k.
  - Mondrian scor.: $A : \mathcal{Z}^n \to \mathbb{R}^n$
    - permutation equivariant - within each cluster.
      $\forall$ perm π_k of $Z[k]$ $\quad A(Z[1], \dots, Z[k]_{\pi_k}, \dots Z[K])[k] = A(Z)[k]_{\pi_k}$



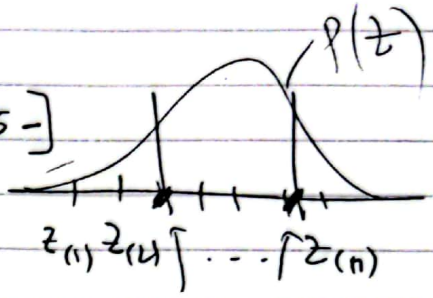
  - Algo: CI in the cluster of $Z_t = (X_t, Y_t)$; has 1-cluster validity

- Systematic approach [Vovk '13], validity conditional on
  → a.k.a PAC & connected to tolerance regions
  - Training set Z
  - Test features X_t → mostly impossibility in finite sample
    some asy. results Lei & Wasserman14
  - Test labels Y_t

  → Apply CI per class - "Mondrian"
    - Vovk, Lei
  → Confidence-interval/Test duality
    - Guan & Tibshirani.

Training-set cond. validity

- Tolerance regions [Wilks '41, Wald '43, Tukey '45]

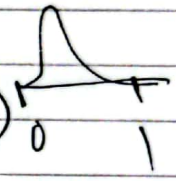


· $Z = (Z_1 \dots Z_n) \sim_{iid} P$ density on $\mathbb{R}$.

· find T s.t. $T: \mathbb{R}^n \to \mathbb{R}$: $\mathbb{E}_Z \overbrace{P_{Z_t}(Z_t \in T(Z))}^{\text{needs id, excl not enough } Z_t \sim P} \geq 1-\alpha$

· $P_Z\left\{ \underbrace{P_{Z_t}(Z_t \in T(Z))}_{\in [0,1]} \in [b,c] \right\} \geq 1-\alpha$

- Claim: $P_{Z_t}(Z_{(r)} \leq Z_t \leq Z_{(s)}) \sim \text{Beta}(s-r, n+1-(s-r))$ T_Z random.
if P_t is continuous.



$[\text{Beta}(a,b) : x^{a-1}(1-x)^{b-1}]$

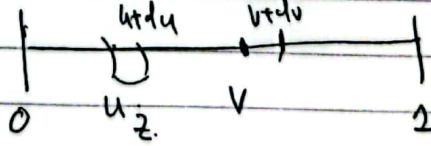
Proof sketch

- statement unchanged if transform $Z_i' := F(Z_i)$, F - cdf of f.
(s/c F is strictly increasing)
thus, assume wlog. $Z_i \sim U[0,1]$.

then: $P_{Z_t}(Z_{(r)} \leq Z_t \leq Z_{(s)}) = Z_{(s)} - Z_{(r)}$.

- Fact: joint distribution of $(Z_{(r)}, Z_{(s)})$ has pdf.
$\propto u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}$.

Heuristically.



$P(Z_{(r)} \in [u, u+du], Z_{(s)} \in [v, v+dv]) \sim u^{r-1}\,du\,(v-u-du)^{s-r-1}\,dv\,(1-v-dv)^{n-s}$

multinomial pmf

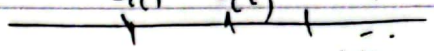
* obtain marginal density of $Z_{(s)} - Z_{(r)}$ by integration, get desired claim ☒

$Q \sim \text{Beta}(s-r, n+1-(s-r))$

if $T(z) = [z_{(r)}, z_{(s)}]$, need to choose r, s such that

$\mathbb{E}\,\text{Beta}(\ldots,\ldots)\, Q \geq \alpha \iff \frac{s-r}{n+1} \geq \alpha$

$\mathbb{P}\left(Q \in [b,c]\right) \geq 1-\alpha.$



More general perspectives: · statistically equivalent blocks (Tukey 1947)

· $T(z) = \bigcup_i \left[z_{(r_{2i-1})}, z_{(r_{2i})}\right]$, for any non-decreasing sequence $(r_i)_{i \geq 1}$

Same coverage as $[z_{(a)}, z_{(s)}]$: $b-a = \sum r_{2i} - \sum r_{2i-1}$

- ~~Sequential~~ (sequential) = construct sequence of ~~tolerance~~ statistically equiv blocks

[Tukey & Fraser]

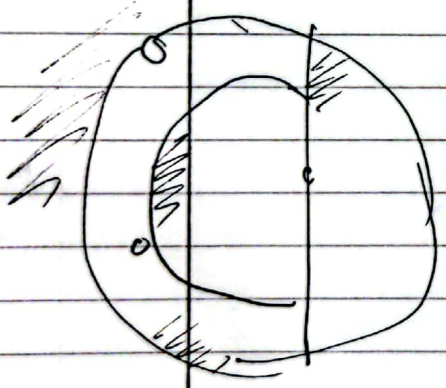
· start w/ fn $a_1 : \mathcal{Z} \to \mathbb{R}$, (assume all values are distinct)

· iteratively for $j \geq 2$

$i_1 = \arg\max_i a_1(z_i)$

- construct function a_j, depe possibly depending on $\{z_{i_1}, \ldots z_{i_{j-1}}\}$

- $i_j = \arg\max_{i \notin \{i_1, \ldots i_{j-1}\}} a_j(z_i)$



" partition space according to the functions $(a_i)_{i \geq 1}$ "

then

· let $A_j = \{z : a_j(z) \geq a_j(z_{i_j})\}$

$B_j = \left(\mathcal{Z} \setminus \bigcup_{i<j} A_i\right) \cap A_j$ are stat. equiv. blocks

~~· assumes iid, stronger than exch~~

· but more flexible ~~than standard~~ se conformal.

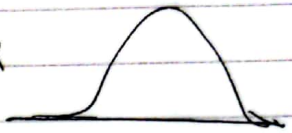
~~(stronger assumption => more flexibility)~~

Clopper-Pearson-CI

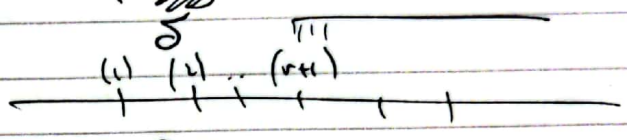
[Patel et al' 2020]

Equivalence of Tolerance regions, Inductive CP, PAC Pred. Sets

For Aim: $P_Z\{P_{Z_t}(Z_t \in \bar{T}(Z)) \geq 1-\delta\} \geq 1-\alpha$



$\Leftarrow$ δ



$T(Z) = [Z_{(r+1)}, +\infty)$.

$Q :=$

Wilks: $P_{Z_t}(Z_t \in T(Z)) = P_{Z_t}(Z_{(r+1)} \leq Z_t) \sim Beta(n-r, r+1)$

Let $F_{Beta(u,v)}$ be cdf of $Beta(u,v)$

need $P(Q \geq 1-\delta) \geq 1-\alpha$

$P(Q \leq 1-\delta) \leq \alpha$ $\quad F_{Beta(n-r, r+1)}(1-\delta) \leq \alpha$ $\cdots (*)$

Claim: $F_{Binomial(n,p)}(r) = F_{Beta(n-r, r+1)}(1-p)$

check by taking derivative wrt p.

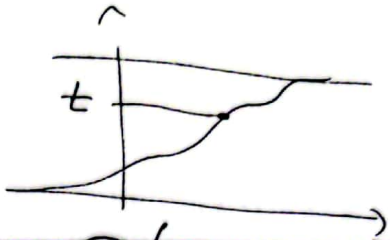
$(*) \Leftrightarrow F_{Bino}(n, \delta)(r) \leq \alpha$ $\quad (**)$

Given (n, r), find largest $\delta = u(r)$ s.t. $(**)$ holds

This $u(r)$ is known as the Clopper-Pearson upper confidence bound (UCB) for p_0

p_0, when $R \sim Bino(n, p_0)$.

Claim: we have $P_R\{p_0 \in [0, u(R)]\} \geq 1-\alpha$



Fact : $\forall$ R.V. R w/ cdf F. $P(F(R) \le t) \le t$, $\forall t \in \mathbb{R}$.

Goal : $P(p_0 \in [0, U(R)]) \ge 1-\alpha$

$p_0 < U(R)$: but $p_0 \to F(r, p_0)$ is decreasing.

$\Leftrightarrow$ $F_{Bino(n,p_0)}(R) > F_{Bino(n,p_0)}(r) = \alpha$. [$r \le n-1$; $r = n$, constant]

$P(F(R) > \alpha) = 1 - P(\underbrace{F(R) \le \alpha}_{\le \alpha}) \ge 1-\alpha$ □

Note: Clopper-Pearson CI relies on general principle of "inverting the CDF" for constructing CIs.

Conclusion : One-sided Tolerance region $\Leftrightarrow$ (one-sided) Clopper-Pearson CI.

Alternative way to present it:

Inductive Conformal prediction produces PAC confidence sets. [Vovk '13]

given map f, let

$T(z) = \{z : f(z) \ge q_{\frac{r+1}{n}}(\{f(z_i), i \in [n]\})\}$

$= \{z : \text{rank}(f(z), \{f(z_i)\}_{i \in [n]}) \ge r+1\}$

↳ ICP w/ CM f

We have shown that : if $\alpha \ge F_{Bino(n,\delta)}(r)$, then

if (z, z_t) are iid.

$P_z\{P_{z_t}\{z_t \in T(z)\} \ge 1-\delta\} \ge 1-\alpha$ □

[this follows b/c. $T(z) = \left[\{f(z)\}_{(r+1)}, \infty\right)$ = our previous construction

& (**) Gobbi

[so, Vovk's 2013 result already follows from ~~Wilks~~ Wilks's 1941 result]

Another perspective

Nested pred. sets / "PAC pred sets". [Park, Bastani '20]

- Consider $(T_\tau)_{\tau \in \mathbb{R}}$; $T_\tau : \mathcal{X} \to \mathcal{Y}$

  $T_\tau(x) = \{z : f(z) \geq \tau\}$ → nested: $T_{\tau_1} \subset T_{\tau_2}$ if $\tau_1 \geq \tau_2$

- $R_\tau = \sum_{i=1}^{n} \mathbb{I}(z_i \notin T_\tau) = \#\{i : f(x_i, y_i) < \tau\}$

- choose $\hat{\tau} = \hat{\tau}(z_1, \dots, z_n)$ s.t. $F_{\text{Bino}(n,\delta)}(R_{\hat{\tau}}) \leq \alpha$. [†]

- Claim: $P_z\{P_{z_{n+1}}\{z_{n+1} \in T_{\hat{\tau}(z)}\} \geq 1-\delta\} \geq 1-\alpha$.

Equivalence to tolerance region, etc:

- Let r be the largest integer s.t. $F_{\text{Bino}(n,\delta)}(r) \leq \alpha$.

  Solve [†] s.t. $\hat{\tau}$ is as large as possible. Then

  $R_{\hat{\tau}} = r \quad = \#\{i : f(z_i) < \hat{\tau}\}$

- Let $f_{(i)}$ be order stats. of $\{f(x_i, y_i)\}$

  then $f_{(r)} < \hat{\tau}$, $f_{(r+1)} \geq \hat{\tau}$

  $\hat{\tau}$ maximal $\Rightarrow$ $\hat{\tau} = f_{(r+1)}$

- $T_{\hat{\tau}} = \{z : f(z) \geq f_{(r+1)}\}$ $\Rightarrow$ same as Wilks's construction. ☒

Conclusion: Saw equivalence of

- 1-sided tolerance regions
- 1-sided Clopper-Pearson confidence interval
- Conditional validity of ICP
- PAC prediction sets

Note: All this can be also extended directly to the supervised setting

Feature-conditional validity

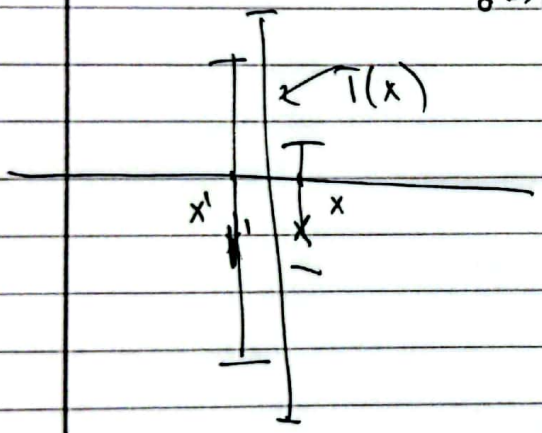
- $Z_i := (X_i, Y_i)$, $\sim_{iid} P$, $Z_t = (X_t, Y_t) \sim P$

for P-almost all x: $P_{Z, Y_t}\{ Y_t \in \bar{T}(Z, X_t) \mid X_t = x \} \geq 1-\alpha$.

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$

Thm [Lei, Wasserman '14] if T has $1-\alpha$ - cond. validity, for any P, then, given any P and non-atom x of P.

$$P\left\{ \lim_{\delta \to 0} \operatorname{ess\,sup}_{x' : \|x'-x\| \leq \delta} \mu\left(T_x(x')\right) = \infty \right\} = 1.$$



μ: Lebesgue measure.

essential supremum (smallest supremum when excluding measure 0 sets)

Possible remedy: asymptotic conditional coverage

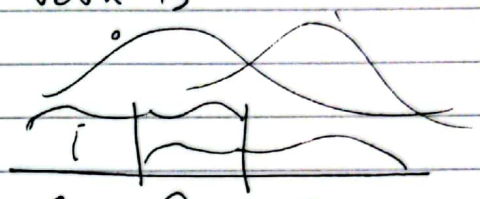
$$P\text{-a.e. } x: \quad \sup_{x \in \operatorname{supp}(x)} \left\{ P_{Z, Y_t}\left(Y_t \in \bar{T}(Z, X_t)\right) - \alpha \right\}_+ \to_P 0.$$

algorithm: [partition $\mathcal{X}$ into hypercubes; within each region, estimate marginal density of Y [if $P(x,y)$ smooth $P(x,y) \simeq p(y) \cdot \text{const}$ in small region]; apply C.P w.l. c.m $\hat{P}$]

Label-conditional validity

$$P(y \in \Gamma(x) \mid y = y) \geq 1-\alpha \quad \forall y \in \mathcal{Y}$$

- special case of Mondrian CP [Vovk et al '05]
  - in the inductive case, class-conditional marginal guarantee stated by Vovk '13



- Lei14 : - K-class classification :
  - min. ambiguity :
    - $\min P(C_{01})$ $\quad P(\_ \mid y=j)$
      s.t. $C_0 \cup C_1 = \mathcal{X}$, $P_j(C_j) \geq 1-\alpha_j \; \forall j=0,1$

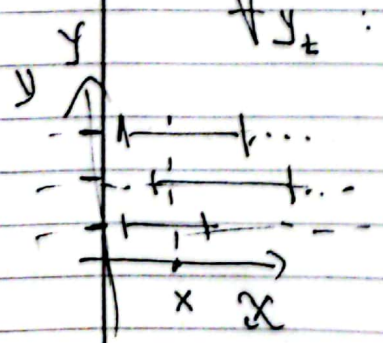
- using Neyman-Pearson lemma :

$$C_0 = \{x : P(y=1 \mid x) \leq t_0\}$$
$$C_1 = \{x : \quad \geq t_1\} \cup C_0$$

- estimation by plug-in, under margin conditions, conv. rate
- finite-sample coverage via I CP.

- Guan Tibshirani '19 : - multiclass
  - split data over classes $j=1,\dots,J$ → for x_t
  - apply standard or icp. → for y_t

$$\forall y_t : \quad P_{Z, X_t \mid Y_t = y_t}\left(X_t \in T_{y_t}(Z)\right) \geq 1-\alpha.$$



- $T(Z, x_t) = \{y_t : x_t \in \Gamma_{y_t}(Z)\}$

[similar to construction of confidence intervals based on tests]
"duality" between hypothesis testing & conf. intervals

by definition: $P_{Z,\, X_t \mid Y_t = y_t}\left(y_t \in T(Z, X_t) \right) \geq 1 - \alpha$

- same guarantee as before, diff^t construction

STAT 991 · Lec 8

Conformal prediction

- other topics:
  - between ICP & CP: cross-CP, jackknife+, sub-sampling
  - more general losses: ICP-type analysis via concentration inequalities

- papers to look at
  - Park et al '20: "PAC pred. sets" → good DeepNN examples.

Calibration

idea: if we predict a probability for an event, then that should be close to the actual probability of observing it.

Def: A - events in a space $\mathcal{X}$ (random event); X observation (outcome of r.v. X) e.g. weather today

e.g. - rain $\in \{0,1\}$ tomorrow
- level of rain $\in [0, \infty)$

- probability predictor / forecaster $p: \mathcal{X} \to [0,1]$. e.g.: human expert predicts prob. of rain. → neural net.

def: p is calibrated if $\forall q \in [0,1]$

$$P(A \mid p(X) = q) = q$$

Notes:- if $p(x)$ can take any value, in general "unattainable" need to relax

- in ML: $A = \{ f(x) = y \}$ (f: classifier; y: observed class)

~~$p(x) = $ top-k softmax scores (sum).~~

. What about pD. consider :

· output of typical classifier NN.

. $s(x) \in \Delta_K$ - K-dimensional simplex.

$s(x)_i \geq 0 \quad \sum_{i=1}^{K} s(x)_i = 1$

"predicted probability" of each class
(has nothing to do with probability)

· Prediction $f(x) \in \arg\max_i s(x)_i$

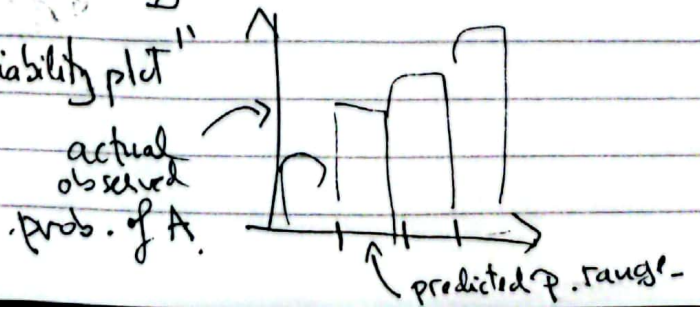
· confidence $p(x) = \max_i s(x)_i$.

History / Context : a tricky notion ; see Lichtenstein et al '1980.

· 1906 : Australian meteorologist Cooke
↳ 1958 : British statistician D.R. Cox [test several Bernoulli parameters]
↳ 1950 : US . meteorol. Brier → Brier score
(- 1960s- : psychology, social sciences → overconfidence
- 1980s- statistics : decision-theory, Bayesian, online, re-calibration
[Dawid, DeGroot, Fienberg, Foster, Vohra ...]
- 1962 : US. meteorol. R. Miller [test several batches of Bernoullis]
:
- 2012+ : Deep learning : → over-confidence.
↳ many approaches, mostly heuristic
↳ guarantees : e.g. for binning [Park et al '20]

. studied under many names : realism, external validity, realism of confidence, appropriateness of confidence, secondary validity
reliability

· Visually: "reliability plot"



A few vignettes/perspectives

* Testing : $P(A \mid p(x) = q) = q \neq q$.

- Suppose we have a completely unstructured, $p(\cdot)$ & a set of values q_i
- consider the RVs $B_i = \mathbb{1}_A \mid p(x) = q_i$.
- Then, it reduces to:
  - $B_i \sim \text{Bernoulli}(p_i)$, $p_i \in [0,1]$
    & some dependence structure.
  - Calibration null : $p_i = q_i \; \forall i$
- simplest case, B_i - indep.
- ~~in this form, is a multiple hypothesis testing problem~~

- Miller'62 : $X_{ij} \sim \text{Ber}(p_{ij}) \quad \forall i$, indep across j.
  test $H_0 : p_{ij} = q_{ij} \; \forall i,j$.
  H_i : fixed i, $\forall j$

to test H_i:
- Chi-squared test statistic : $T_i = \dfrac{\left(\sum_j X_{ij} - \sum_j p_{ij}\right)^2}{\sum_j p_{ij}(1-p_{ij})}$
- by CLT, expect that as $\sum_j p_{ij}(1-p_{ij}) \to \infty$.
  $T_i \Rightarrow_d \chi^2(1)$
- To test H_0 : use $T^\top \hat{\Sigma}^{-1} T$, $\chi^2(1)$ ~~test~~ null dist.
  - $\hat{\Sigma}$: estimated correlation matrix of entries of T
  - [when works ?]

* Scoring rules

- $q \in [0,1]$ - prob. pred $\Rightarrow$; Q : r.v. of probability predictions
- $r(p) = P(A \mid P(x) = p)$
- Brier score: $\mathbb{E}(p - I(A=1))^2$ (*) (over joint dist. of prediction P, event A)
  - p: predicted prob.
  - $I(A=1)$: true event.

$$= \mathbb{E}\{(p - r(p))^2 + r(p)(1 - r(p))\}$$

(bias-var. decomp)

- $(p - r(p))^2$: predicted prob. vs true prob. — calibration
- $r(p)(1 - r(p))$: sharpness
  - "sharp" when this is small, $r(p) \approx 0, 1$.

- more generally, for a K-class prediction problem,
  - let
    - $s : \{1..K\} \times \Delta_K \to \mathbb{R}$ be a scoring rule
    - $s(y, p)$ - Loss when obs. class is $y \in \{1..K\}$ & predict p
  - let q - true pmf of y, & TIBR [again stat. decision theory]

$$R(p, q) = \mathbb{E}_{y \sim q}\, s(y, p)$$

e.g., Brier score: rule $s(y, p) = \|e_y - p\|^2 = \left\| \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \text{at } y - p \right\|^2$

in binary case, enough to look at one coordinate.

square $(y - p)^2$, as before (*) : $y = I(A=1)$

gen: $\mathbb{E}_{y \sim q} \|e_y - p\|^2 = \sum_y q(y)\left[(1 - p(y))^2 + \|p\|^2 - p(y)^2\right] = \|p\|^2 - 2\langle p, q\rangle + 1$

$1 - 2p(y) + \|p\|^2 = \|p - q\|^2 - \|q\|^2 + 1$

def: proper scoring rule: $\forall q \in \Delta_K : q \in \arg\min_{p \in \Delta_K} R(p;q)$

def: Strictly —"— q is unique minimizer.

e.g., Brier score is a strictly proper scoring rule.

$$\arg\min_p \left\{ \|p-q\|^2 - \|q\|^2 + 1 \right\} = p.$$

S-entropy: $R(p,p) = \mathbb{E}_{y \sim p} s(y,p)$.

e.g. Br. sc: $1 - \|p\|^2$

More examples

- logarithm
  $s(y,p) = -\log p(y)$.

$$R(p,q) = \mathbb{E}_{y \sim q} -\log p(y) = \underbrace{\mathbb{E}_{y \sim q} \log\left(\frac{q(y)}{p(y)}\right)}_{KL(q,p)\ \text{KL div}} - \underbrace{\mathbb{E}_{y \sim q} \log\left(q(y)\right)}_{H(q)\ \text{Shannon entropy}}$$

is a str. prop. sc. rule.

- normalized prob.
  spherical sc. rule: $s(y,p) = -\frac{p(y)}{\|p\|}$ · strictly proper.

Stat ggl. Calib: (Lec 8)

$$R(p,q) = \mathbb{E}_{y\sim q} S(y,p)$$

$$= \underbrace{\mathbb{E}_{y\sim q}\{S(y,p) - S(y,q)\}}_{\geq 0 \text{ for proper Bregman rule.}} + \underbrace{\mathbb{E}_{y\sim q} S(y,q)}_{R(q,q)\,\cdot\,\text{entropy}}$$

> 0, unless $p=q$ for s.p.l.r.

- this is just loss fn.
- stat. decision problem: $\min_{f\in\mathcal{F}} R(p_{f(x)}, q(y))$ over $(x,y)\sim\mathcal{D}$

  - f: prob dist w/ softmax scores f. (could say $\min_{p\in\mathcal{F}} \dots p(x) \dots$)
  - q: use 1-hot encoding of y.

- min pred risk pushes us to 0-1. (over-conf)
  - ↳ especially if overfit
  - ↳ esp. if use log loss $\mathbb{E}_{y\sim q} -\log p(y)$.

- Comm. to pred sets / ↑

	pred st. clf.	here
what?	$x \to C(x) \subseteq \mathcal{Y}$	$x \to q(x) \in \Delta_{\lvert\mathcal{Y}\rvert-1}$ prob. dist. over $\mathcal{Y}$
aim:	$y \in C(x)$	$p(y\mid x) \approx q(x)$

* Finite-sample guarantees for binning.

- Suppose (A_i, P) have a joint distribution
  - A_i: event. P: prob. pred.
  - e.g. ML: $(\mathbb{1}\{Y=y\}, P(x))$, (X, Y) data
- Let $B_j = [\cdot, \cdot]$ be a partition of $[0,1]$ (bins). $j = 1, \dots, K$
  - let $p_j = P(A \mid P \in B_j)$
  - ~~Can do~~.
  - Given a sample $Z = \{(A_i, P_i),\ i = 1..n\}$
    - achieve conditional coverage for (p_j), $j \in [K]$:
      by constructing CI (e.g. Clopper-Pearson)
      using $Z_j = \{(A_i, P_i) : P_i \in B_j\}$
      $$P(p_j \in CI(Z_j) \mid P \in B_j) \geq 1-\alpha.$$
    - Can also ensure coverage of discretization
      # of $\tilde{P}$: equal to p_j w.p. $P(P \in B_j)$ when $P \in B_j$.
      ~~train~~
  - e.g. ML: - $b(x)$ - bin of x
    - $c(x) = P_{(X,Y)}(f(X) = Y \mid P(X) \in B_{b(x)})$ - per-bin accuracy
    - guarantee $P_X(c(x) \in CI(x)) \geq 1-\alpha$.

- Summary: - calibration
  - perspectives: - testing
    - scoring rules
    - CIs after binning

others: re-calibration.