validation of SBI (cont'd)

- given test data $\{\hat{X}_i \sim p(X)\}_{i=1}^{N}$ or $\{\hat{X}_i = p(X \mid Y = \text{fixed})\}_{i=1}^{N}$

  $\{X_i^* \sim p^*(X)\}_{i=1}^{N'}$ or $\{X_i^* = p^*(X \mid Y = \text{fixed})\}_{i=1}^{N'}$

case (1) $N' \approx N \gg 1 \Rightarrow$ compare the distribution of the samples $\{\hat{X}\}, \{X^*\}$ using MMD or FID or density/coverage

case (2) $N' = 0$ (no ground truth) $\Rightarrow$ compare diversity of different approximations via Vendi score
[if $N' \gg 1$, also compare against GT Vendi score]

case (3) $N' = 1$ important case in practice

- in SBI, create GT via forward simulation, $\{(Y_i \sim p^S(Y), X_i = \phi(Y_i, \varepsilon))\}$

  $\Rightarrow Y_1$ is a GT example for $p(Y \mid X = X_1)$ with $N' = 1$

  fixed

- weather forecasts: $p(Y = \text{rain tomorrow} \mid X = \text{weather up to now})$

"80% rain probability"

$Y_1 = p^*(Y = \text{rain} \mid X = \text{weather so far}) \hat{=}$ actual weather

is a GT sample with $N' = 1$

- idea: "calibration" merge instances with same predicted confidence in a joint test set among all days with "80% rain prob."

it should have rained in 80% of the cases

applied to classification: $p(Y = k \mid x)$ is 80%, and answer is right

80% of the time $\Rightarrow$ "well calibrated"

$> 80\%$ — " — $\Rightarrow$ "under confident"

$< 80\%$ — " — $\Rightarrow$ "overconfident"

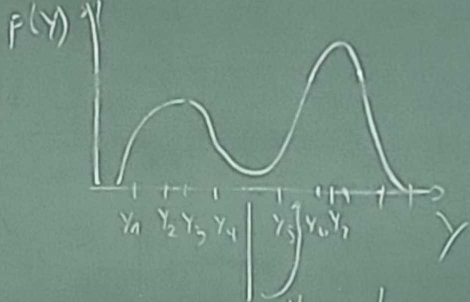
- realization: sort predicted sample (consider the inverse problem $p(Y \mid X)$ with $Y \in \mathbb{R}$

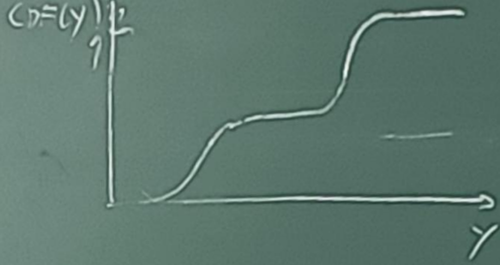
$\{Y_i = \hat{Y}_i \sim p(Y \mid X = \text{fixed})\}_{i=1}^{N} \cup \{Y_0 = Y^* \sim p^*(Y \mid X)\}$

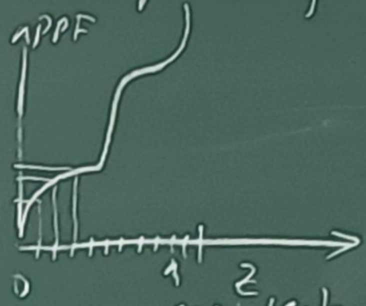
sort into $(Y_{[0]}, \ldots, Y_{[N]})$ ← sorted order

$Y_0 \rightsquigarrow Y_{[k]}$ calibration $\hat{=}$ $k \sim \text{uniform}(0, N)$

- let $p(Y)$ some 1-D prob, $CDF(Y)$ corresponding cumulative dist. f.







equidistant quantiles $\frac{1}{(N+1)}, \dots, \frac{N}{(N+1)}$

all intervals have same probability mass

$\Rightarrow$ the interval, where a new sample lands is uniformly distributed

- if quantiles are unknown, a sorted sample $Y_{[n]} \sim p(Y)$ is a good approx. probability mass in every interval is approximately equal

alg. ① given GT Y_i^*, sample from model $\{\hat{Y}_k \sim p(Y \mid X_i)\}_{k=1}^{M}$

② sort the joined set $\{Y_{i0} = Y^*, \hat{Y}_{i1}, \dots, \hat{Y}_{iM}\} \Rightarrow (Y_{i[0]}, \dots, Y_{i[M]})$

③ $c_i = \frac{[k]}{M}$ ← index of Y^* in sorted order $\quad c \in [0, 1]$

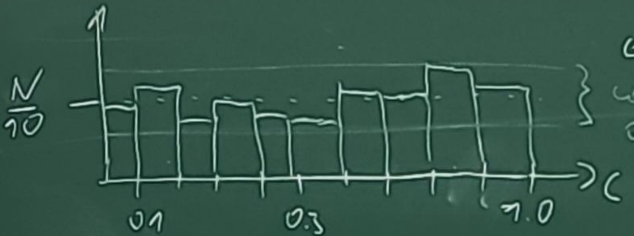
repeat this for many GT instances (X_i, Y_i) $\Rightarrow$ sample $\{c_i\}_{i=1}^{N}$

- evaluate $\{C_1\}$ via histogram test: if model is calibrated (null hypothesis)

$$C_1 \sim \text{Binomial}\left(N, p = \frac{1}{L}\right)$$

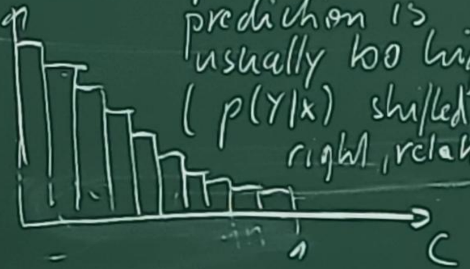
L ← # of bins

$L = 10$



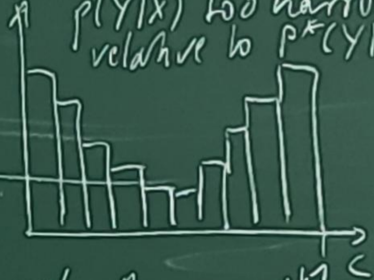
well calibrated

- typical histograms when uncalibrated



prediction is usually too high ($p(Y|X)$ shifted to the right, relative to $p^*(Y|X)$)

overconfidence
$p(Y|X)$ too narrow relative to $p^*(Y|X)$



underconfidence
$p(Y|X)$ too wide

- caveat: "well-calibrated" does not imply "accurate"
  a bad model that knows that it is bad, is still calibrated

toy example: t time, sample mean on day t $\mu_t \sim N(0,1)$, outcome $Y_t \sim N(\mu_t, 1)$

$\Rightarrow$ marginal distribution $Y_t \sim p^*(Y_t) \sim N(0, \sigma^2 = 2)$

if model predicts $Y_t \sim p(Y) \approx p^*(Y_t)$, it is calibrated, but predicting $X_t \sim p(X|\mu_t) = p^*(X|\mu_t)$ is better

both are calibrated but variances differ

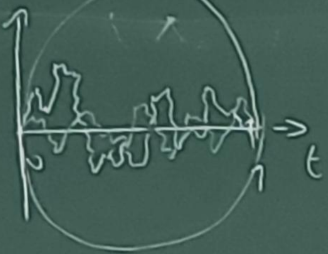
• alternative visualization via empirical CDF of $\{c_i\}_{i=1}^{N}$ $\quad c_i \in [0,1]$

$$CDF(t) = \frac{1}{N}\sum_{i=1}^{N} 1[c_i \le t], \quad \text{define} \quad r(t) = CDF(t) - t$$

$[CDF^*(t)]$

$t \in [0,1]$

draw $r(t)$

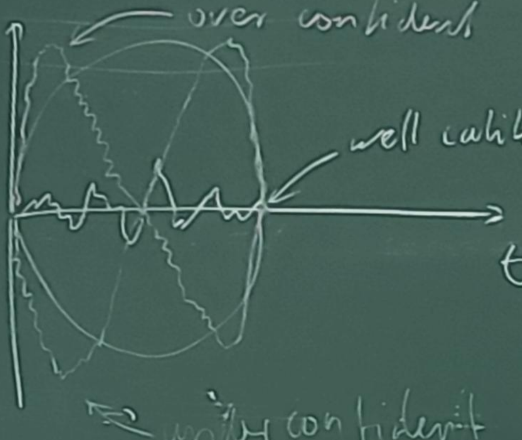


- one curve per element of $Y \in \mathbb{R}^D$
- if null hypothesis "well calibrated" is true curves are "Brownian bridges" ≙ random walk with fixed start and end point
- confidence intervals of Brownian bridge are analytically known $\quad \text{stddev}(t) = \sqrt{\frac{1}{N} t(1-t)}$

(in practice, draw $\frac{\sqrt{N}}{2} r(t)$ with $\text{stddev}(t) = \sqrt{t(1-t)}$

$2 \leftarrow$ implies 95% confidence region

typical calibration errors



joint calibration checks: instead of testing $Y_i = \{Y_{i1}, \dots, Y_{i10}\}$ one feature at a time, check the entire vector

$\Rightarrow$ reduce problem to 1-D via "energy" or "surprisal" distribution

$$e(Y \mid X_1) = -\log p(Y \mid X_1)$$

(energy if $p(Y) = \frac{1}{Z} \exp(-e(Y)/kT)$ (Gibbs distribution)

$$-\log p(Y) = \frac{e(Y)}{kT} + \log Z \;)$$

define energy distribution via "how often do we see $e(Y) = E$"

$$\boxed{p(E \mid X_1) = \int_Y \delta(e(Y \mid X_1) - E)\, p(Y \mid X_1)\, dY}$$

example: if $Y \sim \mathcal{N}(0, \mathbb{I}_D)$, then $-\log p(Y) = \frac{\|Y\|^2}{2} + \text{const}$

$$p(E) = p\left(\frac{\|Y\|^2}{2} = E\right) = 2\chi^2_D(2E)$$

$E = \text{const} = \{Y \mid \frac{\|Y\|^2}{2} = E$, i.e. surface of a sphere with radius $\sqrt{2E}\}$

alg: ① for $i = 1, \dots, N$

(a) $\{\hat{Y}_{ik} \sim p(Y \mid X_i)\}_{k=1}^{M}$ and Y_i^* the GT from test set

(b) $e_{ik} = -\log(\hat{Y}_{ik} \mid X_i)$, $e_{i0} = -\log(Y_i^* \mid X_i)$

(c) sort into $(e_{i[0]}, \dots, e_{i[M]})$, let $[k]$ index of Y_i^* after sorting

(d) define $C_i = \frac{[k]}{M}$

② use histogram or CDF methods to analyse distribution of $\{C_i\}_{i=1}^{N}$

Tilman Gneiting

weather forecast $p(Y = \text{rain} \mid X = \text{weather so far, date})$

model ignores weather so far ⇒ makes predictions only according to date

prediction is calibrated, if $p(Y \mid \text{date})$ is the correct statistics for "date"

but much less accurate if it exploited weather correlation

between subsequent days