UNCERTAINTY QUANTIFICATION

DEEL MASTERCLASS

JOSEBA DALMAU

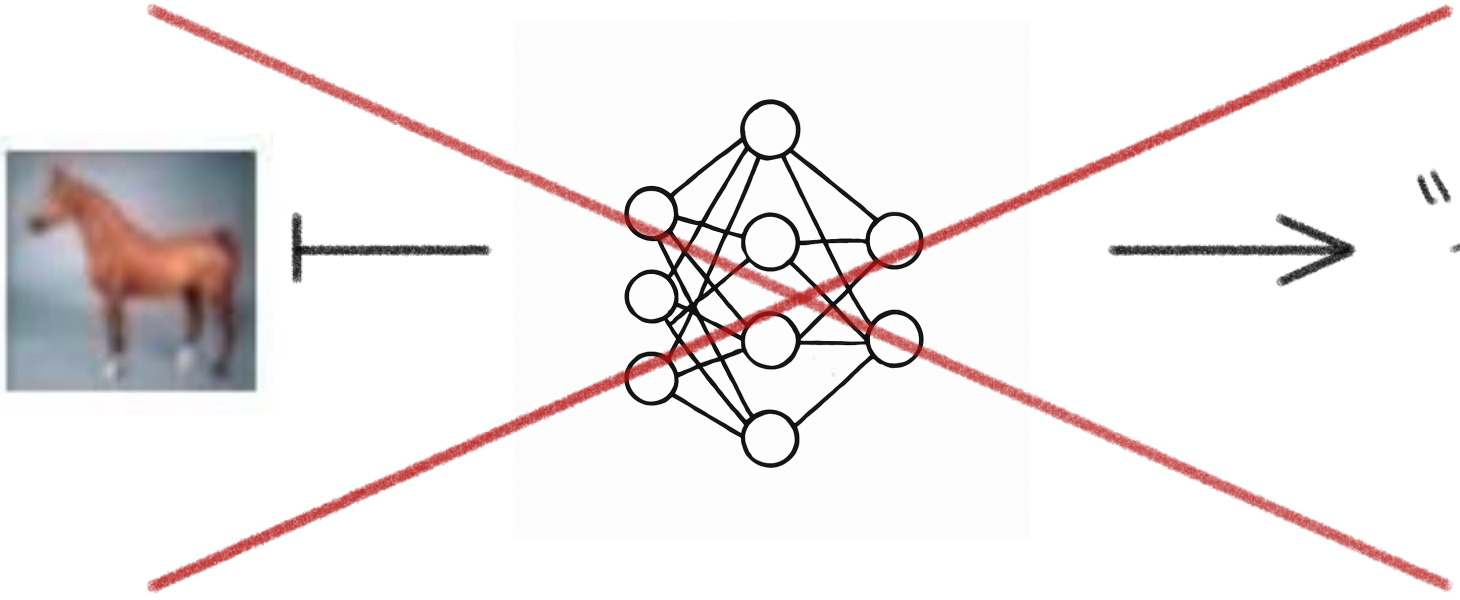
DEEL
Dependable, Explainable & Embeddable Learning

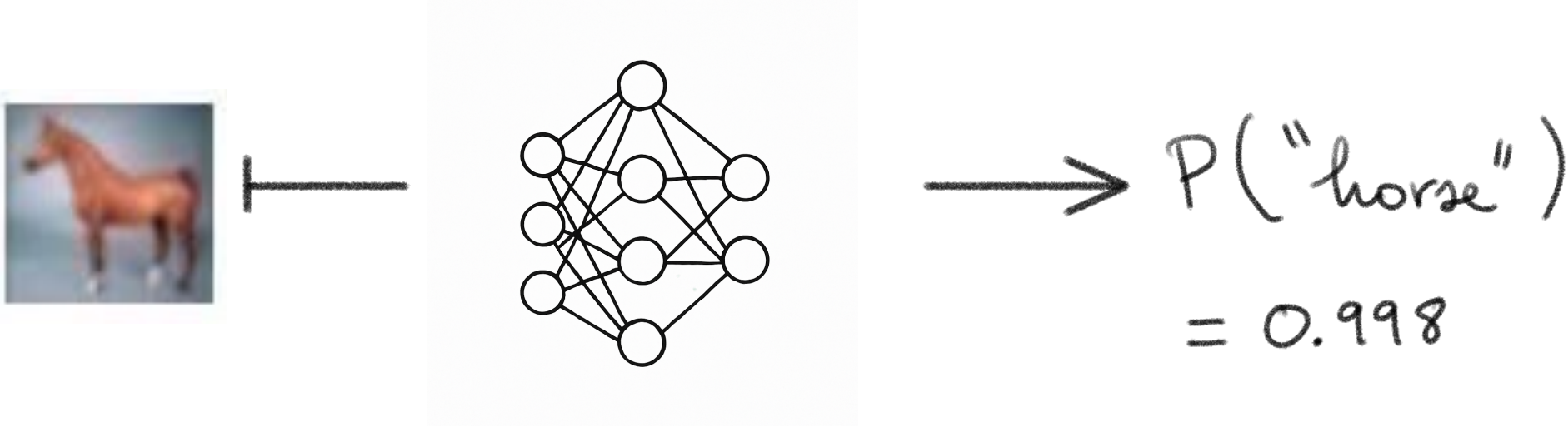
IRT SAINT EXUPÉRY

Uncertainty Quantification In Deep Learning

- Can we trust the predictions of our models ?
- Under what circumstances ?
- To what extent ?

NNs ARE ILL-CALIBRATED





NNs ARE ILL-CALIBRATED




P("horse") = 0.998

P("horse") = 0.998

P("horse") = 0.998

P("horse") = 0.998

Proportion of correct answers IS NOT 0.998

Classical UQ Techniques

- Bayesian methods
- Ensembling
- Dropout
- ...

Non Post-hoc

Without Guarantees

Conformal Prediction:

Objective and Guarantee

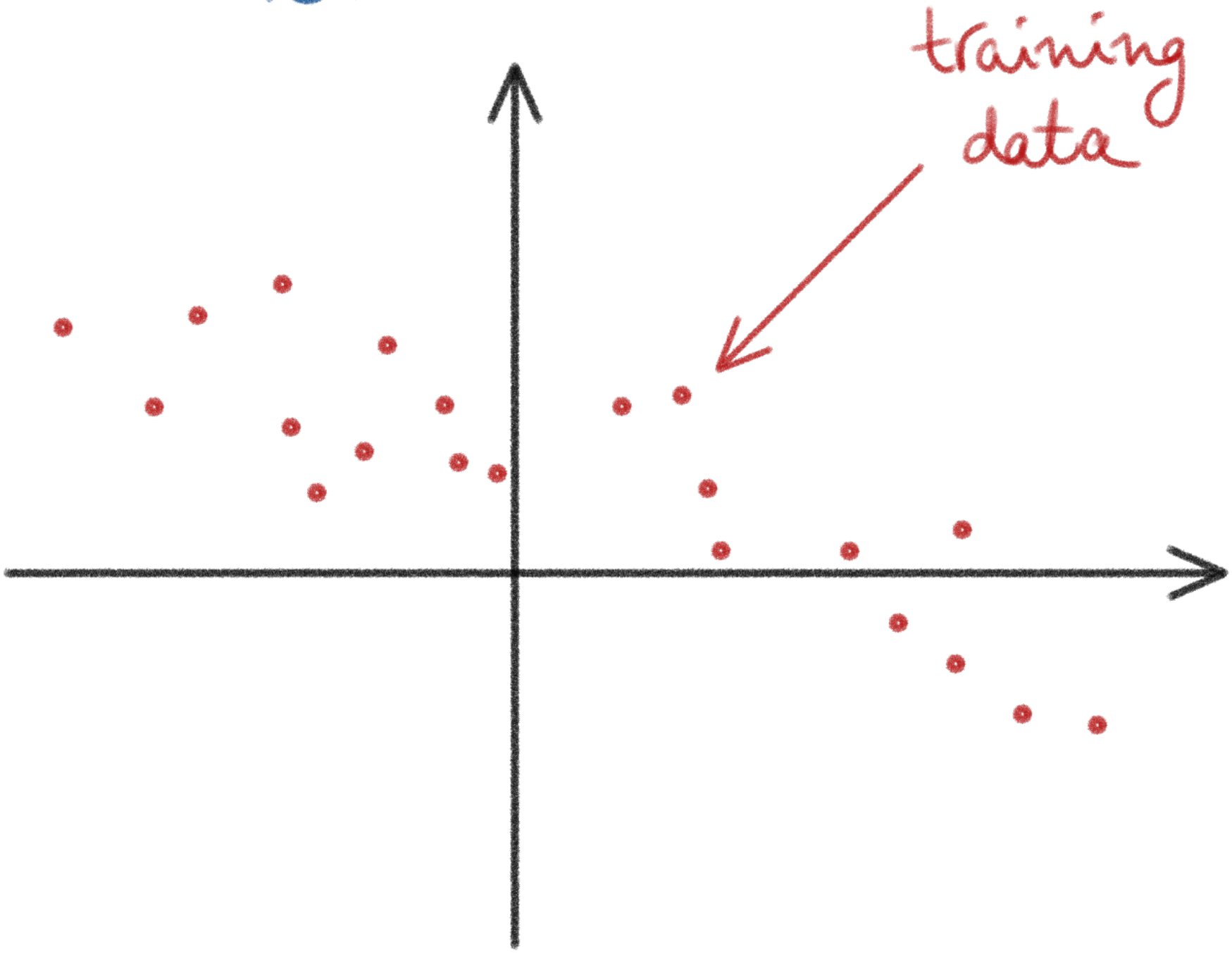
Given: a predictor $\hat{f} : \mathcal{X} \longrightarrow \mathcal{Y}$
and a nominal error rate α.

Build: $\hat{C}_\alpha : \mathcal{X} \longrightarrow \mathcal{P}(\mathcal{Y})$

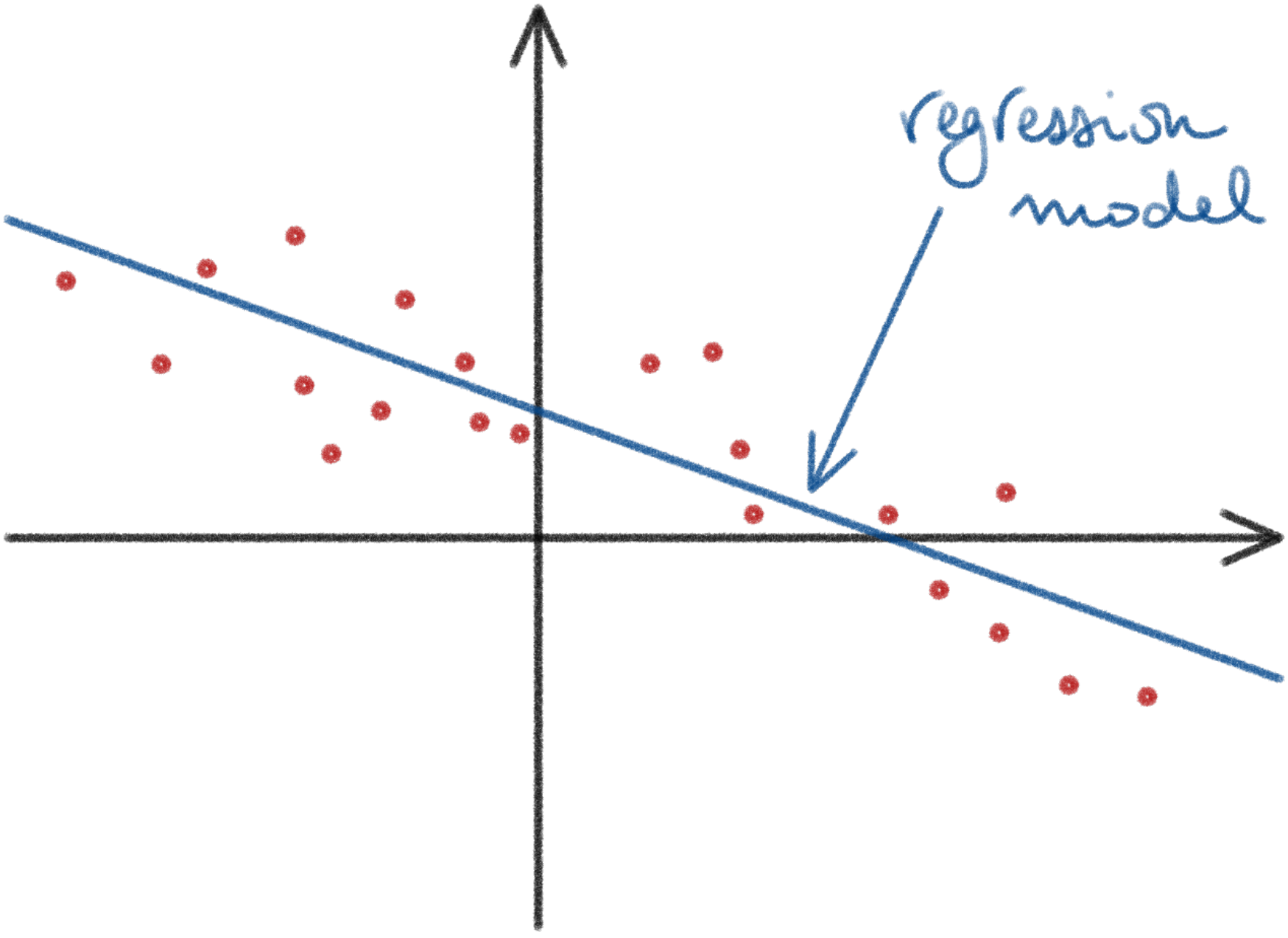
With the following guarantee:

$$\mathbb{P}\left(Y_{test} \in \hat{C}_\alpha(X_{test})\right) \geq 1 - \alpha$$

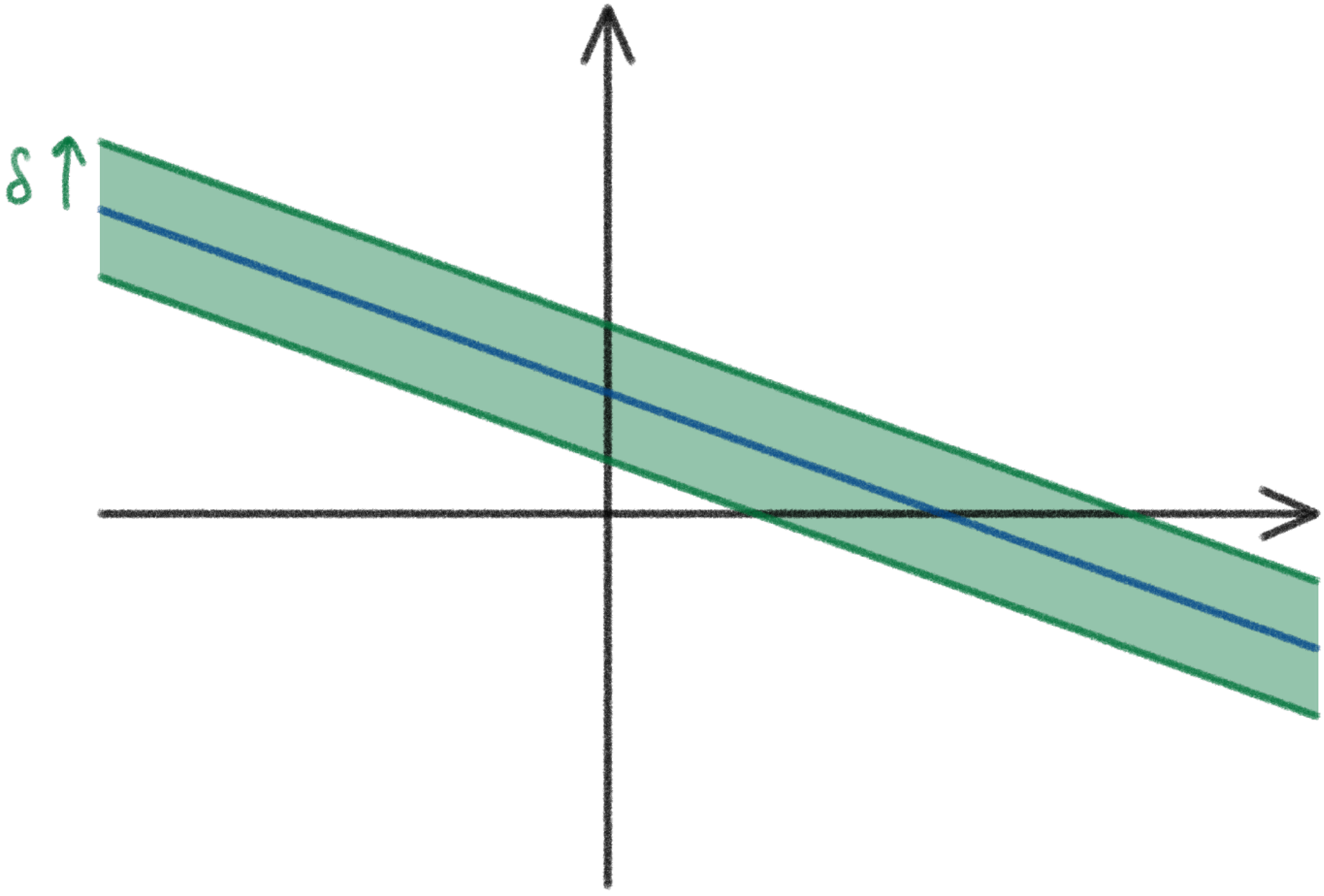
Intuition



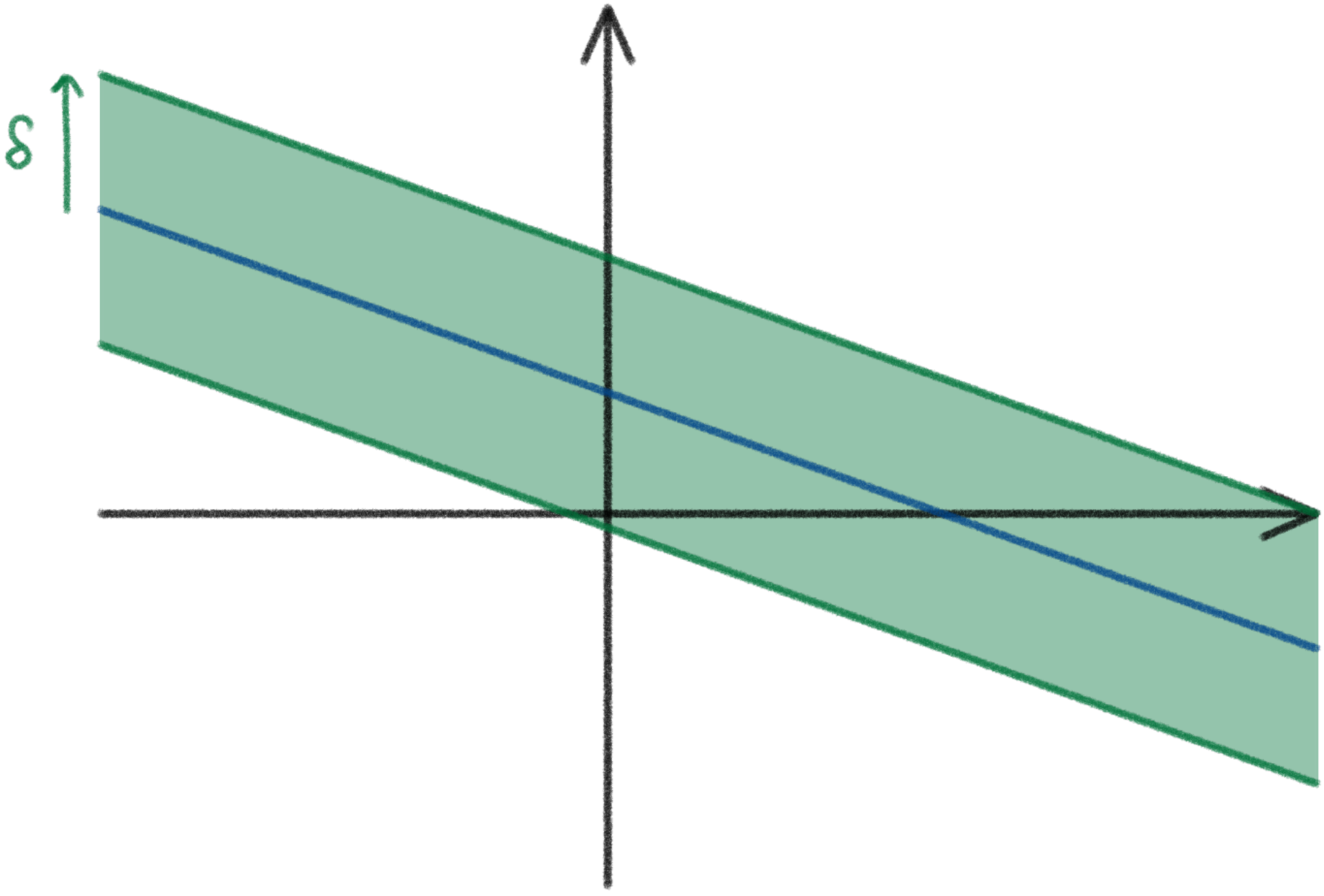
Intuition



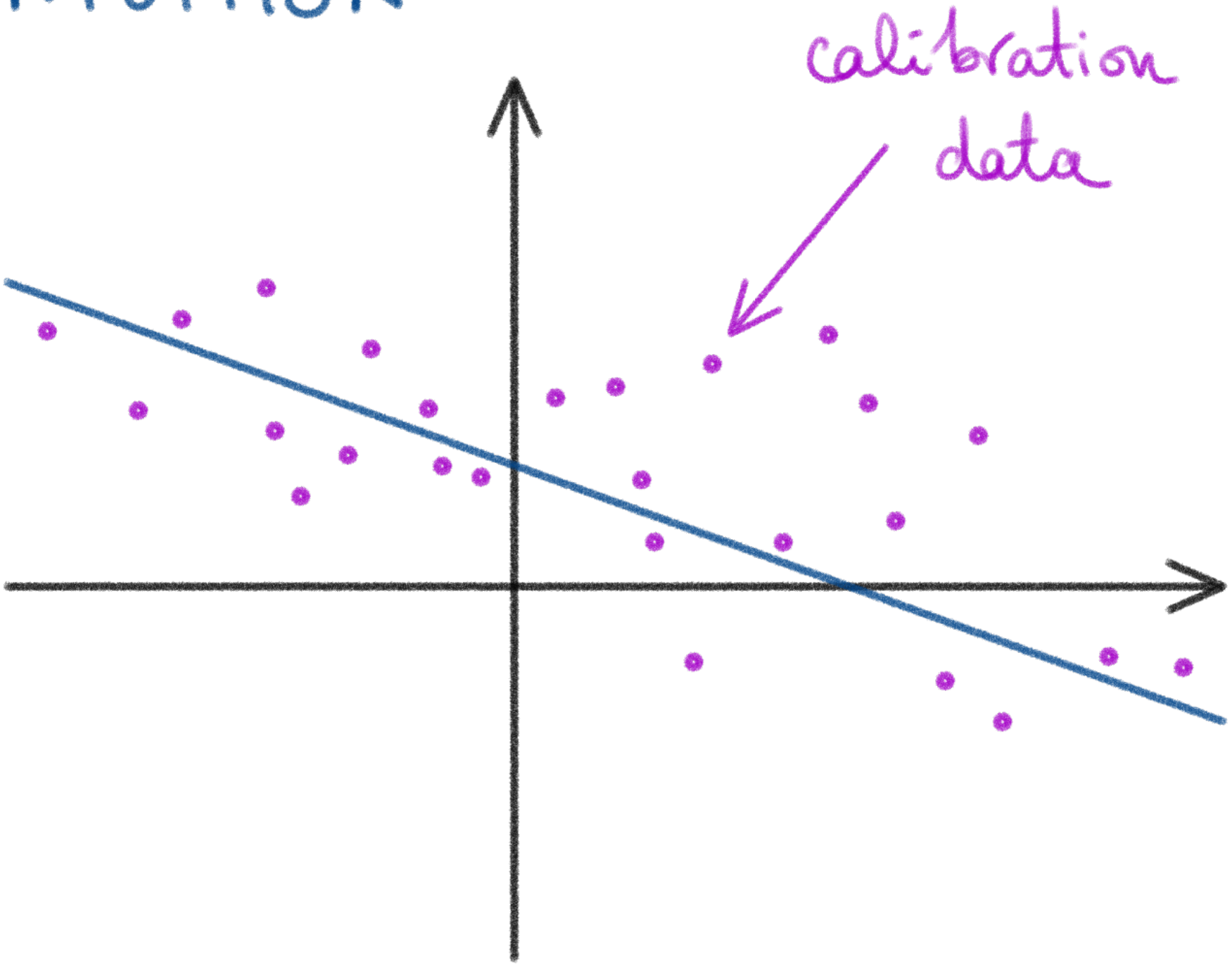
Intuition

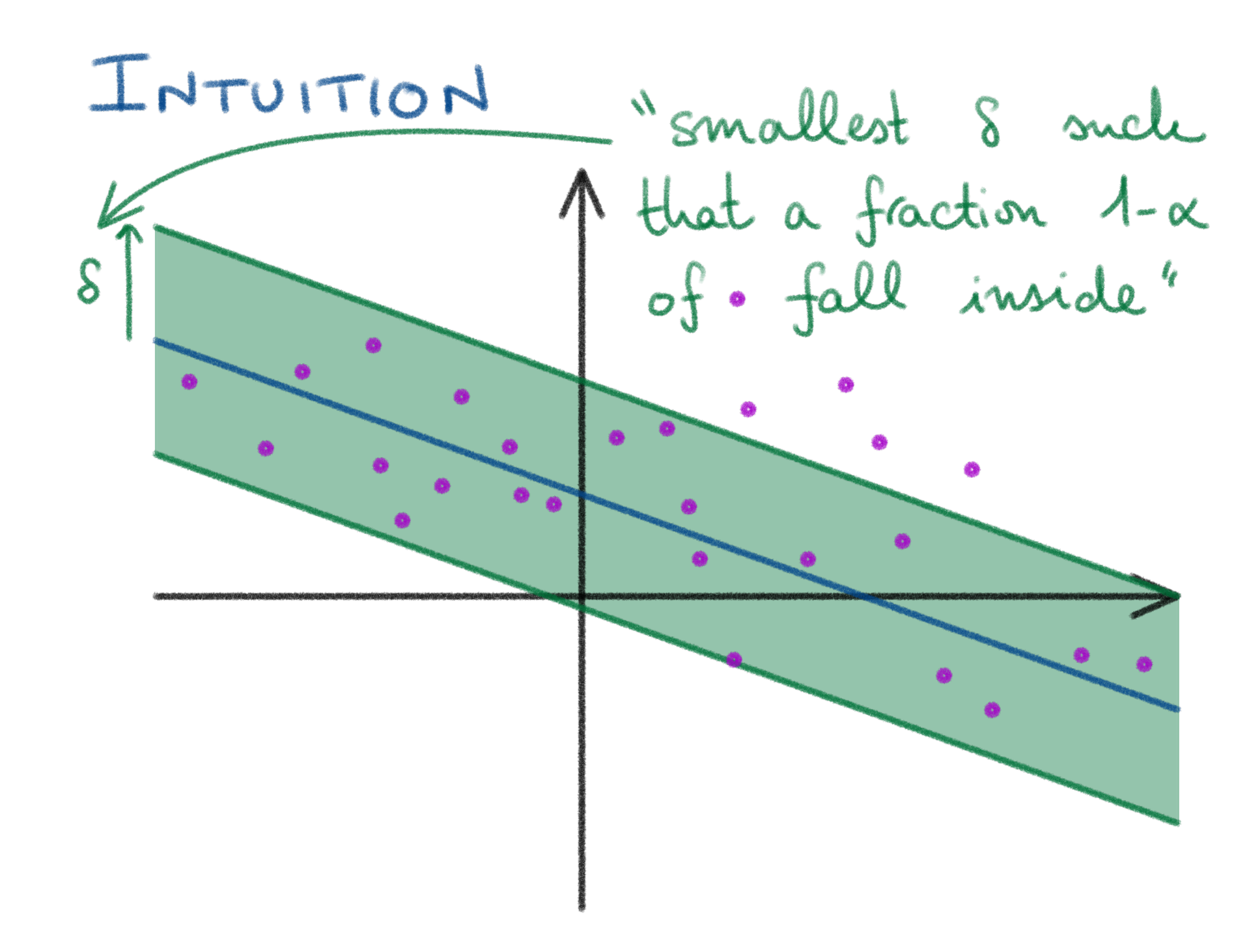


Intuition



Intuition




INTUITION
"smallest δ such that a fraction 1-α of • fall inside"
δ

Outline

→ Conformal Regression

- Theorem
- Jackknife+ / CV+
- CQR

→ Conformal Classification

- LAC
- APS / RAPS

Conformal Regression

Given: $\hat{f} : \mathcal{X} \to \mathcal{Y}$ and α.

Build: $\hat{C}_\alpha : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$

With the following guarantee:

$$\mathbb{P}\left(Y_{test} \in \hat{C}_\alpha(X_{test})\right) \geq 1 - \alpha$$

Conformal Regression

Given: $\hat{f} : \mathcal{X} \to \mathbb{R}$ and α.

Build: $\hat{C}_\alpha : \mathcal{X} \to \mathcal{P}(\mathbb{R})$

usually an interval!

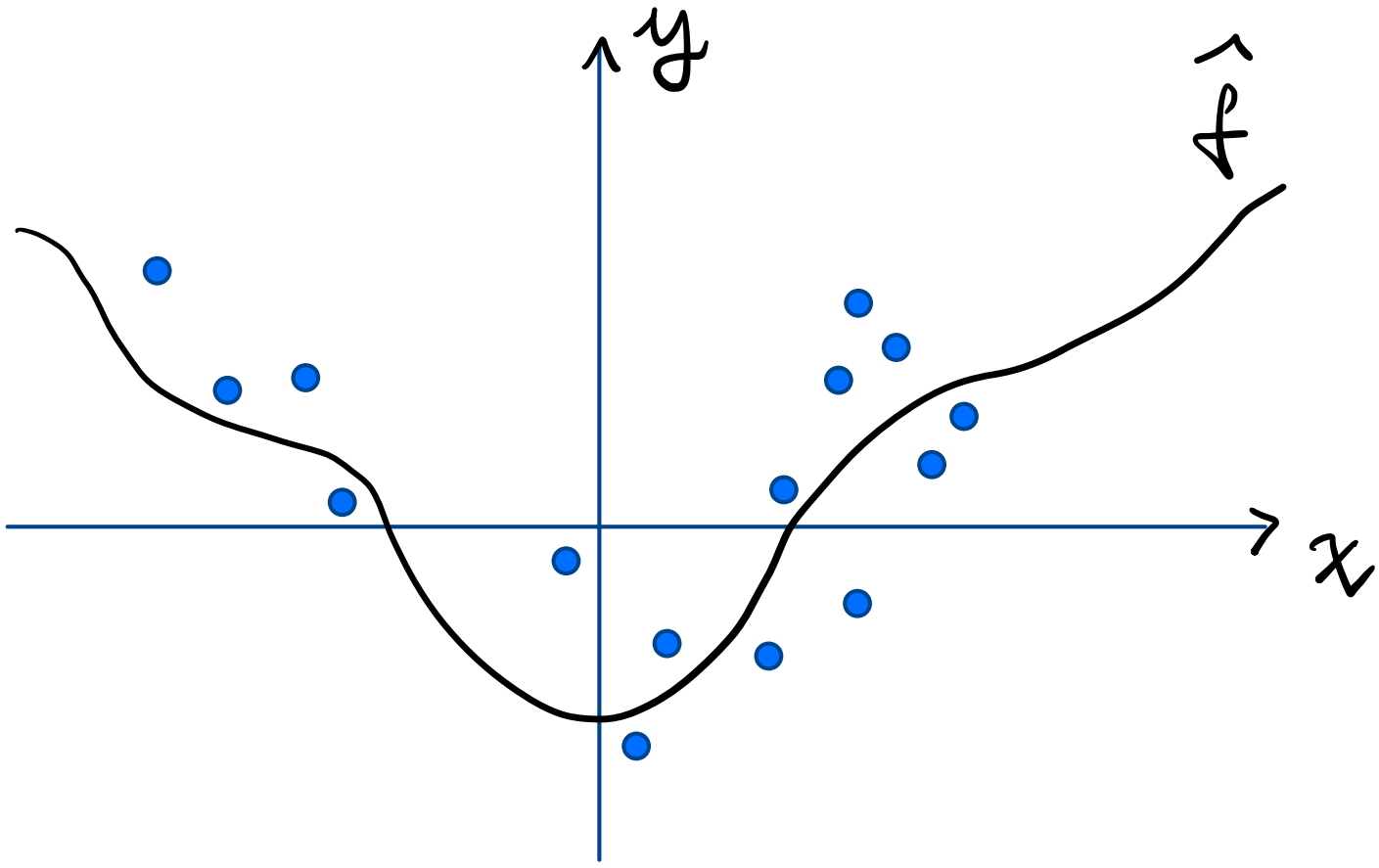
With the following guarantee:

$$\mathbb{P}\left(Y_{test} \in \hat{C}_\alpha(X_{test})\right) \geq 1 - \alpha$$

Calibration

We use a calibration set

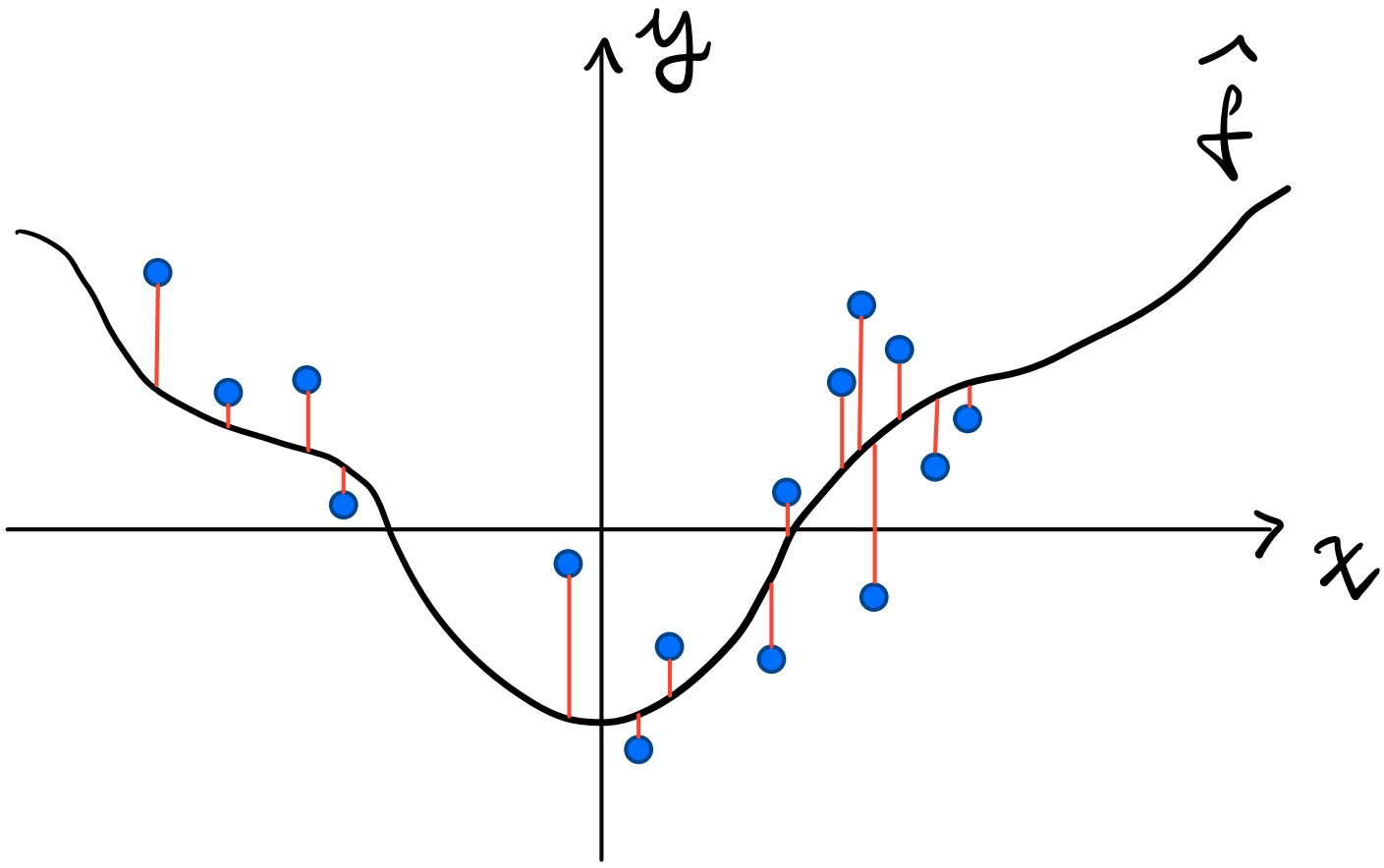
$$D_{calib} = \{(x_i, y_i)\}_{i=1}^{n}$$



CALIBRATION

We measure the scores (errors)

$$s_i = |y_i - \hat{f}(x_i)|$$



Calibration

We compute:

$\hat{s}_\alpha :=$ the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th

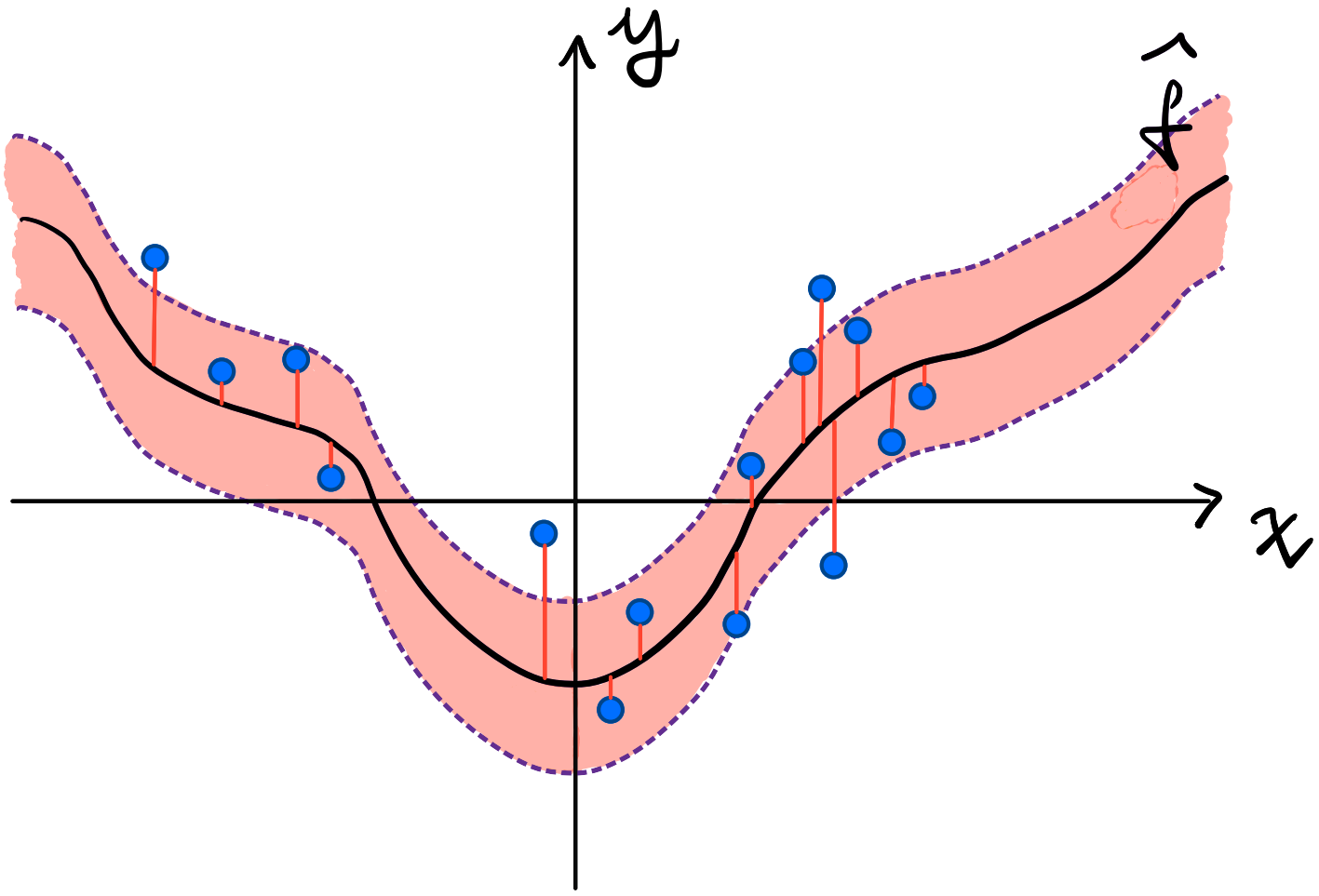
quantile of the scores $s_1, \ldots, s_n$

i.e. the $\lceil (n+1)(1-\alpha) \rceil$-th

smallest score.

CALIBRATION

We predict $\hat{C}_\alpha(x) = [\hat{f}(x) - \delta_\alpha, \hat{f}(x) + \delta_\alpha]$



Exercise

1. Write down the pseudocode of the Split Conformal Regression algorithm

2. Should we add any constraints on the nominal error rate α?

Hint: there may be a problem when choosing the "$\lceil (n+1)(1-\alpha) \rceil$-th smallest score"

GUARANTEE (Vovk et al 1999)

Theorem.- Given a calibration set $\{(X_i, Y_i)\}_{i=1}^{n}$ and a test point (X_{n+1}, Y_{n+1}), if the scores $S_i = |Y_i - \hat{f}(X_i)|$ are exchangeable, then the interval

$$\hat{C}_\alpha(X_{n+1}) = \left[\hat{f}(X_{n+1}) - \delta_\alpha \,,\; \hat{f}(X_{n+1}) + \delta_\alpha\right]$$

satisfies:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1-\alpha$$

EXCHANGEABILITY

The random variables $(X_1, X_2, \dots, X_n)$ are exchangeable if:

For any $\pi \in \{$ permutations of $\{1, \dots, n\}\}$

$$X_{\pi(1)}, \dots, X_{\pi(n)} \overset{\text{law}}{=} X_1, \dots, X_n$$

Exercise.- Give an example of an array of r.v.s that are exchangeable but not i.i.d.

EXCHANGEABILITY

The random variables $(X_1, X_2, \ldots, X_n)$ are exchangeable if:

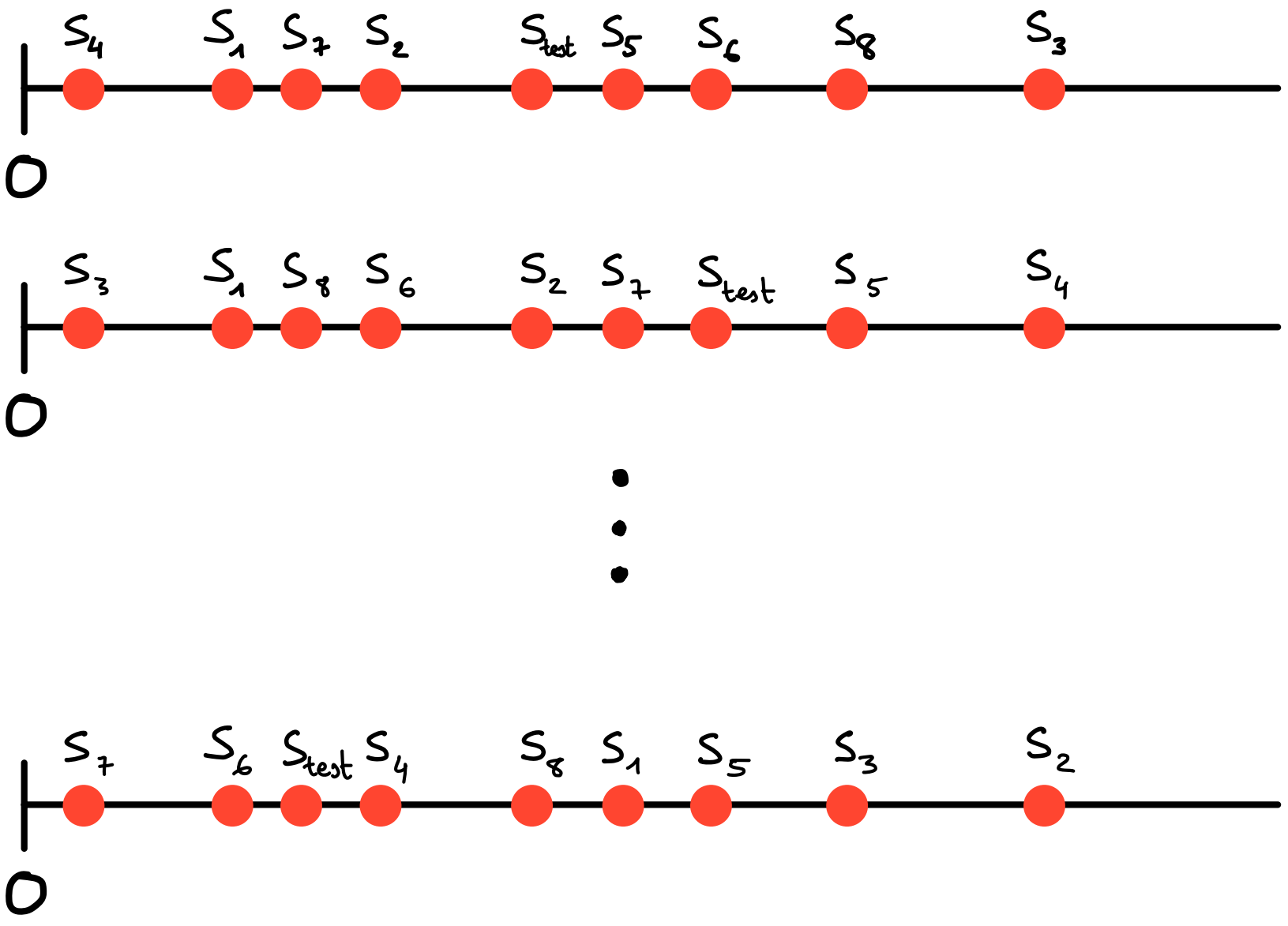
For any $\pi \in \{$ permutations of $\{1, \ldots, n\}\}$

$$X_{\pi(1)}, \ldots, X_{\pi(n)} \overset{\text{law}}{=} X_1, \ldots, X_n$$

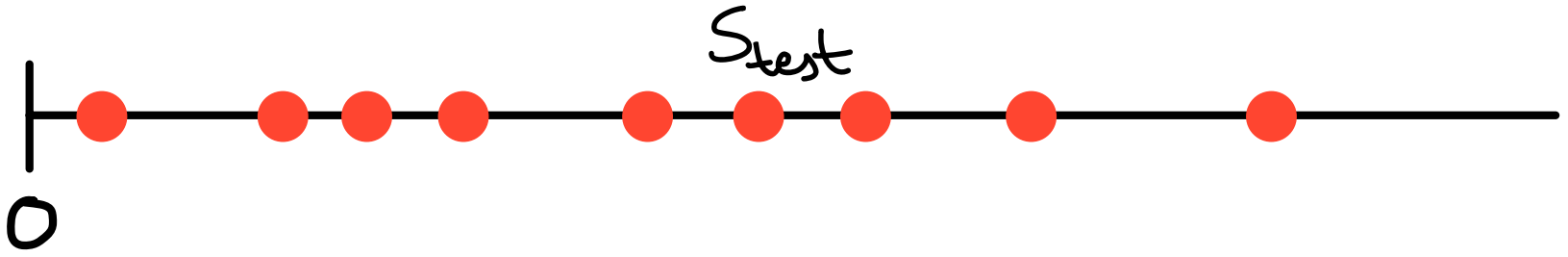
Example: $Z_1, \ldots, Z_n$ i.i.d. $\Rightarrow X_i = \dfrac{Z_i}{\sum_{j=1}^{n} |Z_j|}$

exchangeable

Proof


S_4 S_1 S_7 S_2 S_{test} S_5 S_6 S_8 S_3
0
S_3 S_1 S_8 S_6 S_2 S_7 S_{test} S_5 S_4
0
S_7 S_6 S_{test} S_4 S_8 S_1 S_5 S_3 S_2
0

Proof



$$\mathbb{P}(\text{Rank of } S_{test} = k)$$

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

$$\Rightarrow \text{Rank of } S_{test} \sim \text{Unif}(\{1, \ldots, n+1\})$$

Proof

Rank of $S_{test} \sim \text{Unif}(\{1, \dots, n+1\})$

$$\Rightarrow \mathbb{P}\left(\begin{matrix}\text{Rank of} \\ S_{test}\end{matrix} \leq k\right) = \frac{k}{n+1}$$

We choose the smallest k s.t.

$$\frac{k}{n+1} \geq 1-\alpha \quad \text{i.e.} \quad k^* = \lceil (n+1)(1-\alpha) \rceil$$

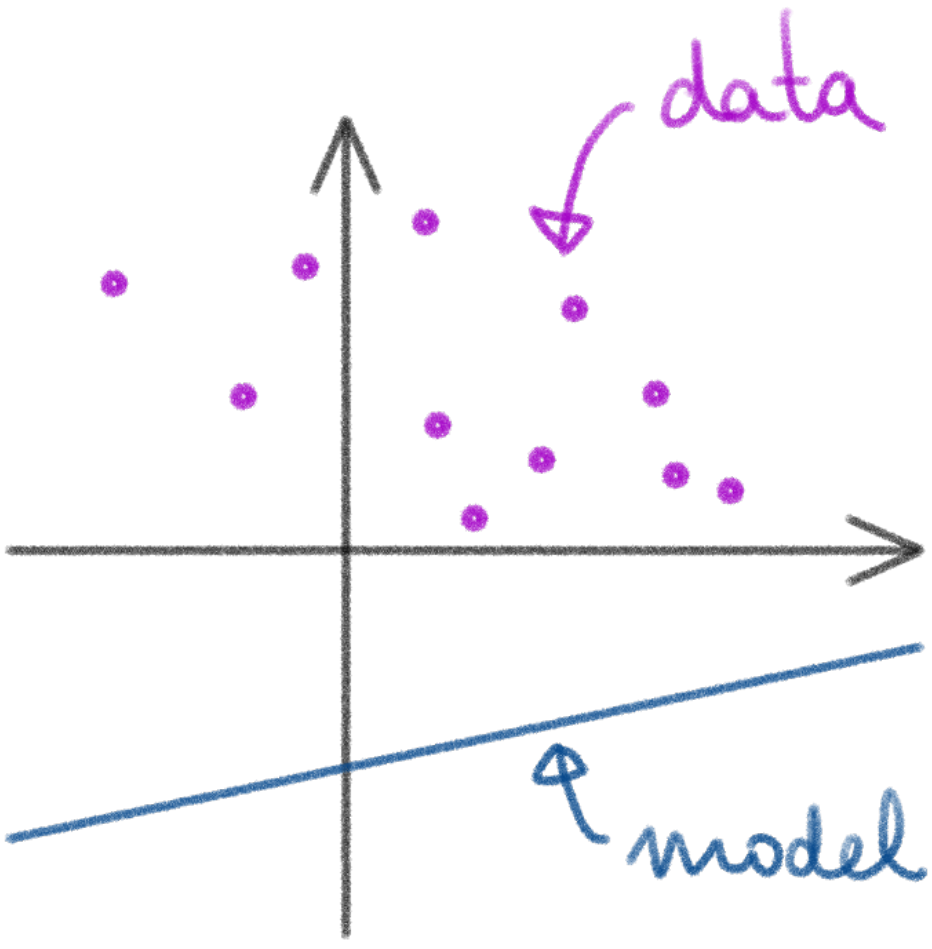
PROOF

Then : $\mathbb{P}\left(\text{Rank of } S_{test} \leq K^*\right) \geq 1-\alpha$

Rank of $S_{test} \leq K^*$

$\Longleftrightarrow$ $S_{test} \leq$ the K^*-th smallest score $S_1, \dots, S_n$

Too Good To Be True?



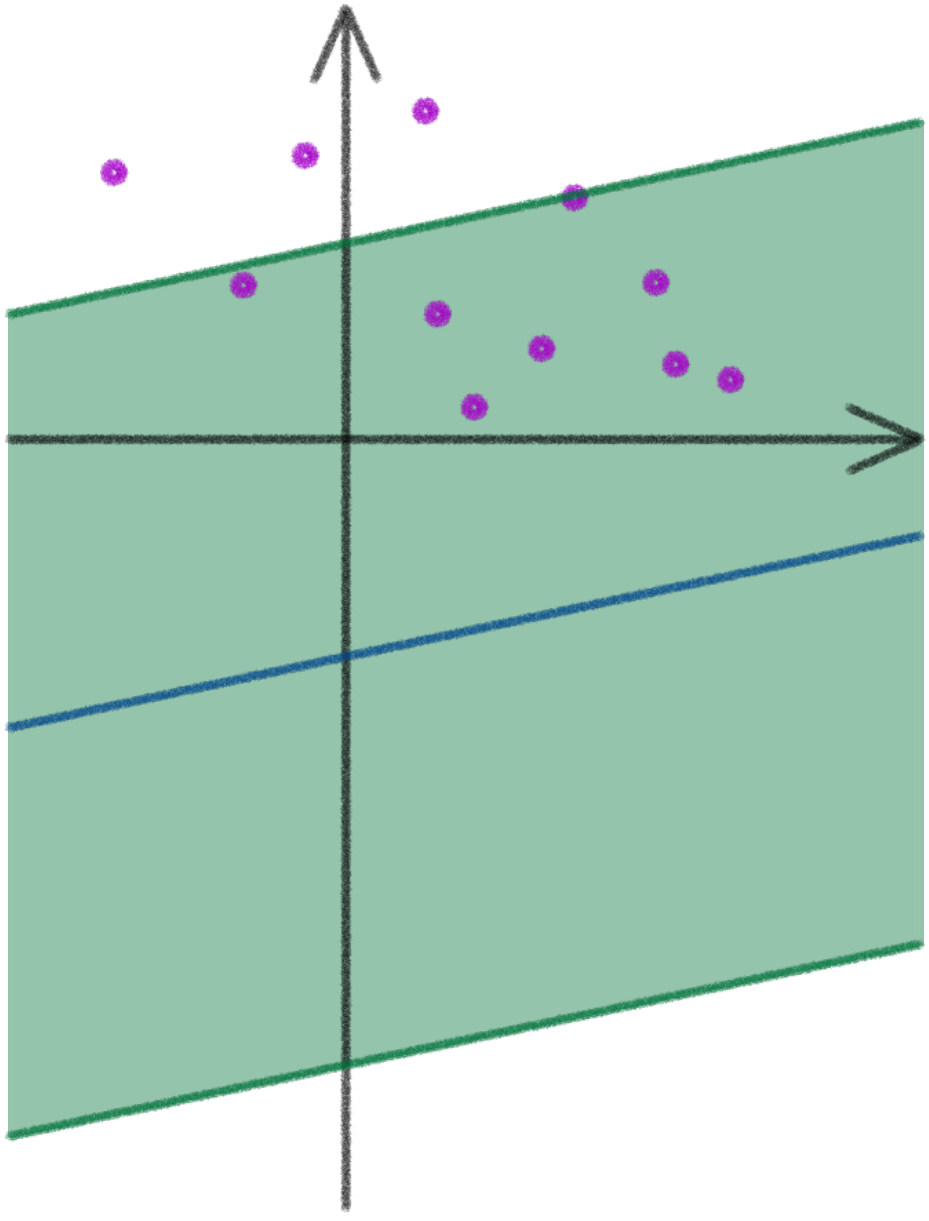
No assumptions on the model $\hat{f}$

No assumptions on the data distribution $\mathbb{P}$.

Too Good To Be True?

$\hat{f}$ bad

$\Rightarrow$ $\hat{C}_\alpha$ very large



Advantages

- Post-hoc
- Distribution-free
- Minimal hypotheses
- Finite-sample guarantee

Limitations

1. Calibration - marginal guarantee
2. Non - conditional guarantee
3. Need for calibration data
4. Bad with heteroscedastic data

1. Calibration-marginal guarantee

Calibration - marginal guarantee:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1-\alpha$$

how is this set constructed?

Calibration-marginal guarantee:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1-\alpha$$

$$\hat{C}_\alpha(X_{n+1}) = \left[\hat{f}(X_{n+1}) - \delta_\alpha, \hat{f}(X_{n+1}) + \delta_\alpha\right]$$

$\delta_\alpha = \lceil (n+1)(1-\alpha) \rceil$ smallest score

$$S_1 = |\hat{f}(X_1) - Y_1| \quad , \ldots, \quad S_n = |\hat{f}(X_n) - Y_n|$$

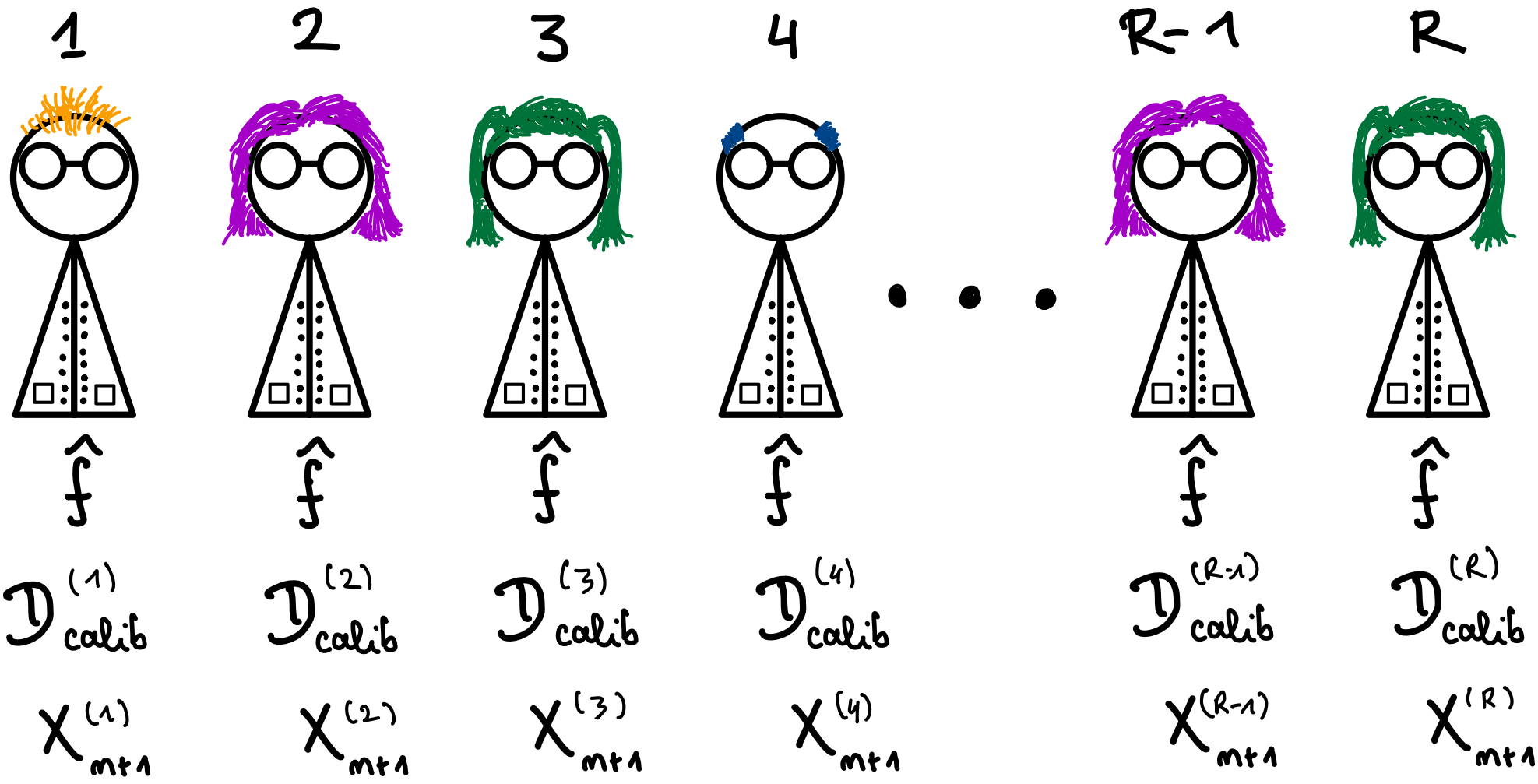
Calibration-marginal guarantee:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1-\alpha$$

On average over the draw of the calibration set!

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid \{(X_i, Y_i)\}_{i=1}^n\right) \geq 1-\alpha \; ?$$

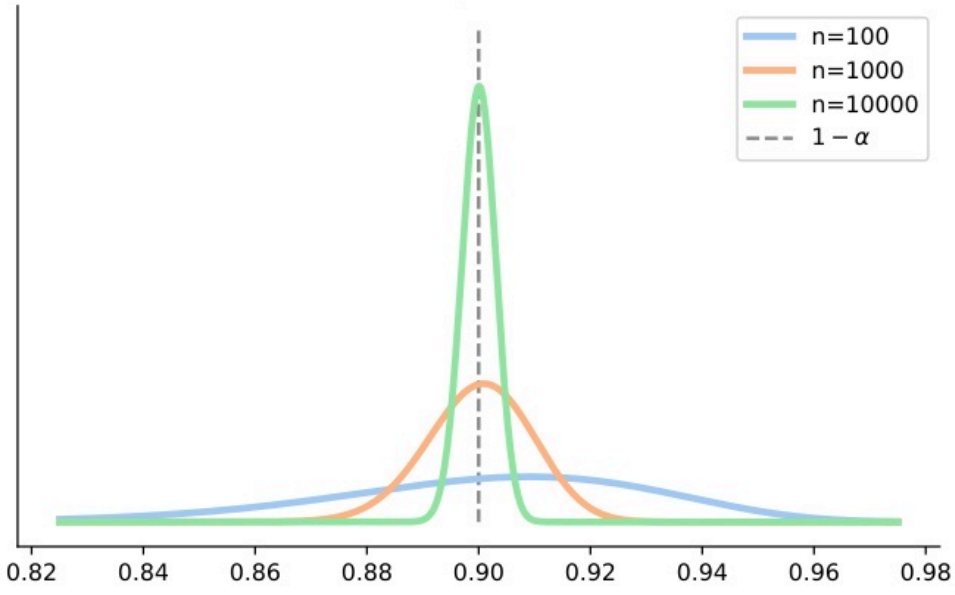
Calibration-marginal guarantee:



Calibration-Conditional Coverage

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1}) \mid \{(X_i, Y_i)\}_{i=1}^n\right) \sim \text{Beta}(n+1-\ell, \ell)$$

with $\ell = \lfloor (n+1)\alpha \rfloor$



*

We can pick $\delta > 0$, $\varepsilon > 0$ and choose $n(\varepsilon) = |\mathcal{D}^{\text{calib}}|$ so that with proba $1-\delta$

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1}) \mid \mathcal{D}^{\text{calib}}\right) \geq 1 - \alpha - \varepsilon$$

*

ϵ	0.1	0.05	0.01	0.005	0.001
$n(\epsilon)$	22	102	2491	9812	244390

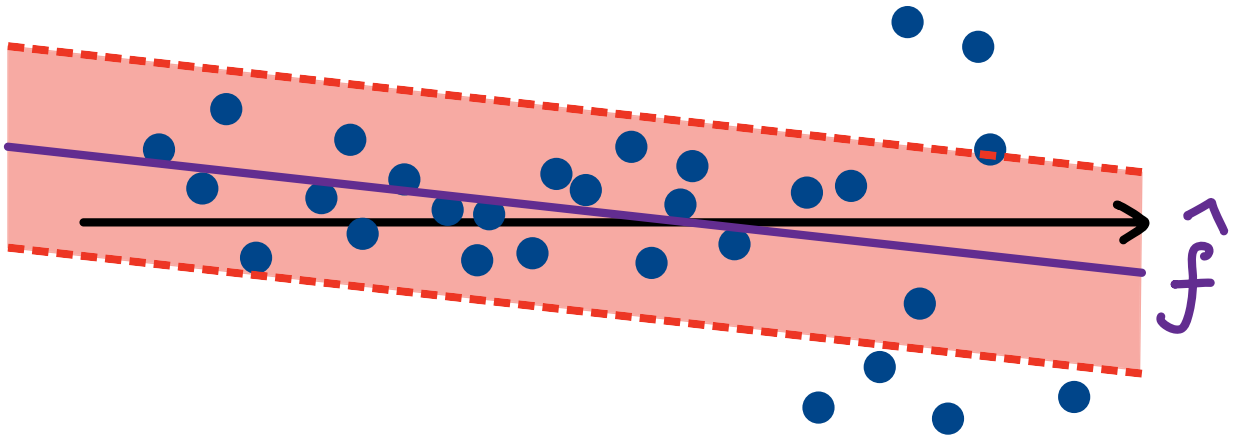
* Angelopoulos et al, 2022

2. NON-CONDITIONAL GUARANTEE

Non-conditional guarantee

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1-\alpha$$ ✓

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \,\middle|\, X_{n+1} = x\right) \geq 1-\alpha$$ ✗



CP EVALUATION

2 main metrics :

→ Coverage :

$$\frac{1}{|D^{test}|} \sum_{(x,y) \in D^{test}} \mathbb{1}_{y \in \hat{\mathcal{C}}_\alpha(x)}$$

should be close to $1-\alpha$

⚠ We are approximating

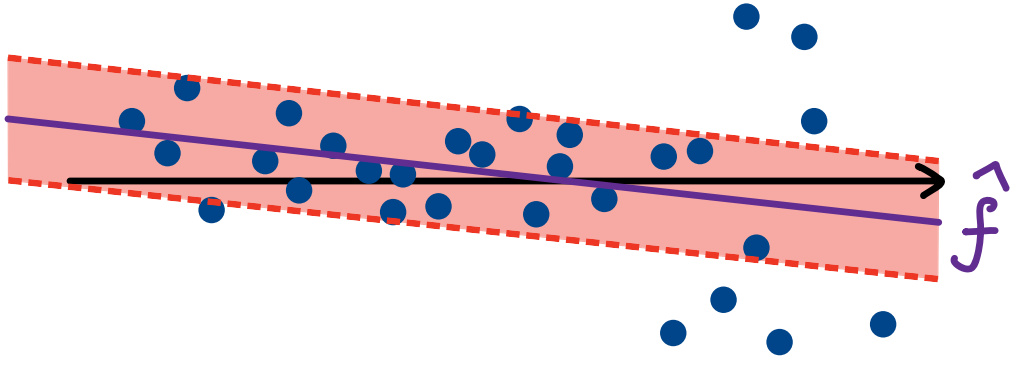
$$\mathbb{P}\left(Y_{n+1} \in \hat{\mathcal{C}}_\alpha(X_{n+1}) \mid D^{cal}\right)$$

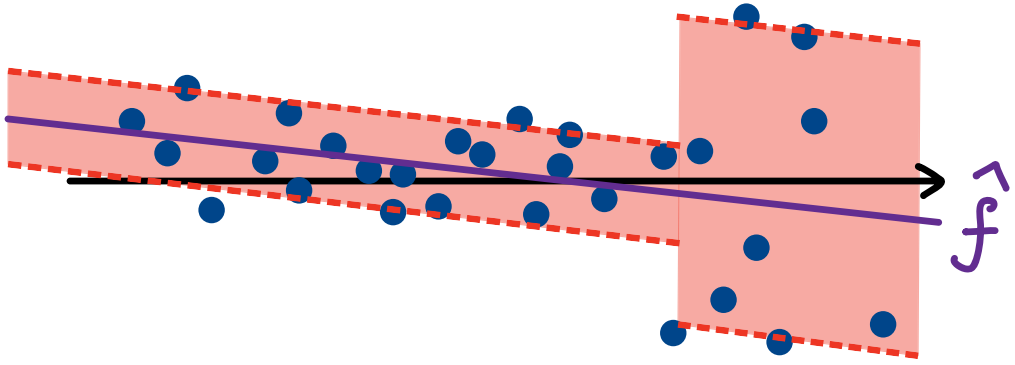
CP EVALUATION

→ average size of prediction sets:

$$\frac{1}{|D^{test}|} \sum_{(x,y) \in D^{test}} |\mathcal{C}_\alpha(x)|$$

Smaller isn't always better!





Impossibility of Conditional Coverage

Theorem (Vovk 2012)

If $\hat{C}_\alpha$ is such that

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} = x) \geq 1-\alpha \quad \forall x$$

$$\Rightarrow \hat{C}_\alpha = \mathbb{R} \quad \text{(except on some very particular cases)}$$

GROUP-BALANCED CP (VOVK 2012)

Assume $\mathcal{X} = \bigcup_{g=1}^{G} \mathcal{X}_g$ (partition)

we want:

$$\mathbb{P}\left(Y_{n+1} \in \hat{\mathcal{C}}_\alpha(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_g\right) \geq 1-\alpha$$

Exercise.- Write an algorithm that achieves the above coverage by calibrating on each of the $\mathcal{X}_g$ independently.

Solution :- Define $D_g^{cal} = \{(X,Y) \in D^{cal} : X \in \mathcal{X}_g\}$

(i) Compute the non-conformity scores:

$$S_i^{(g)} = |Y_i - \hat{f}(X_i)| \quad \text{for } i = 1, \ldots, n_g := |D_g^{cal}|$$

(ii) Compute the quantiles:

$$\delta^{(g)} = \lceil (n_g+1)(1-\alpha) \rceil \text{ smallest score from } S_1^{(g)}, \ldots, S_{n_g}^{(g)}$$

(iii) Build the prediction set: for g s.t. $X_{n+1} \in \mathcal{X}_g$

$$C_\alpha(X_{n+1}) = [\hat{f}(X_{n+1}) - \delta^{(g)}, \hat{f}(X_{n+1}) + \delta^{(g)}]$$

CP WITH CONDITIONAL GUARANTIES

(Gibbs et al. 2024)

$$\mathbb{P}\left(Y_{n+1} \in \hat{\mathcal{C}}_\alpha(X_{n+1}) \mid X_{n+1}\right) \geq 1-\alpha \quad \forall x$$

$$\Longleftrightarrow$$

$$\mathbb{P}\left[f(X_{n+1}) \left(\mathbb{1}_{Y_{n+1} \in \hat{\mathcal{C}}_\alpha(X_{n+1})} - (1-\alpha) \right) \right] = 0$$

for all measurable f.

CP WITH CONDITIONAL GUARANTIES

Gibbs et al. 2024

$$\mathbb{P}\left[f(X_{n+1}) \left(\mathbb{1}_{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})} - (1-\alpha) \right) \right] = 0$$

for all $f \in \mathcal{F}$.

Choosing the right $\mathcal{F}$:

→ group-balanced coverage (overlapping groups)

→ coverage under covariate shift

3. NEED FOR CALIBRATION DATA

Split Conformal

Theorem.- Given a calibration set $\{(X_i, Y_i)\}_{i=1}^{n}$ and a test point (X_{n+1}, Y_{n+1}), if the scores $S_i = |Y_i - \hat{f}(X_i)|$ are exchangeable, then the interval

$$\hat{C}_\alpha(X_{n+1}) = \left[\hat{f}(X_{n+1}) - \delta_\alpha \,,\; \hat{f}(X_{n+1}) + \delta_\alpha\right]$$

satisfies:

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right) \geq 1-\alpha$$

Split Conformal

$\hat{f}$ model (obtained by optimizing on D^{train})

$D^{calib} = \{(X_i, Y_i)\}_{i=1}^{n}$ calibration data

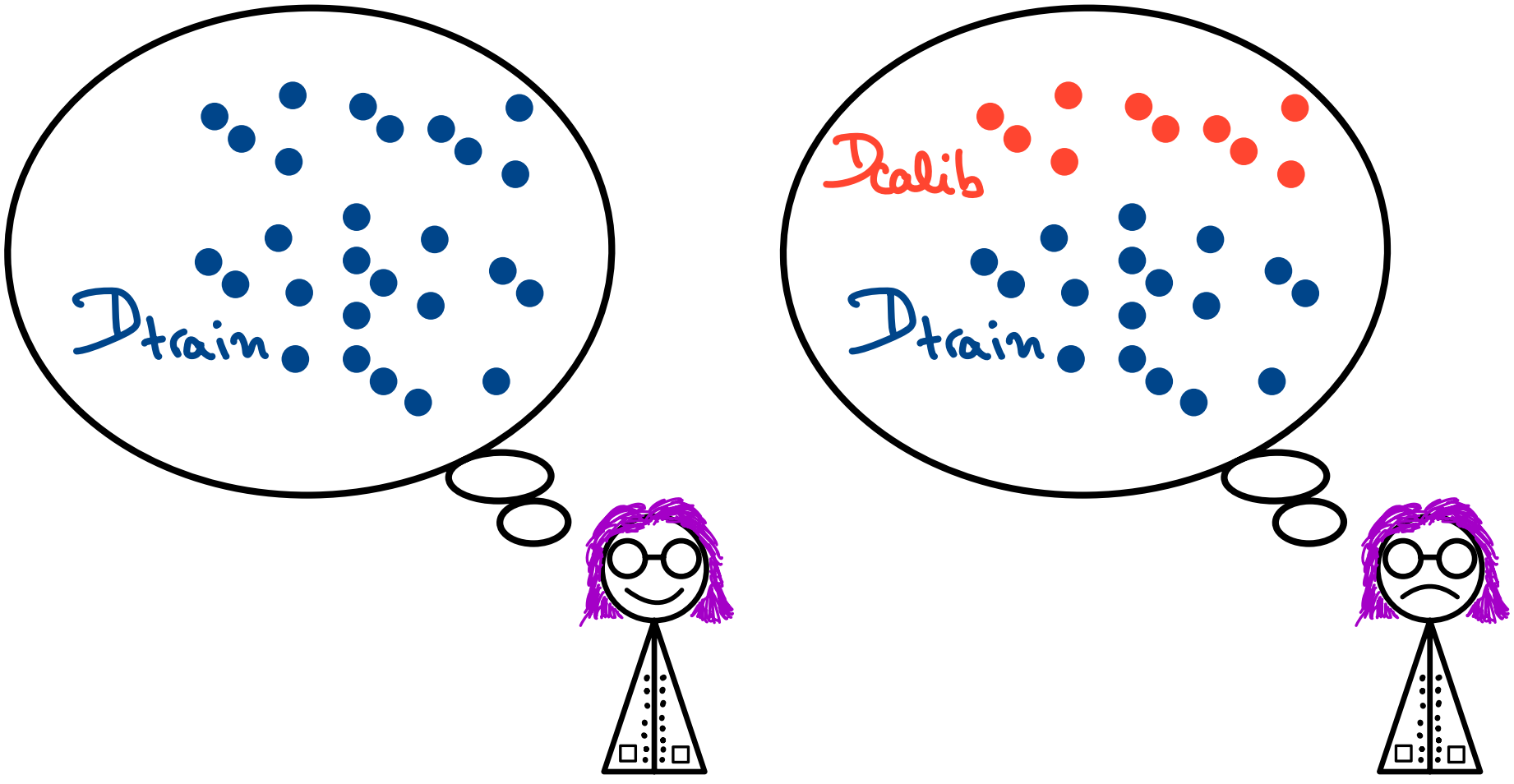
(X_{n+1}, Y_{n+1}) test point.

$S_i = |Y_i - \hat{f}(X_i)|, \quad i = 1, \ldots, n+1$

exchangeable

LIMITATIONS

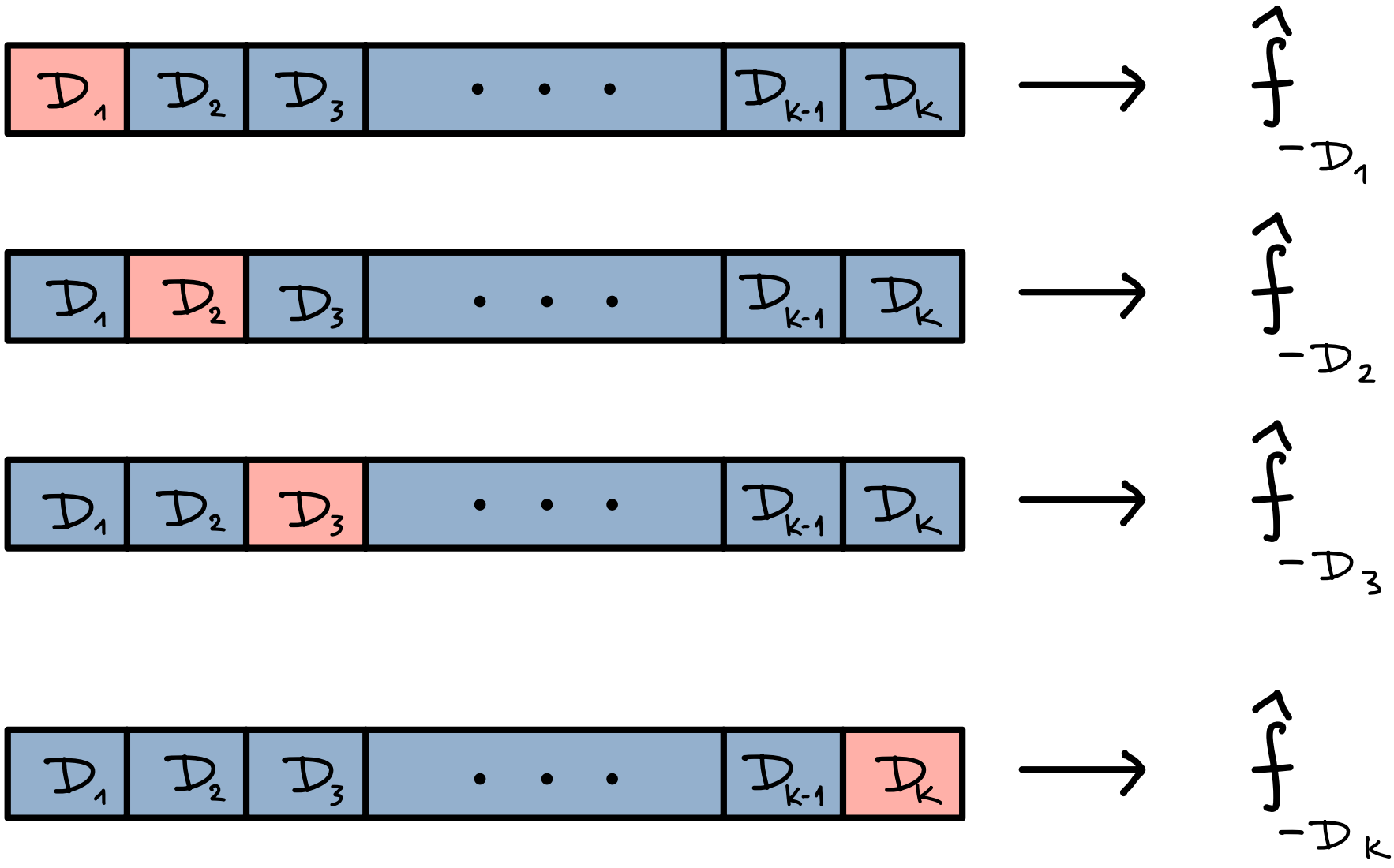
Need for calibration data



JACKNIFE+, CROSS-VALIDATION+

(Barber et al. 2019)

Partition the data:



JACKNIFE+, CROSS-VALIDATION+

Calibration:

$$S_i^{CV} = \left| Y_i - \hat{f}_{-D_{ind(i)}}(X_i) \right|, \quad i = 1, \ldots, n$$

JACKNIFE+ , CROSS-VALIDATION+

Inference:

$\hat{\ell}_\alpha(x) := \lfloor \alpha(n+1) \rfloor$ - th smallest value of

$$\hat{f}_{-S_{ind(1)}}(x) - S^{cv}_1, \ldots, \hat{f}_{-S_{ind(n)}}(x) - S^{cv}_n$$

$\hat{u}_\alpha(x) := \lceil (1-\alpha)(n+1) \rceil$ - th smallest value of

$$\hat{f}_{-S_{ind(1)}}(x) + S^{cv}_1, \ldots, \hat{f}_{-S_{ind(n)}}(x) + S^{cv}_n$$

$$\hat{C}_\alpha(x) = \left[\hat{\ell}_\alpha(x), \hat{u}_\alpha(x) \right]$$

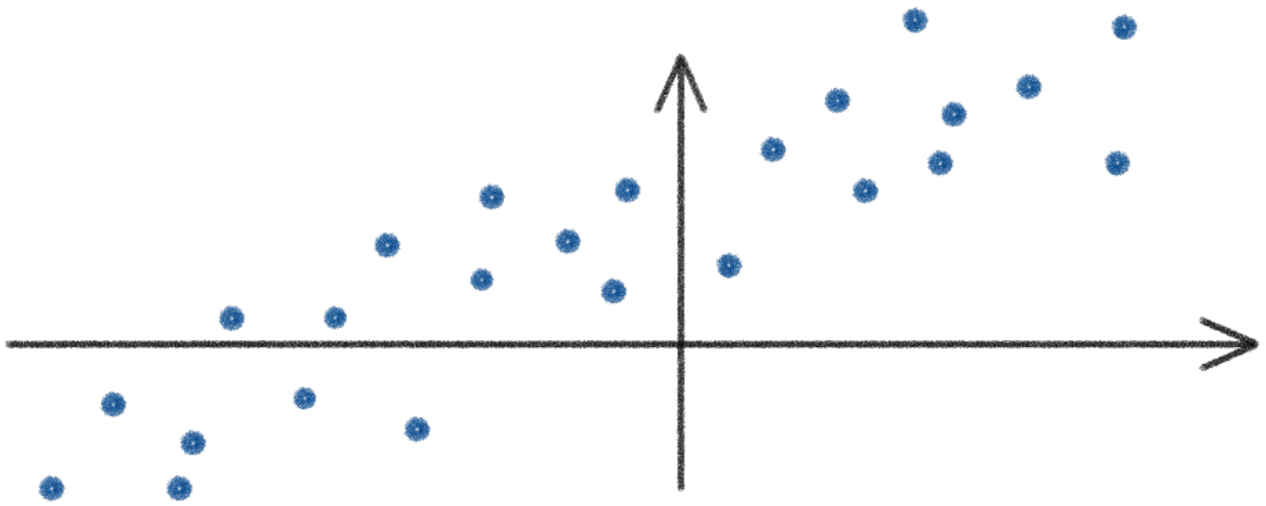
JACKNIFE+ , CROSS-VALIDATION+

Guarantee:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right) \geq 1-2\alpha$$

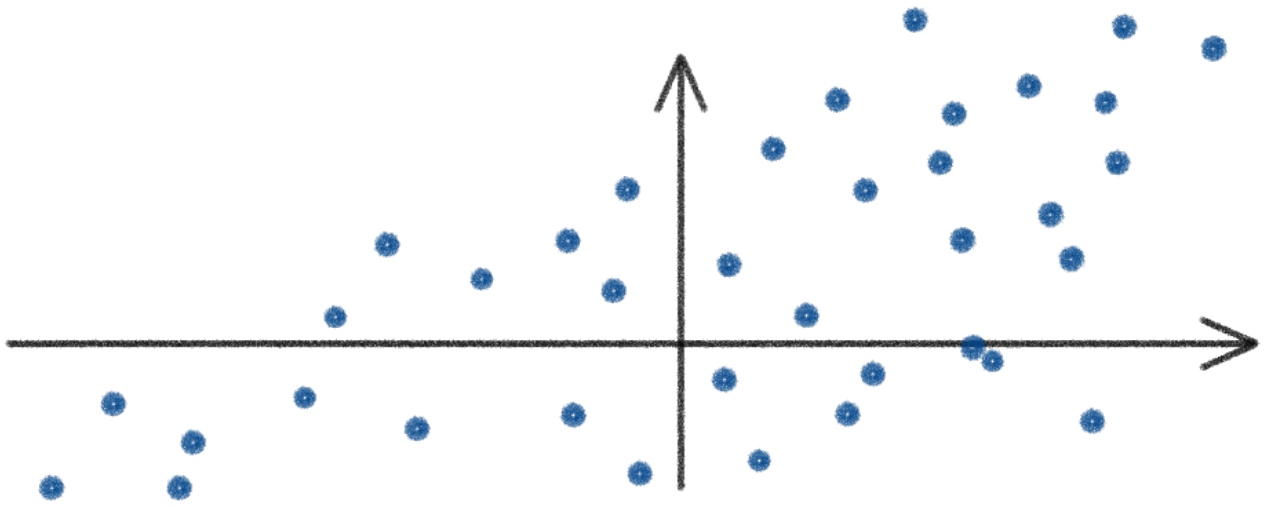
4. HETEROSCEDASTIC DATA

Homoscedastic data



$$Var(Y \mid X = x) = \text{constant}$$

Heteroscedastic data



$$Var(Y \mid X = x) = g(x)$$

LOCALLY ADAPTIVE CP (Papadopoulos et al 2008)

Train a second model $\hat{\sigma}$ such that

$$\hat{\sigma}(X) \approx |\hat{f}(X) - Y|$$

(i) Compute δ_α from the non-conformity scores:

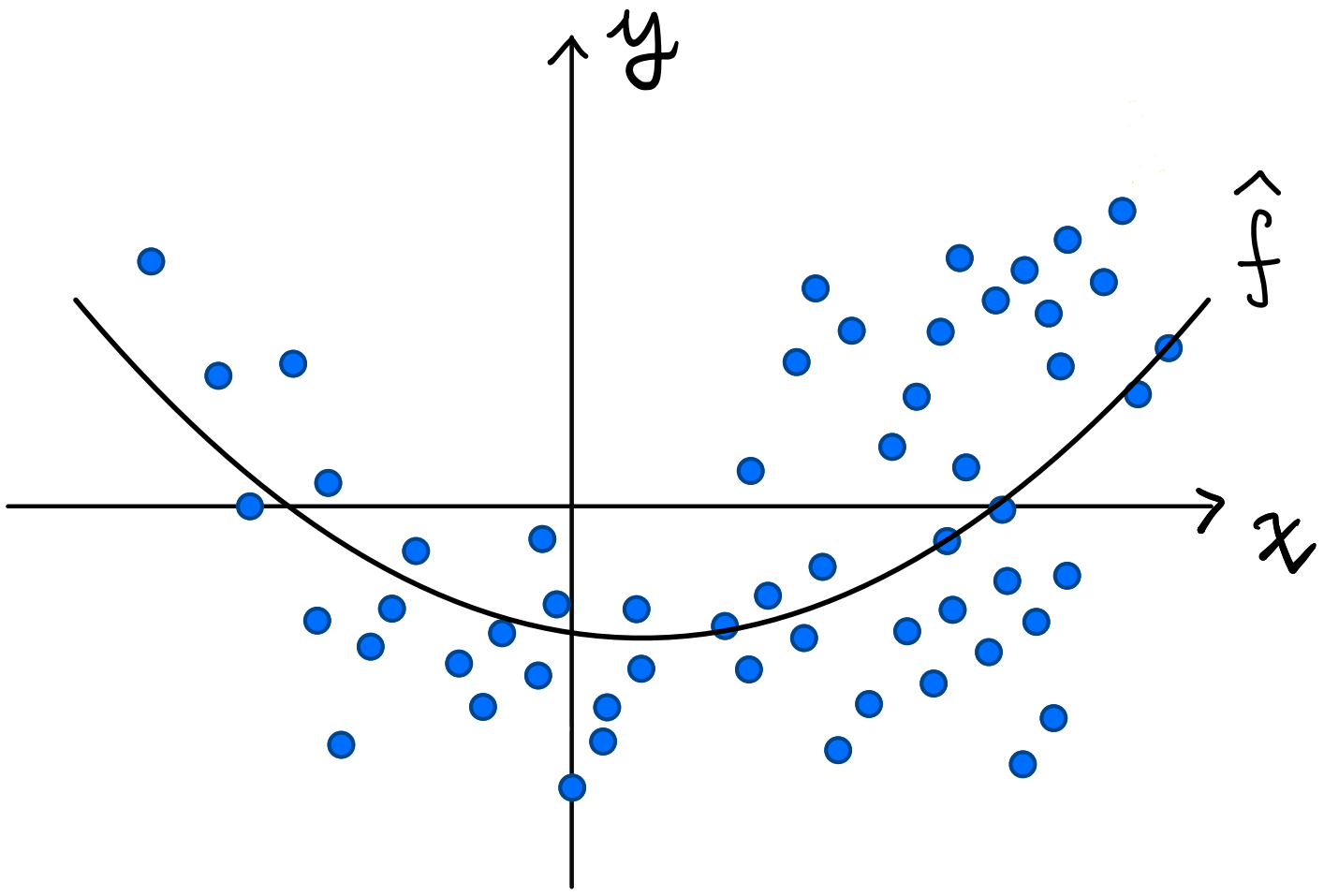
$$S_i = \frac{|\hat{f}(X_i) - Y_i|}{\hat{\sigma}(X_i)}$$

(ii) Build the prediction intervals as:

$$\hat{C}_\alpha(X) = [\hat{f}(X) - \hat{\sigma}(X)\delta_\alpha \, , \, \hat{f}(X) + \hat{\sigma}(X)\delta_\alpha]$$

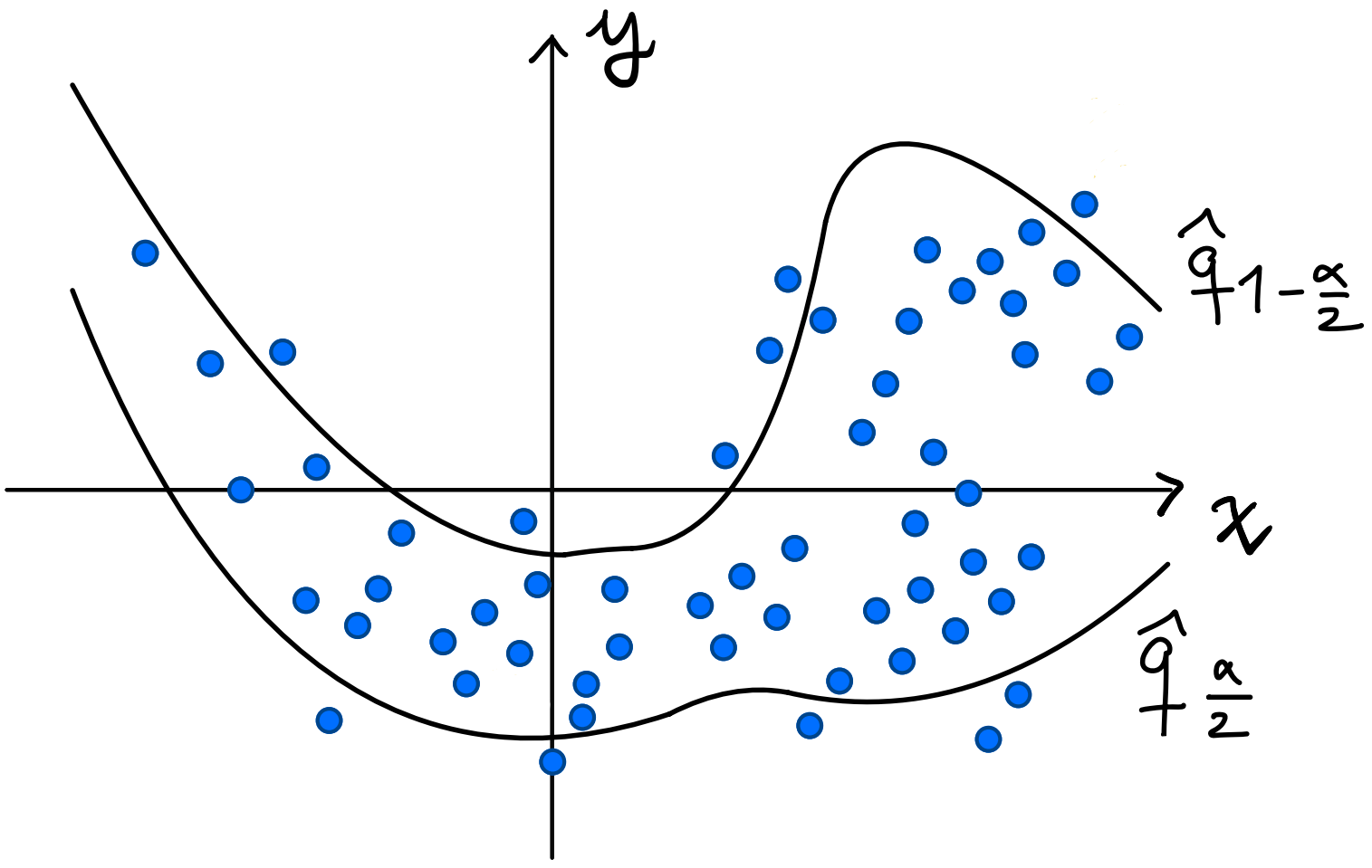
CQR: Conformal Quantile Regression

(Romano et al 2019)



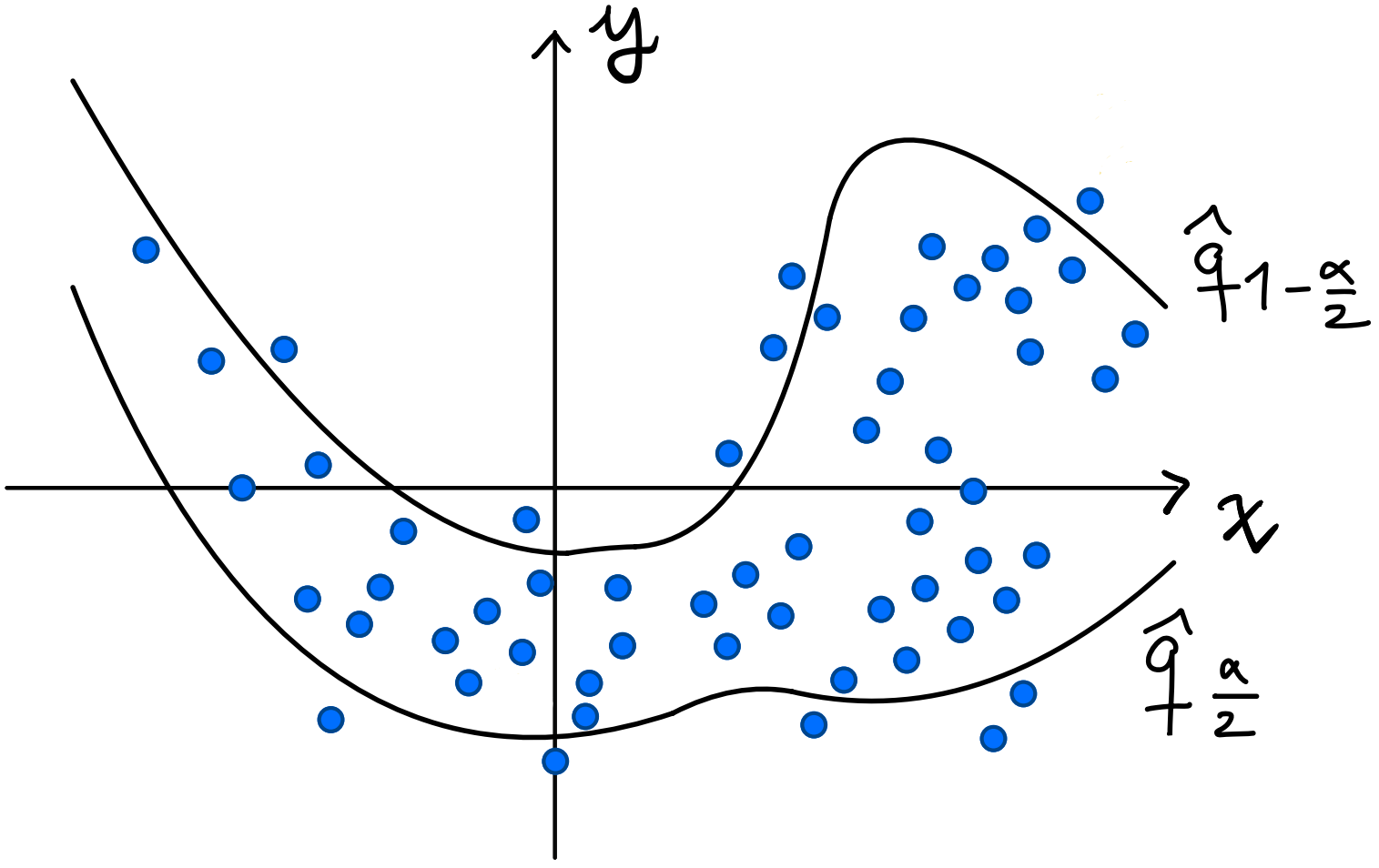
CQR

$$\hat{q}_t(x) \simeq \mathbb{P}(Y \leq t \mid X = x)$$



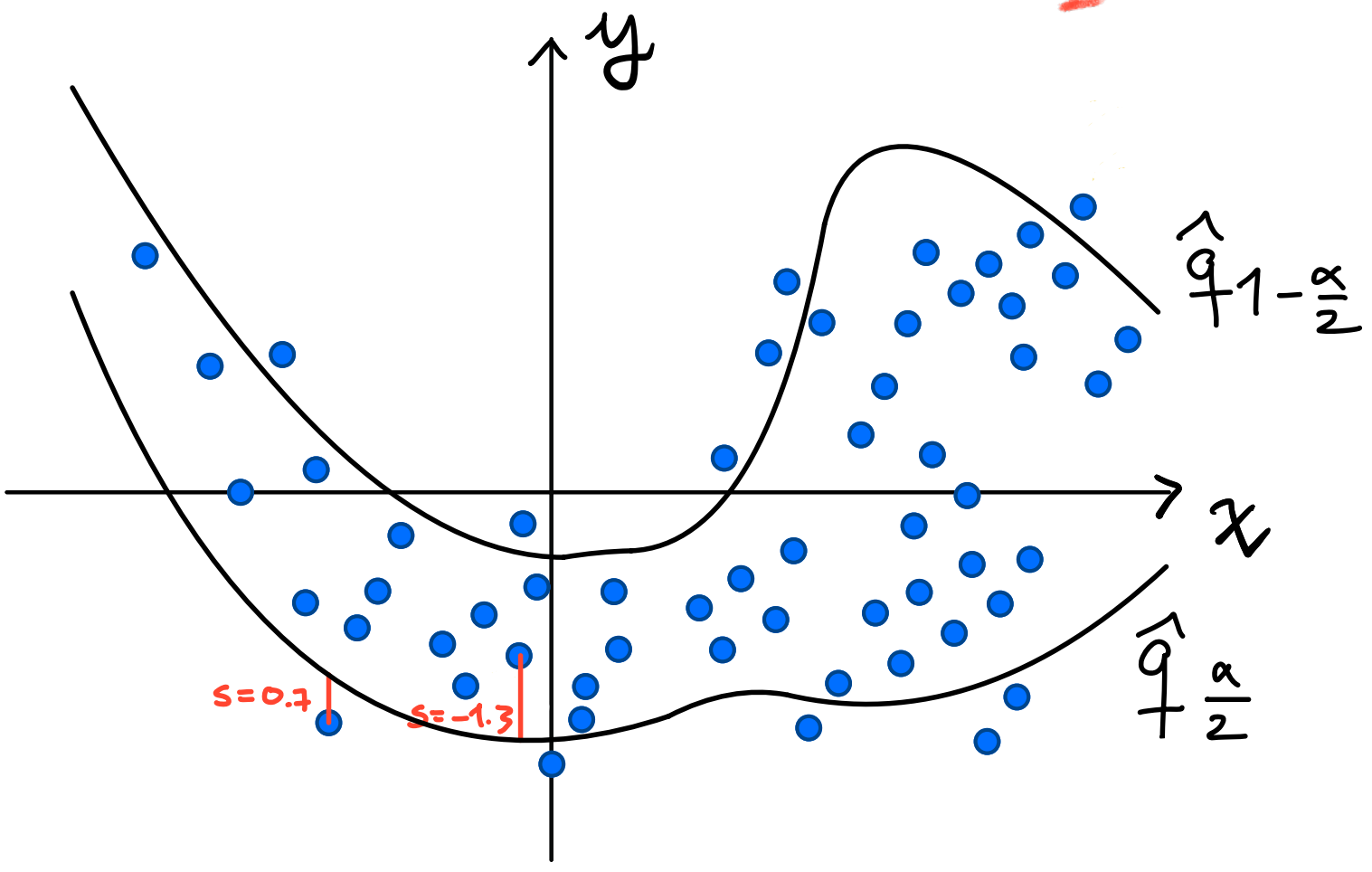
CQR

$$\mathbb{P}\left(Y \in \left[q_{\frac{\alpha}{2}}(x), q_{1-\frac{\alpha}{2}}(x)\right]\right) = ?$$



CQR: Calibration

$$s_i = \max\left\{ q_{\frac{\alpha}{2}}(x_i) - y_i,\ y_i + q_{1-\frac{\alpha}{2}}(x_i) \right\}$$



CQR: Calibration

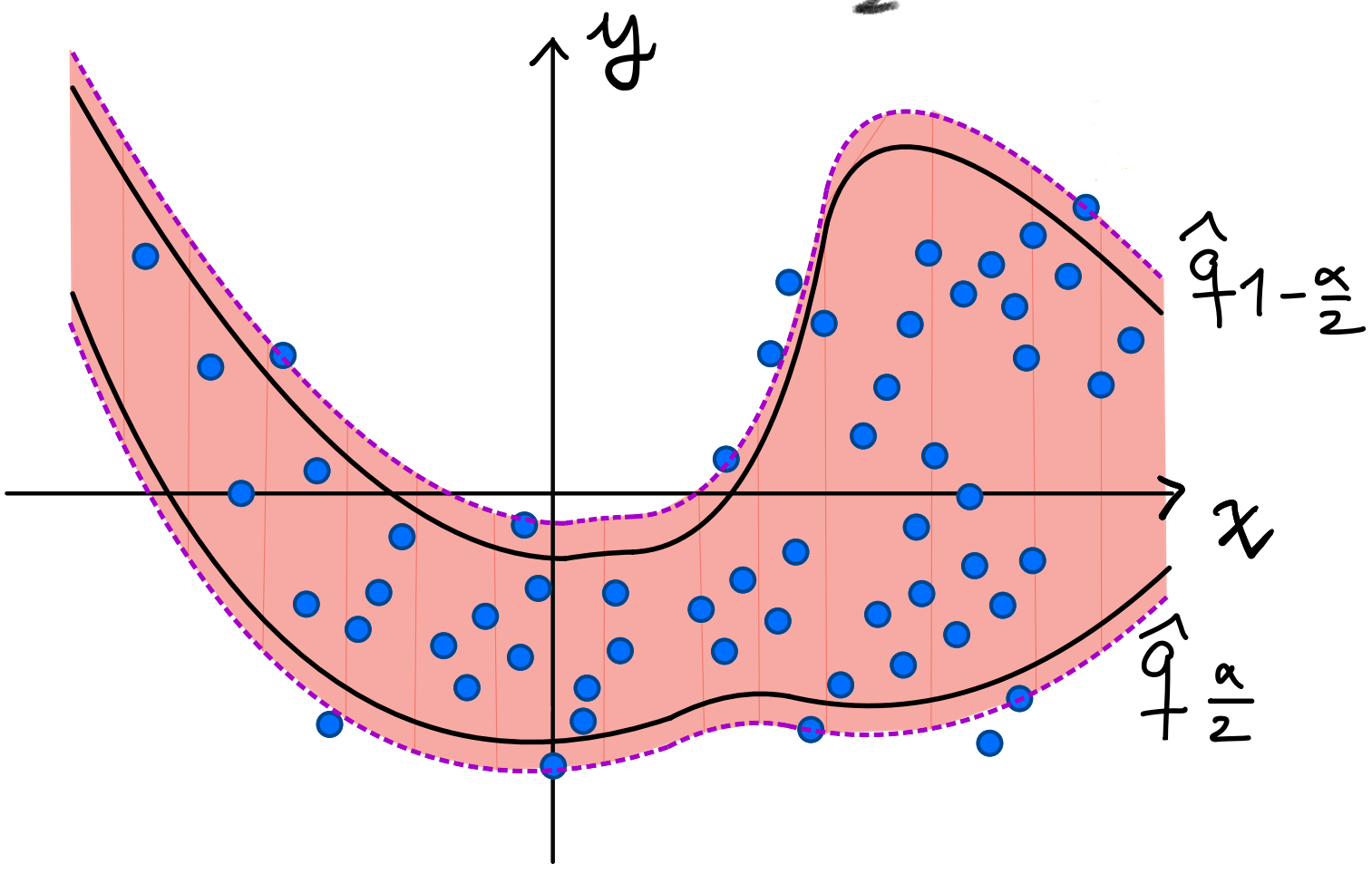
Calibration:

$$S_\alpha := \text{the } \left\lceil \frac{(n+1)(1-\alpha)}{n} \right\rceil \text{- th}$$

quantile of the scores $S_1, \ldots, S_n$

CQR : Calibration

$$C_\alpha(x) = \left[\hat{q}_{\frac{\alpha}{2}}(x) - \delta_\alpha,\ \hat{q}_{1-\frac{\alpha}{2}}(x) + \delta_\alpha\right]$$



General Formulation

Given : a predictor $\hat{f} : \mathcal{X} \longrightarrow \mathcal{Y}$

a nominal error rate α

a scoring function $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$

e.g. in regression $s(y, y') = |y - y'|$

Split conformal : Calibration set $\{(X_i, Y_i)\}_{i=1}^n$

$\Rightarrow$ non-conformity scores $R_i = s(\hat{Y}_i, Y_i)$

$\Rightarrow$ $q_\alpha = \left\lceil \frac{(1-\alpha)(n+1)}{n} \right\rceil$-quantile of the $\{R_i\}_{i=1}^n$

$\Rightarrow$ $\mathcal{C}(x) := \{y : s(\hat{y}, y) \leq q_\alpha\}$

Conformal Classification

- Least Ambiguous Set-Valued Classifiers (LAC)
- Adaptive Prediction Sets (APS)
- Regularized Adaptive Prediction Sets (RAPS)

LAC (Sadinle et al 2019)

Given: $\hat{\pi}$ softmax classifier

α nominal error rate

Use the nonconformity scores

$$s(\hat{\pi}, y) = \underbrace{1 - \hat{\pi}_y}$$

i.e. 1 - the probability of the true label.

APS (Romano et al 2020)

Given : $\hat{\pi}$ softmax classifier

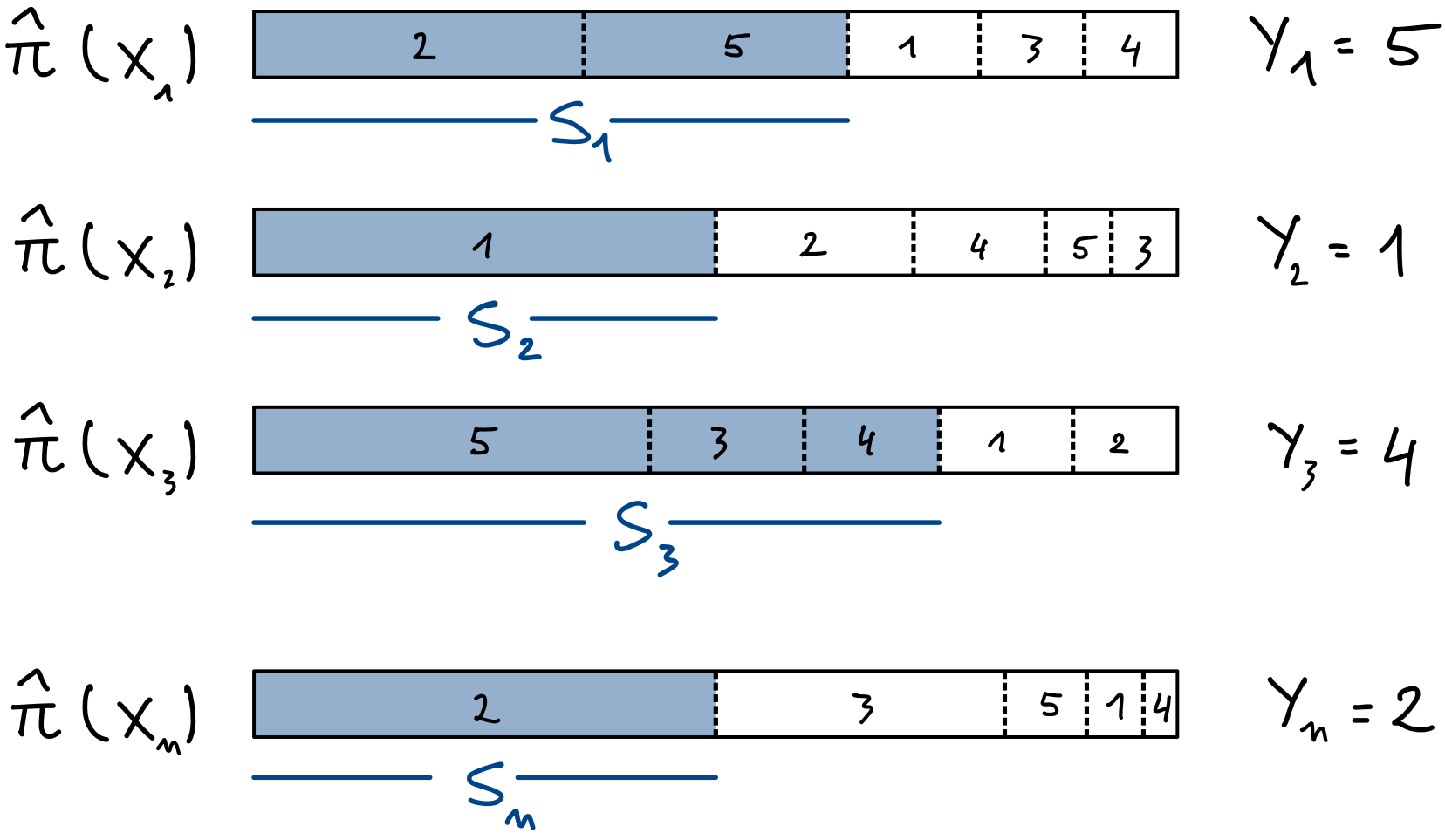
α nominal error rate

Rank: $\hat{\pi}_{(1)}(x) \geq \hat{\pi}_{(2)}(x) \geq \cdots \geq \hat{\pi}_{(K)}(x)$

$$L(x, \hat{\pi}, z) := \min_{c \in \{1, \ldots, K\}} \left\{ \hat{\pi}_{(1)}(x) + \cdots + \hat{\pi}_{c}(x) \geq z \right\}$$

APS

Calibration:



APS

Calibration:

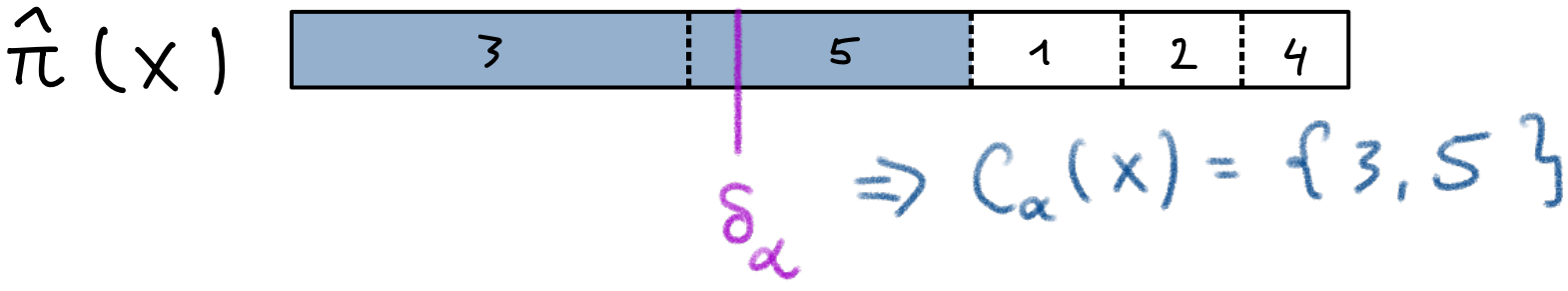
$$S_\alpha := \text{the } \left\lceil \frac{(n+1)(1-\alpha)}{n} \right\rceil \text{-th}$$

quantile of the scores $S_1, \ldots, S_n$

APS

Inference: $\hat{C}_\alpha(x) = \{(1), (2), \ldots, (k)\}$

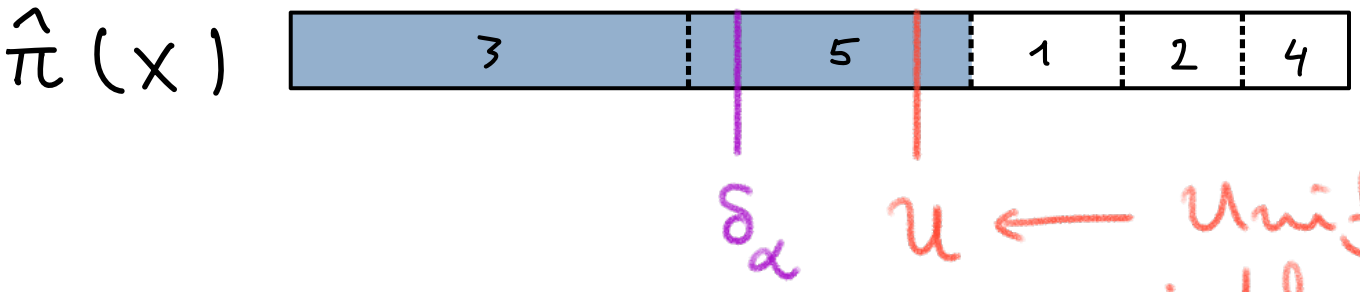
where $k = \min\{i : \hat{\pi}_{(1)} + \cdots + \hat{\pi}_{(i)} \geq \delta_\alpha\}$



Guarantee:

$$\mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1})) \geq 1 - \alpha$$

APS WITH RANDOMIZATION



$$\Rightarrow \quad \hat{\mathcal{C}}_\alpha(x) := \begin{cases} \{3\} & \text{if} \quad u \leq \delta_\alpha \\ \{3, 5\} & \text{if} \quad u > \delta_\alpha \end{cases}$$

Same guarantee, tighter prediction sets!

RAPS (Angelopoulos et al 2020)

Given : $\hat{\pi}$ softmax classifier

$\propto$ nominal error rate

Rank: $\hat{\pi}_{(1)}(x) \geq \hat{\pi}_{(2)}(x) \geq \cdots \geq \hat{\pi}_{(K)}(x)$

$$S_i := \hat{\pi}_{(1)} + \cdots + \hat{\pi}_{(k)} + \lambda(k - k_{reg} + 1)$$

k such that $(k) = Y_i$

rank of Y_i

hyper-parameters

RAPS

Prediction set:

$$\hat{\mathcal{C}}_\alpha(x) = \{(1), \dots, (k)\}$$

where

$$k = \max\{i : \pi_{(1)} + \dots + \pi_{(i)} + \lambda(i - k_{reg} + 1)^+ \leq \delta_\alpha\} + 1$$

More stable than APS if softmax is "noisy" i.e. many classes w. low proba.

Kahoot time!

WRAP-UP

1. List the main advantages of CP.

2. List the main limitations of CP.

3. Give an example of a use-case where you might use CP.

4. Write down a question about sth you did not fully understand / would like to know more about.