Notes on Conformal Risk Control

Catherine Chen cyc2152

November 2023

1 Distribution-free, Risk-controlling Prediction Sets

1.1 Setting and Notation

$(X_i, Y_i)_{i=1,\ldots,m} \sim$ i.i.d. s.t. features vectors $X_i \in \mathcal{X}$ and response $Y_i \in \mathcal{Y}$.

Split data: training and calibration set: $\{\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{cal}}\}$ form a partition of $\{1, \ldots, m\}$, with $n = |\mathcal{I}_{\text{cal}}|$. w.l.o.g., $\mathcal{I}_{\text{cal}} = \{1, \ldots, n\}$.

Fit predictive model on $\mathcal{I}_{\text{train}}$ denote $\hat{f}:\mathcal{X} \to \mathcal{Z}$.

Let $\mathcal{T} : \mathcal{X} \to \mathcal{Y}'$ be a set-valued function (a tolerance region) that maps a feature vector to a set-valued prediction typically constructed from the predictive model, $\hat{f}$. Suppose there exists a collection of such set-valued predictors indexed by a one-dimensional parameter λ taking values in a closed set $\Lambda \subset \mathbb{R} \cup \{\pm\infty\}$ that are nested, i.e. larger values of λ lead to larger sets:

$$\lambda_1 < \lambda_2 \Longrightarrow \mathcal{T}_{\lambda_1}(x) \subset \mathcal{T}_{\lambda_2}(x).$$

Note: $\lambda \to \infty \Rightarrow$ more conservative, i.e. larger set

Notion of error: $L(y, \mathcal{S}) : \boldsymbol{y} \times \boldsymbol{y'} \to \mathbb{R}_{\geq 0}$, loss function on prediction sets. i.e. $L(y, \mathcal{S}) = \mathbb{1}_{\{y \in \mathcal{S}\}}$. The loss function must satisfy the following nesting property:

$$\mathcal{S} \subset \mathcal{S}' \Longrightarrow L(y, \mathcal{S}) \geq L(y, \mathcal{S}').$$

That is, larger sets lead to smaller loss.

Note: $\lambda \to \infty \Rightarrow$ more conservative, i.e. larger set $\Rightarrow$ smaller loss

Define the risk of a set-valued predictor $\mathcal{T}$ to be

$$R(\mathcal{T}) = \mathbb{E}[L(Y, \mathcal{T}(X))]$$

Consider the risk of the tolerance functions from the family $\{\mathcal{T}_\lambda\}_{\lambda \in \Lambda}$.

$R(\lambda)$ is shorthand for $R(\mathcal{T}_\lambda)$.

Assume that there exists an element $\lambda_{\max} \in \Lambda$ such that $R(\lambda_{\max}) = 0$.

1.2 Procedure

Goal: find a set function whose risk is less than some user-specified threshold α. Analyze collection of functions $\{\mathcal{T}_\lambda\}_{\lambda \in \mathcal{T}}$ and estimate their risk on data not used for model training, $\mathcal{I}_{\text{cal}}$. Then show that by choosing the value of λ in a certain way, we can guarantee that the procedure has risk less than α with high probability.

Pointwise upper confidence bound (UCB) for the risk function for each λ:

$$P(R(\lambda) \leq \underbrace{\widehat{R}^{+}(\lambda)}_{\text{UCB}}) \geq 1 - \delta$$

where $\widehat{R}^{+}(\lambda)$ may depend on $(X_1, Y_1), \ldots, (X_n, Y_n)$. Choose $\hat{\lambda}$ as the smallest value of λ s.t. the entire confidence region to the right of λ falls below the target risk level α :

$$\hat{\lambda} \triangleq \inf \left\{ \lambda \in \Lambda : \widehat{R}^{+}\left(\lambda'\right) < \alpha, \forall \lambda' \geq \lambda \right\}$$

1.3 Simplified Hoeffding Bound

1.3.1 Theorem 1: Validity of UCB Calibration

Let $(X_i, Y_i)_{i=1,\ldots,n}$ *be an i.i.d. sample, let* $L(\cdot,\cdot)$ *be a loss satisfying the monotonicity condition:*

$$\mathcal{S} \subset \mathcal{S}' \Longrightarrow L(y, \mathcal{S}) \geq L(y, \mathcal{S}'),$$

and let $\{\mathcal{T}_\lambda\}_{\lambda\in\Lambda}$ *be a collection of set predictors satisfying the nesting property in*

$$\lambda_1 < \lambda_2 \Longrightarrow \mathcal{T}_{\lambda_1}(x) \subset \mathcal{T}_{\lambda_2}(x).$$

Let $R : \Lambda \to \mathbb{R}$ *be a continuous monotone nonincreasing function such that* $R(\lambda) \leq \alpha$ *for some* $\lambda \in \Lambda$. *Suppose* $\widehat{R}^+(\lambda)$ *is a random variable for each* $\lambda \in \Lambda$ *such that*

$$P(R(\lambda) \leq \underbrace{\widehat{R}^+(\lambda)}_{\text{UCB}}) \geq 1 - \delta$$

holds pointwise for each λ. *Then, for* $\hat{\lambda} \triangleq \inf\left\{\lambda \in \Lambda : \widehat{R}^+(\lambda') < \alpha, \forall \lambda' \geq \lambda\right\}$,

$$P\left(R\left(\mathcal{T}_{\hat{\lambda}}\right) \leq \alpha\right) \geq 1 - \delta$$

That is, $\mathcal{T}_{\hat{\lambda}}$ *is a* $(\alpha, \delta) - RCPS$.

Proof. Consider the smallest λ that controls the risk:

$$\lambda^* \triangleq \inf\{\lambda \in \Lambda : R(\lambda) \leq \alpha\}$$

Suppose $R(\hat{\lambda}) > \alpha \implies \hat{\lambda} < \lambda^*$ by the definition of λ^* and the monotonicity and continuity of $R(\cdot)$.

Then $R(\hat{\lambda}) > \alpha \implies \hat{\lambda} < \lambda^* \implies \widehat{R}^+(\lambda^*) < \alpha$ by the definition of $\hat{\lambda}$.

But, since $R(\lambda^*) = \alpha$ (by continuity) and by the coverage property

$$P(R(\lambda) \leq \underbrace{\widehat{R}^+(\lambda)}_{\text{UCB}}) \geq 1 - \delta,$$

this happens with probability at most δ since the coverage property implies

$$P(R(\hat{\lambda}) > \widehat{R}^+(\lambda)) < \delta \implies P(R(\hat{\lambda}) > \alpha > \widehat{R}^+(\lambda^*)) < \delta \implies P(R(\hat{\lambda}) > \alpha) < \delta \implies P(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta$$

1.3.2 Hoeffding's Inequality

Suppose the loss is bounded above by one. Then,

$$P(\widehat{R}(\lambda) - R(\lambda) \leq -x) \leq \exp\left\{-2nx^2\right\}.$$

This implies an upper confidence bound

$$\widehat{R}^+_{sHoef}(\lambda) = \widehat{R}(\lambda) + \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}.$$

*Applying **Theorem 1** with*

$$\begin{aligned}\hat{\lambda} = \hat{\lambda}^{sHoef} &\triangleq \inf\left\{\lambda \in \Lambda : \widehat{R}^+_{\text{sHoef}}(\lambda') < \alpha, \forall \lambda' \geq \lambda\right\} \\ &= \inf\left\{\lambda \in \Lambda : \widehat{R}(\lambda) < \alpha - \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}\right\},\end{aligned}$$

we can generate an RCPS.

1.3.3 Theorem 2: RCPS from Hoeffding's Inequality

In the setting of Theorem 1, assume also that the loss is bounded by one. Then, $\mathcal{T}_{\hat{\lambda}\ sHoef}$ *is a* $(\alpha, \delta) - RCPS$.

1.4 Hoeffding-Bentkus Bound

In general, a UCB can be obtained if the lower tail probability of $\widehat{R}(\lambda)$ can be controlled, which is nearly tight for binary loss function.

1.4.1 Proposition 2:

Suppose $g(t;R)$ is a nondecreasing function in $t \in \mathbb{R}$ for every R :

$$P(\widehat{R}(\lambda) \le t) \le g(t; R(\lambda))$$

Then, $\widehat{R}^{+}(\lambda) = \sup\{R : g(\widehat{R}(\lambda); R) \ge \delta\}$ satisfies

$$P(R(\lambda) \le \hat{R}^{+}(\lambda)) \ge 1 - \delta.$$

This result shows how a tail probability bound can be inverted to yield a UCB. Thus $g(\widehat{R}(\lambda); R)$ is a conservative p-value for testing the one-sided null hypothesis $H_0 : R(\lambda) \ge R$.

Proof. Let G denote the CDF of $R(\lambda)$.

If $R(\lambda) > R^{+}(\lambda)$, then by definition, $g(\widehat{R}(\lambda); R(\lambda)) < \delta$, since $\widehat{R}^{+}(\lambda) = \sup\{R : g(\widehat{R}(\lambda); R) \ge \delta\}$.

As a result,

$$P(R(\lambda) > \widehat{R}^{+}(\lambda)) \le P(g(\widehat{R}(\lambda); R(\lambda)) < \delta) \le P(G(\widehat{R}(\lambda)) < \delta).$$

Let $G^{-1}(\delta) = \sup\{x : G(x) \le \delta\}$. Then,

$$P(G(\widehat{R}(\lambda)) < \delta) \le P(\widehat{R}(\lambda) < G^{-1}(\delta)) \le \delta.$$

This implies that $P(R(\lambda) > \widehat{R}^{+}(\lambda)) \le \delta$ and completes the proof.

1.4.2 Proposition 3: Hoeffding's Inequality Tighter Version

Suppose the loss is bounded above by one. Then, for any $t < R(\lambda)$,

$$P(\widehat{R}(\lambda) \le t) \le \exp\{-nh_1(t; R(\lambda))\}$$

where $h_1(t;R) = t\log(t/R) + (1-t)\log((1-t)/(1-R))$.

Note: The weaker Hoeffding inequality is implied by Proposition 3 using the fact that $h_1(t;R) \ge 2(t-R)^2$.

1.4.3 Proposition 4: Bentkus' Inequality

Suppose the loss is bounded above by one. Then,

$$P(\widehat{R}(\lambda) \le t) \le eP(\mathrm{Binom}(n, R(\lambda)) \le \lceil nt \rceil),$$

where $\mathrm{Binom}(n,p)$ denotes a binomial random variable with sample size n and success probability p.

Note: Bentkus inequality implies that the Binomial distribution is the worst case up to a small constant. The Bentkus inequality is nearly tight if the loss function is binary, in which case $n\widehat{R}(\lambda)$ is binomial.

Putting **Propositions 3** and **4** together, we obtain a lower tail probability bound for $\widehat{R}(\lambda)$:

$$g^{\mathrm{HB}}(t; R(\lambda)) \triangleq \min\left(\exp\{-nh_1(t; R(\lambda))\}, eP(\mathrm{Binom}(n, R(\lambda)) \le \lceil nt \rceil)\right).$$

By **Proposition 2**, we obtain a $(1-\delta)$ upper confidence bound for $R(\lambda)$ as

$$\widehat{R}^{+}_{\mathrm{HB}}(\lambda) = \sup\left\{R : g^{\mathrm{HB}}(\widehat{R}(\lambda); R) \ge \delta\right\}.$$

1.4.4 Theorem 3: RCPS from the Hoeffding-Bentkus Bound

In the setting of Theorem 1, assume additionally that the loss is bounded by one. Obtain $\hat{\lambda}^{\mathrm{HB}}$ *from* $\widehat{R}^{+}_{\mathrm{HB}}(\lambda)$ *as* $\hat{\lambda} \triangleq \inf\left\{\lambda \in \Lambda : \widehat{R}^{+}\left(\lambda'\right) < \alpha, \forall \lambda' \geq \lambda\right\}$*. Then,* $\mathcal{T}_{\hat{\lambda}^{\mathrm{HB}}}$ *is a* (α, δ)*-RCPS.*

1.5 Waudby-Smith-Ramdas Bound

For non-binary loss functions, and bound that is adaptive to the variance via online inference and martingale analysis.

1.5.1 Proposition 5 (Waudby-Smith-Ramdas Bound)

Let $L_i(\lambda) = L\left(Y_i, T_\lambda\left(X_i\right)\right)$ *and*

$$\hat{\mu}_i(\lambda) = \frac{1/2 + \sum_{j=1}^{i} L_j(\lambda)}{1+i}, \hat{\sigma}_i^2(\lambda) = \frac{1/4 + \sum_{j=1}^{i} \left(L_j(\lambda) - \hat{\mu}_j(\lambda)\right)^2}{1+i}, v_i(\lambda) = \min\left\{1, \sqrt{\frac{2\log(1/\delta)}{n\hat{\sigma}_{i-1}^2(\lambda)}}\right\}.$$

Further, let

$$\mathcal{K}_i(R;\lambda) = \prod_{j=1}^{i}\left\{1 - v_j(\lambda)\left(L_j(\lambda) - R\right)\right\}, \quad \widehat{R}_{\mathrm{WSR}}^{+}(\lambda) = \inf\left\{R \geq 0 : \max_{i=1,\ldots,n} \mathcal{K}_i(R;\lambda) > \frac{1}{\delta}\right\}.$$

Then, $\widehat{R}_{\mathrm{WSR}}^{+}(\lambda)$ *is a*$(1-\delta)$ *upper confidence bound for* $R(\lambda)$.

Proof. Let $\mathcal{K}_i = \mathcal{K}_i(R(\lambda);\lambda), \mathcal{F}_0$ be the trivial sigma-field and $\mathcal{F}_i$ be the sigma-field generated by $(L_1(\lambda), \ldots, L_i(\lambda))$. Then, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_n$ is a filtration. By definition, $v_i(\lambda) \in \mathcal{F}_{i-1}$ is a predictable sequence and $\mathcal{K}_i \in \mathcal{F}_i$. Since $\mathbb{E}\left[L_i(\lambda)\right] = R(\lambda)$,

$$\mathbb{E}\left[\mathcal{K}_i \mid \mathcal{F}_{i-1}\right] = \mathbb{E}\left[\mathcal{K}_{i-1}(1 - v_i(\lambda)\left(L_i(\lambda) - R(\lambda)\right)) \mid \mathcal{F}_{i-1}\right] = \mathcal{K}_{i-1}\mathbb{E}\left[1 - v_i(\lambda)\left(L_i(\lambda) - R(\lambda)\right) \mid \mathcal{F}_{i-1}\right] = \mathcal{K}_{i-1}$$

In addition, since $v_i \in [0,1]$ and $(L_i(\lambda) - R(\lambda)) \in [-1,1]$, each component $1 - v_i(\lambda)\left(L_i(\lambda) - R(\lambda)\right) \geq 0$. Thus, $\{\mathcal{K}_i : i = 1, \ldots, n\}$ is a non-negative martingale with respect to the filtration $\{\mathcal{F}_i : i = 1, \ldots, n\}$.

Ville's Inequality
Let $X_0, X_1, X_2, \ldots$ *be a non-negative supermartingale. Then, for any real number* $a > 0$,

$$\mathrm{P}\left[\sup_{n\geq 0} X_n \geq a\right] \leq \frac{\mathrm{E}\left[X_0\right]}{a}$$

By Ville's inequality,

$$P\left(\max_{i=1,\ldots,n} \mathcal{K}_i \geq \frac{1}{\delta}\right) \leq \delta.$$

However, since $v_i \geq 0, \mathcal{K}_i(R;\lambda)$ is increasing in R almost surely for every i. By definition of $\widehat{R}_{\mathrm{WSR}}^{+}(\lambda)$, if $\widehat{R}_{\mathrm{WSR}}^{+}(\lambda) < R(\lambda)$, then $P\left(\max_{i=1,\ldots,n} \mathcal{K}_i \geq 1/\delta\right)$. Therefore,

$$P\left(\widehat{R}_{\mathrm{WSR}}^{+}(\lambda) < R(\lambda)\right) \leq P\left(\max_i \mathcal{K}_i \geq \frac{1}{\delta}\right) \leq \delta.$$

This proves that $\widehat{R}_{\mathrm{WSR}}^{+}(\lambda)$ is a valid upper confidence bound of $R(\lambda)$.

1.5.2 Theorem 4: RCPS From the Waudby-Smith-Ramdas Bound

In the setting of Theorem 1, assume additionally that the loss is bounded by 1. Then, $\mathcal{T}_{\hat{\lambda}\mathrm{WSR}}$ *is a*$(\alpha,\delta)-$ *RCPS.*

1.6 Unbounded Losses

1.6.1 Proposition A.1 (Impossibility of Valid UCB for Unbounded Losses in Finite Samples)

Let $\mathcal{F}$ be the class of all distributions supported on $[0, \infty)$ with finite mean, and $\mu(F)$ be the mean of the distribution F. Let $\hat{\mu}^+$ be any function of $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} F$ such that $P\left(\hat{\mu}^+ \geq \mu(F)\right) \geq 1 - \delta$ for any n and $F \in \mathcal{F}$. Then, $P\left(\hat{\mu}^+ = \infty\right) \geq 1 - \delta$.

Proof. It is clear that $\mathcal{F}$ satisfies the conditions

1. For every $F \in \mathcal{F}$, $\mu_F = \int_{-\infty}^{\infty} z dF$ exists and is finite.
2. For every real m, there is an $F \in \mathcal{F}$ with $\mu_F = m$.
3. $\mathcal{F}$ is convex, that is, if F , $G \in \mathcal{F}$, π is a positive fraction, and $H = \pi F(1 - \pi)G$ then $H \in \mathcal{F}$.

If $P_F(C[\mu_F]) \geq 1 - \delta$ for all $F \in \mathcal{F}$, then $P_F(C[\mu_F]) \geq 1 - \delta$ for all m and all $F \in \mathcal{F}$. For any such $\hat{\mu}^+$, $[0, \hat{\mu}^+]$is a $(1 - \delta)$ confidence interval of $\mu(F)$. By their Corollary 2, we know that for any $\mu \in \{\mu(F) : F \in \mathcal{F}\}$ and $F \in \mathcal{F}$

$$P_F\left(\mu \in \left[0, \hat{\mu}^+\right]\right) \geq 1 - \delta \Longleftrightarrow P_F\left(\mu \leq \hat{\mu}^+\right) \geq 1 - \delta.$$

The proof is complete by letting $\mu \to \infty$.