Lecture at IHP (January 19th, 2024) — Sébastien Gerchinovitz

Guaranteed prediction sets via concentration inequalities

1. Introduction

Split conformal prediction:

$(X_1, Y_1), \dots, (X_n, Y_n), (X, Y)$ exchangeable

$s(x, y)$ "non-conformity score" (e.g., for regression $s(x,y) = |y - f(x)|$ with pretrained f)

$S_i := s(X_i, Y_i)$, $i = 1, \dots, n$

$\rightsquigarrow$ $S_{(1)} \le S_{(2)} \le \dots \le S_{(n)}$ "order statistics"

Prediction set (for a fixed $\alpha \in \left[\frac{1}{n+1}, 1\right)$):

$$C(x) := \left\{ y \in \mathcal{Y} : s(x,y) \le S_{(\lceil (1-\alpha)(n+1) \rceil)} \right\}$$

Marginal coverage guarantee: $\mathbb{P}(Y \in C(X)) \ge 1 - \alpha$

↳ w.r.t. to $\underbrace{(X_1, Y_1), \dots, (X_n, Y_n)}_{\mathcal{D}_n}, (X, Y)$

Question: can we derive calibration-conditional guarantees of the form

$$\mathbb{P}\left[\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) \ge 1 - \alpha \right] \ge 1 - \delta \quad ?$$

↑ for most calibration data $\mathcal{D}_n = (X_i, Y_i)_{1 \le i \le n}$ the coverage on test points is $\ge 1 - \alpha$

Somewhat stronger than

$$\mathbb{E}\left[\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) \right] = \mathbb{P}(Y \in C(X)) \ge 1 - \alpha.$$

2. A "tolerance region" viewpoint on split conformal prediction

2.1. Reminder: tolerance regions

Wilks (1941), Wald (1943), Scheffé and Tukey (1945), Tukey (1947), etc.

Suppose $Z_1, \dots, Z_n$ are i.i.d. copies of a real-valued r.v. Z

F: cdf of Z

$\mathcal{L}$: law of Z

Problem: construct a "tolerance region" $[\varphi_1(Z_1,\dots,Z_n), \varphi_2(Z_1,\dots,Z_n)] = [T_1, T_2]$ such that $\mathbb{P}\big(\mathcal{L}([T_1,T_2]) \geq 1-\alpha\big) \geq 1-\delta$.

$[T_1,T_2]$ contains (at least) a $1-\alpha$ fraction of the population mass.

for most observed n-samples

One-sided alternative: $\mathbb{P}\big(\mathcal{L}((-\infty; T_1]) \geq 1-\alpha\big) \geq 1-\delta$.

$\Leftrightarrow T_1 \geq \text{quantile}_{1-\alpha}(\mathcal{L})$

Solution: T_1 and T_2 of the form $T_i = Z_{(k_i)}$ (order statistics)

Theorem (Wilks '41, Scheffé and Tukey '45) "Wilks' method"

- If F is continuous, $\mathcal{L}([Z_{(r)}, Z_{(n+1-r)}]) \sim \text{Beta}(n+1-2r, 2r)$
  $\mathcal{L}((-\infty; Z_{(k)}]) \sim \text{Beta}(k, n+1-k)$
- If F is not continuous, the random masses $\mathcal{L}(\cdot)$ above stochastically dominate the Beta distributions above. E.g.,

  (stochastically dominated by $\overset{\text{sto}}{\geq}$ below)

  $\mathbb{P}\big(\mathcal{L}((-\infty; Z_{(k)}]) \geq t\big) \geq \mathbb{P}(B \geq t)$ with $B \sim \text{Beta}(k, n+1-k)$

  In particular, if $\mathbb{P}(B \geq 1-\alpha) \geq 1-\delta$, then this also holds for $\mathcal{L}((-\infty; Z_{(k)}])$

Proof ingredient : simulate the Z_i's by inverse transform sampling, and reduce the problem to the case $\mathcal{L} = \mathcal{U}([0,1])$, up to an inequality if F is not continuous.

2.2. Application to split conformal predictors (see Vovk 2012)

In all the sequel, we assume that $\underbrace{(X_1,Y_1),\dots,(X_n,Y_n)}_{\mathcal{D}_n}$ are independent copies of (X,Y) and that $\alpha \in \left[\frac{\ln(1/\delta)}{n}, 1\right)$

Recall that $C(x) = \{y \in \mathcal{Y} : s(x,y) \le s_{(k)}\}$

so that $\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) = \mathbb{P}(s(X,Y) \le s_{(k)} \mid \mathcal{D}_n)$

$= \mathcal{L}((-\infty, s_{(k)}))$ with $\mathcal{L} :=$ law of $s(X,Y)$

$\overset{\text{ste}}{\gtrsim} \text{Beta}(k, n+1-k)$

$s_{(k)} = k$-th order statistic of $s(X_i,Y_i)$, $i=1,\dots,n$

To get $\mathbb{P}\left[\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) \ge 1-\alpha\right] \ge 1-\delta$, we thus pick $k \in \{1,\dots,n\}$ such that

$$\text{quantile}_\delta\left(\text{Beta}(k, n+1-k)\right) \ge 1-\alpha$$

↳ feasible iff $\alpha \ge 1-\delta^{1/n}$ (e.g., if $\alpha \ge \frac{\ln(1/\delta)}{n}$)

cf. Appendix A for details

$\Leftrightarrow$ cdf of Beta/Bin distributions

$$\Leftrightarrow \text{Beta}_{k,n+1-k}(1-\alpha) \le \delta$$

$$\Leftrightarrow \text{Bin}_{n,1-\alpha}(k-1) \ge 1-\delta$$

Lemma (e.g., Vovk 2012)

$\forall n \ge 1, \forall k \in \{1,\dots,n\}, \forall p \in [0,1]$,

$$\text{Bin}_{n,p}(k-1) = 1 - \text{Beta}_{k,n+1-k}(p)$$

We thus pick $k := 1 + \text{quantile}_{1-\delta}\left(\text{Bin}(n, 1-\alpha)\right)$ and set

$$C(x) := \{y : s(x,y) \le s_{(k)}\}.$$

Then: $\mathbb{P}\left[\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) \ge 1-\alpha\right] \ge 1-\delta$.

Note that the choice of $k = \lceil (n+1)(1-\alpha) \rceil$ recommended by the split CP method only yields: (when $S(X,Y)$ has a continuous cdf)

$$\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) \sim \text{Beta}(k, n+1-k), \text{ whose expectation equals } \frac{k}{n+1} \approx 1-\alpha$$

Therefore, the coverage $\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n)$ is "often" below $1-\alpha$.

(but not much below, since, by Bernstein's inequality (see Appendix A for details),

$$\mathbb{P}\left[\mathbb{P}(Y \in C(X) \mid \mathcal{D}_n) \geq 1 - \alpha - \underbrace{\left[\sqrt{\frac{2\alpha \ln 1/\delta}{n}} - \frac{2 \ln 1/\delta}{n} \right]}_{\ll \alpha \text{ if } n \gg \frac{\ln(1/\delta)}{\alpha}} \right] \geq 1 - \delta$$

3. Beyond binary losses

Two key references:

RCPS algorithm

- Bates et al. (2021), Distribution-free, risk-controlling prediction sets.
- Park et al. (2021), PAC confidence sets for DNNs via calibrated prediction.

3.1. Setting: set-valued prediction with nested sets

Let

- loss function: $L(y, S) \geq 0$ be nonincreasing in $S \subset \mathcal{Y}'$, for any $y \in \mathcal{Y}$
- $(C_\lambda(\cdot))_{\lambda \in \Lambda}$ be a family of set-valued predictors, with $\Lambda \subset \mathbb{R}$ closed and satisfying the nesting property: $\lambda_1 \leq \lambda_2 \Rightarrow \forall x \in \mathcal{X}, C_{\lambda_1}(x) \subset C_{\lambda_2}(x)$
- assume also: $\lim_{\lambda \to \sup(\Lambda)} L(y, C_\lambda(x)) = 0$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

Risk: Let (X,Y) be a random pair in $\mathcal{X}\times\mathcal{Y}$. We set: for all $\lambda\in\Lambda$,

$$R(\lambda) := \mathbb{E}\left[L(Y,C_\lambda(X))\right] = \int L(y,C_\lambda(x))\, dP_{X,Y}(x,y) \underbrace{< +\infty}_{\text{assumption}}$$

Ex 1: conformal prediction!

$$C_\lambda(x) = \{y\in\mathcal{Y} : s(x,y)\le\lambda\}$$ ← NB: any $(C_\lambda(x))_{\lambda\in\Lambda}$ is of this form if $\lim_{\lambda\downarrow\lambda_0} C_\lambda(x) = C_{\lambda_0}(x)$

$$L(y,S) = \mathbb{1}_{y\notin S}$$

$$R(\lambda) = \mathbb{E}\left[L(Y,C_\lambda(X))\right] = \mathbb{P}(Y\notin C_\lambda(X))$$

Ex 2: multiclass classification with class-varying loss

$$\mathcal{Y} = \{1,\dots,K\}$$

For $S\subset\mathcal{Y}$, $L(y,S) = \ell_y\cdot\mathbb{1}_{y\notin S}$ (some mistakes are more serious than others)

$C_\lambda(x) \overset{\text{e.g.}}{=}$ {all classes that a trained DNN predicts as likely (softmax $\ge 1-\lambda$)}

Ex 3: multilabel classification

$x\in\mathcal{X}$, $y\subset\{1,\dots,K\}$ (multiple classes for object x)

For $S\subset\{1,\dots,K\}$, $L(y,S) = \dfrac{|y\setminus S|}{|y|} = 1 - \underbrace{\dfrac{|y\cap S|}{|y|}}_{\text{"recall"}}$

(NB: in this example, S and y live in the same space $= \mathcal{P}(\{1,\dots,K\})$)

Other examples: image segmentation (close to Ex. 3)
hierarchical classification
protein structure prediction
etc... as long as • $(C_\lambda(i))_{\lambda \in \mathbb{R}}$ are nested
• $L(y, S)$ is nonincreasing in S

3.2. The RCPS algorithm (RCPS = Risk Controlling Prediction Sets)

Let $\alpha > 0$ and $\delta \in (0,1)$ be two predefined "risk levels".

Let (X_i, Y_i), $1 \le i \le n$, be $\perp\!\!\!\perp$ copies of (X, Y). We set $\mathcal{D}_n := (X_i, Y_i)_{1 \le i \le n}$

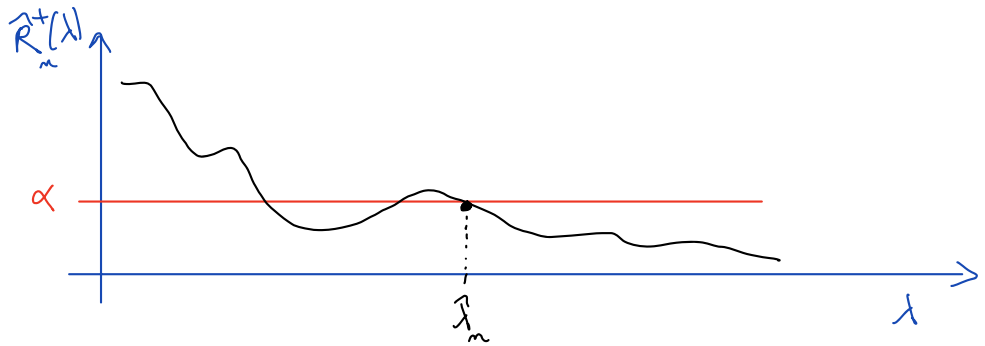
Assume $\lambda \mapsto \hat{R}_n^+(\lambda)$ is a random function (built from $\mathcal{D}_n$) such that:

$$\forall \lambda \in \Lambda, \quad \mathbb{P}\left(R(\lambda) \le \hat{R}_n^+(\lambda)\right) \ge 1 - \delta. \qquad (1)$$

↳ pointwise high-probability bound

Then, choose $\hat{\lambda}_n$ from $\mathcal{D}_n$ such that (assuming existence)

$$\forall \lambda' \ge \hat{\lambda}_n, \quad \hat{R}_n^+(\lambda') \le \alpha \qquad (2)$$



(in some cases, $\hat{R}_n^+(\lambda)$ is nonincreasing)

Examples: we assume $L(y,s)$ bounded in, say, $[0,1]$.

- Using Hoeffding's inequality and setting $\hat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n L(Y_i, C_\lambda(X_i))$,

$$\mathbb{P}\left(R(\lambda) \le \underbrace{\hat{R}_n(\lambda) + \sqrt{\frac{\ln(1/\delta)}{2n}}}_{\hat{R}_n^+(\lambda) \text{ satisfies (1)}} \right) \ge 1-\delta$$

- If instead we use an "empirical Bernstein bound" (Maurer and Pontil '2009), we obtain, with $\hat{\sigma}^2(\lambda) = \frac{1}{n-1}\sum_{i=1}^n \left(L(Y_i, C_\lambda(X_i)) - \hat{R}_n(\lambda)\right)^2$

$$\mathbb{P}\left(R(\lambda) \le \underbrace{\hat{R}_n(\lambda) + \hat{\sigma}(\lambda)\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)}}_{\hat{R}_n^+(\lambda) \text{ satisfies (1)}} \right) \ge 1-\delta$$

- Many other non-asymptotic concentration inequalities! (beyond bounded loss)
- Let us just mention a tight $\hat{R}_n^+(\lambda)$ when $L(y,s) = \mathbb{1}_{y \notin s} \in \{0,1\}$:

$$\hat{R}_n^+(\lambda) = \sup\left\{ p \in [0,1] : \underbrace{\text{Bin}_{n,p}}_{\text{cdf of Bin}(n,p)}\left(n\hat{R}_n(\lambda)\right) \ge \delta \right\} \qquad (3)$$

See, e.g., Langford (2005, Theorem 3.3) for a proof, or Appendix A below.

NB: In that case, RCPS reduces to Wilks' method! [see Section 3.3 below]

Theorem (Bates et al. 2021) Assume $(X_i, Y_i)_{1 \le i \le n}$ are i.i.d. copies of $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, and that $\lambda \mapsto \hat{R}_n^+(\lambda)$ satisfies (1) with $\alpha > 0$, $\delta \in (0,1)$. Then, any $\hat{\lambda}_n$ such that (2) holds satisfies

$$\mathbb{P}\left(R(\hat{\lambda}_n) \le \alpha \right) \ge 1-\delta \qquad \left[\mathbb{P} \text{ is w.r.t random draw of } (X_i, Y_i)_{1 \le i \le n}\right]$$

If R is right-continuous, the conclusion also holds with

$$\hat{\lambda}_n := \inf\left\{ \lambda \in \Lambda : \forall \lambda' \ge \lambda,\ \hat{R}_n^+(\lambda') \le \alpha \right\}$$

Proof: We can assume wlog that $\{\lambda \in \Lambda : R(\lambda) > \alpha\} \neq \emptyset$.

Define

$$\lambda^* := \sup\{\lambda \in \Lambda : R(\lambda) > \alpha\} < +\infty \quad \text{since} \quad \lim_{\lambda \to \sup(\Lambda)} R(\lambda) = 0$$

We distinguish two cases:

Case 1: $R(\lambda^*) > \alpha$

By (1), we have, with probability $\geq 1-\delta$,

$$\alpha < R(\lambda^*) \leq \hat{R}_n^+(\lambda^*)$$

so that, by (2): $\hat{\lambda}_n > \lambda^*$, which implies $R(\hat{\lambda}_n) \leq \alpha$. (by def of λ^*)

Case 2: $R(\lambda^*) \leq \alpha$

Let $\lambda \in \Lambda$ such that $\lambda < \lambda^*$. In particular: $R(\lambda) > \alpha$.

As before, with probability $\geq 1-\delta$,

$$\alpha < R(\lambda) \leq \hat{R}_n^+(\lambda)$$

so that $\hat{\lambda}_n > \lambda$ on this event.

Therefore, $\mathbb{P}(\hat{\lambda}_n \geq \lambda^*) = \lim_{\lambda \uparrow \lambda^*} \mathbb{P}(\hat{\lambda}_n > \lambda) \geq 1-\delta$.

In particular, $\mathbb{P}\left(R(\hat{\lambda}_n) \leq \alpha\right) \geq \mathbb{P}\left(R(\hat{\lambda}_n) \leq R(\lambda^*)\right) \geq 1-\delta$

This includes the first part of the theorem.

The second part follows as a consequence. ∎

3.3. For Bernoulli losses, RCPS with (3) computes Wilks' tolerance region!

We assume here that $L(y,S) = \mathbb{1}_{y \notin S}$, so that $\hat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \notin C_\lambda(X_i)}$

$\hat{R}_n^+(\lambda) := \sup\left\{ p : \mathrm{Bin}_{n,p}\left(n\hat{R}_n(\lambda)\right) \geq \delta \right\}$ satisfies (1).
↑ by, e.g., Langford (2005, Thm 3.3)

We also assume that $\forall x \in \mathcal{X}, \forall \lambda_0 \in \Lambda, \lim_{\lambda \downarrow \lambda_0} C_\lambda(x) = C_{\lambda_0}(x)$ (often true)

Note that: (the case $n\hat{R}_n(\lambda) = n$ is treated separately)

$$\hat{R}_n^+(\lambda) \leq \alpha \iff \mathrm{Bin}_{n,\alpha+\varepsilon}\left(n\hat{R}_n(\lambda)\right) < \delta \quad \forall \varepsilon > 0 \text{ small enough}$$

$$\iff \mathrm{Bin}_{n,1-\alpha-\varepsilon}\left(n - n\hat{R}_n(\lambda) - 1\right) > 1-\delta \quad \forall \varepsilon > 0 \text{ small enough}$$

$$\iff \mathrm{Bin}_{n,1-\alpha}\left(n - n\hat{R}_n(\lambda) - 1\right) \geq 1-\delta$$

($p \mapsto \mathrm{Bin}_{n,p}(k)$ is continuous and decreasing in p, for $k \leq n-1$)

$$\iff n - n\hat{R}_n(\lambda) - 1 \geq \underbrace{\min\left\{ k \in \{0,\dots,n\} : \mathrm{Bin}_{n,1-\alpha}(k) \geq 1-\delta \right\}}_{= \text{quantile}_{1-\delta}\left(\mathrm{Bin}(n,1-\alpha)\right)}$$

$$\iff \sum_{i=1}^{n} \mathbb{1}_{Y_i \in C_\lambda(X_i)} \geq 1 + \text{quantile}_{1-\delta}\left(\mathrm{Bin}(n,1-\alpha)\right) =: k_{n,\alpha,\delta}$$

$$\iff \lambda \geq \lambda_{(k_{n,\alpha,\delta})} \quad \text{where} \begin{cases} \lambda_i := s(X_i,Y_i) := \min\{\lambda \in \Lambda : Y_i \in C_\lambda(X_i)\} \\ \lambda_{(1)} \leq \lambda_{(2)} \leq \dots \leq \lambda_{(n)} \end{cases}$$

(assuming existence)

Thus: $\underbrace{\inf\left\{ \lambda : \hat{R}_n^+(\lambda) \leq \alpha \right\}}_{\text{RCPS}} = \underbrace{\lambda_{(k_{n,\alpha,\delta})}}_{\text{tolerance region seen in Section 1.2.}}$

Appendix A: Some useful facts about Binomial and Beta distributions

We write:

- $t \in [0,1] \mapsto \mathrm{Beta}_{a,b}(t)$ for the cumulative distribution function (cdf) of the $\mathrm{Beta}(a,b)$ distribution ($a, b > 0$).
- $k \in \{0,1,\dots,n\} \mapsto \mathrm{Bin}_{n,p}(k)$ for the cdf of the binomial $\mathrm{Bin}(n,p)$ distribution ($n \geq 1$, $p \in [0,1]$).

Lemma (e.g., Task 2012, Lemma 3 in arXiv version)

$\forall n \geq 1$, $\forall k \in \{1,\dots,n\}$, $\forall p \in [0,1]$,

$$\mathrm{Bin}_{n,p}(k-1) = 1 - \mathrm{Beta}_{k,n+1-k}(p)$$

Proof: Let $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1])$, and set $U_{(1)} \leq \dots \leq U_{(n)}$.

We know that $X := \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq p\}} \sim \mathrm{Bin}(n,p)$ and $U_{(k)} \sim \mathrm{Beta}(k, n+1-k)$ (if P_{Beta} with $\nu = \mathrm{Unif}([0,1])$).

Therefore,

$$\begin{aligned} \mathrm{Bin}_{n,p}(k-1) &= 1 - \mathbb{P}(X \geq k) \\ &= 1 - \mathbb{P}(U_{(k)} \leq p) \\ &= 1 - \mathrm{Beta}_{k,n+1-k}(p). \end{aligned}$$

■

Fact: for $n \geq 1$ and $\alpha, \delta \in (0,1)$, we have:

$$\exists k \in \{1,\dots,n\} \text{ s.t. } \mathrm{quantile}_{\delta}(\mathrm{Beta}(k, n+1-k)) \geq 1-\alpha \qquad \text{(i)}$$

$$\Leftrightarrow \mathrm{quantile}_{\delta}(\mathrm{Beta}(n,1)) \geq 1-\alpha \qquad \text{(ii)}$$

$$\Leftrightarrow \alpha \geq 1 - \delta^{1/n} \qquad \text{(iii)}$$

$$\Leftarrow \alpha \geq \frac{\ln(1/\delta)}{n} \qquad \text{(iv)}$$

Proof: we have a natural coupling of all distributions $\mathrm{Beta}(k, n+1-k)$, $k \in \{1,\dots,n\}$, given by the order statistics $U_{(1)} \leq \dots \leq U_{(n)}$ with $U_1, \dots, U_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0,1])$

(recall that $U_{(k)} \sim \text{Beta}(k, n+1-k)$ for each $k \in \{1, \dots, n\}$)

Therefore,

$$\text{Beta}(n,1) = \text{law}(U_{(n)}) \overset{\text{st}}{\geq} \text{law}(U_{(k)}) = \text{Beta}(k, n+1-k)$$

so that

$$\text{quantile}_\delta(\text{Beta}(n,1)) \geq \text{quantile}_\delta(\text{Beta}(k,n+1-k))$$

This proves (i) $\Leftrightarrow$ (ii).

Furthermore,

$$\begin{aligned}
(ii) &\Leftrightarrow \text{Beta}_{n,1}(1-\alpha) \leq \delta \\
&\Leftrightarrow \mathbb{P}(U_{(n)} \leq 1-\alpha) \leq \delta \\
&\Leftrightarrow \mathbb{P}(\max_{1 \leq i \leq n} U_i \leq 1-\alpha) \leq \delta \\
&\Leftrightarrow (1-\alpha)^n \leq \delta \\
&\Leftrightarrow \alpha \geq 1 - \delta^{1/n}
\end{aligned}$$

which proves (ii) $\Leftrightarrow$ (iii).

Finally, note that

$$1 - \delta^{1/n} = 1 - \exp\left(-\frac{\ln(1/\delta)}{n}\right) \leq \frac{\ln(1/\delta)}{n}$$

so that (iv) $\Rightarrow$ (iii). ∎

Theorem (e.g., Thm 3.3. by Langford 2005): *A tight high-probability bound on p*

Let $n \geq 1$, $p^* \in [0,1]$, and $Z \sim \text{Bin}(n, p^*)$

Then, the r.v.

$$\hat{p}_n^+ := \sup\left\{ p \in [0,1] : \underbrace{\text{Bin}_{n,p}(Z)}_{\text{cdf of } \text{Bin}(n,p)} \geq \delta \right\}$$

satisfies:

$$\mathbb{P}(p^* \leq \hat{p}_n^+) \geq 1 - \delta.$$

Proof: since $Z \sim \mathrm{Bin}_{n,p^*}$, we have $\underbrace{\mathrm{Bin}_{n,p^*}(Z)}_{\text{cdf of } \mathrm{Bin}(n,p^*)} \overset{\text{sto}}{\geq} \mathrm{Unif}([0,1])$

so that $\mathbb{P}\left(\mathrm{Bin}_{n,p^*}(Z) \geq \delta\right) \geq 1-\delta$

Noting that $\left\{\mathrm{Bin}_{n,p^*}(Z) \geq \delta\right\} \subset \left\{p^* \leq \hat{p}_n^+\right\}$ concludes the proof. ∎

Lemma (concentration of beta distribution) [See Vovk'2012 for a somewhat equivalent statement]

Let $n \geq 1$, $\alpha \in \left[\frac{1}{n+1}, 1\right)$, $k := \lceil (n+1)(1-\alpha) \rceil \in \{1, \dots, n\}$.

Then, for $X \sim \mathrm{Beta}(k, n+1-k)$ and any $\delta \in (0,1)$,

$$\mathbb{P}\left(X \geq 1-\alpha - \sqrt{\frac{2\alpha \ln(1/\delta)}{n}} - \frac{2\ln(1/\delta)}{n}\right) \geq 1-\delta.$$

Proof (sketch): let $\varepsilon := \sqrt{\frac{2\alpha\ln(1/\delta)}{n}} + \frac{2\ln(1/\delta)}{n}$, and assume that $\alpha+\varepsilon<1$.

$$\begin{aligned}
\mathbb{P}(X \geq 1-\alpha-\varepsilon) &= 1 - \mathrm{Beta}_{k,n+1-k}(1-\alpha-\varepsilon) \\
&= \mathrm{Bin}_{n,1-\alpha-\varepsilon}(k-1) \quad \text{by a lemma above} \\
&= \mathbb{P}(S \leq k-1) \quad \text{for } S \sim \mathrm{Bin}(n, 1-\alpha-\varepsilon) \\
&= 1 - \mathbb{P}(S \geq k)
\end{aligned}$$

But,

$$\begin{aligned}
\mathbb{P}(S \geq k) &= \mathbb{P}(n-S \leq n-k) \\
&= \mathrm{Bin}_{n,\alpha+\varepsilon}(n-k) \\
&\leq \delta,
\end{aligned}$$

where the last inequality follows from $\mathrm{Bin}_{n,\alpha+\varepsilon}(n-k) \leq \mathrm{Bin}_{n,p}(n-k)$ with $p := \frac{n-k}{n} + \sqrt{\frac{2\cdot\frac{n-k}{n}\ln(1/\delta)}{n}} + \frac{2\ln(1/\delta)}{n} < \alpha+\varepsilon$ (by the choice of $k = \lceil (n+1)(1-\alpha) \rceil$) and from $\mathrm{Bin}_{n,p}(n-k) \leq \delta$ by Bernstein's inequality. ∎

References [authors (year), title]

Wilks (1941), Determination of sample sizes for setting tolerance limits.

Wald (1943), An extension of Wilks' method for setting tolerance limits.

Scheffé and Tukey (1945), Non-parametric estimation I. Validation of order statistics.

Tukey (1947), Non-parametric estimation II. Statistically equivalent blocks and tolerance regions – the continuous case.

Vovk (2012), Conditional validity of inductive conformal predictors.

Bates et al. (2021), Distribution-free, risk-controlling prediction sets.

Park et al. (2021), PAC confidence sets for DNNs via calibrated prediction.

Maurer and Pontil (2009), Empirical Bernstein bounds and sample variance penalization.

Langford (2005), Tutorial on practical prediction theory for classification