Notes on Conformal Risk Control

Catherine Chen cyc2152

December 2023

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal Risk Control." arXiv, Apr. 29, 2023. Accessed: Dec. 14, 2023. [Online]. Available: http://arxiv.org/abs/2208.02814

Abstract: Control the expected value of any monotone loss function via conformal prediction, that is tight up to an $O(1/n)$ factor.

1 Introduction

1.1 Setting and Notation

Consider a *calibration dataset* $(X_i, Y_i)_{i=1,\ldots,n} \sim$ i.i.d. s.t. features vectors $X_i \in \mathcal{X}$ and response $Y_i \in \mathcal{Y}$. Conformal prediction seeks to bound the *miscoverage*, of a new test point (X_{n+1}, Y_{n+1})

$$\mathbb{P}\left(Y_{n+1} \notin \mathcal{C}\left(X_{n+1}\right)\right) \leq \alpha \tag{1}$$

where α is a user-specified error rate, and $\mathcal{C}$ is a function of the model and calibration data that produces the prediction set.

Objective: Provide guarantee called, *conformal risk control*, of the form

$$\mathbb{E}\left[\ell\left(\mathcal{C}\left(X_{n+1}\right), Y_{n+1}\right)\right] \leq \alpha, \tag{2}$$

for any bounded *loss function* ℓ that has an inverse relation with $\mathcal{C}(X_{n+1})$.

Remark: Recover conformal miscoverage guarantee with $\ell\left(\mathcal{C}\left(X_{n+1}\right), Y_{n+1}\right) = \mathbb{1}\left\{Y_{n+1} \notin \mathcal{C}\left(X_{n+1}\right)\right\}$.

Conformal risk control seeks to find a threshold $\hat{\lambda}$ that controls the proportion of missed classes:

$$\mathbb{E}\left[\ell\left(\mathcal{C}_{\hat{\lambda}}\left(X_{n+1}\right), Y_{n+1}\right)\right] = \mathbb{E}\left[1 - \frac{\left|Y_{n+1} \cap \mathcal{C}_{\lambda}\left(X_{n+1}\right)\right|}{\left|Y_{n+1}\right|}\right]$$

Note: the threshold λ will be defined s.t. it increases with $|\mathcal{C}_\lambda(x)|$; in other words, as λ grows, $\mathcal{C}_\lambda(x)$ becomes more conservative.

1.2 Algorithm and Preview of Main Results

Given base model f, post-process the predictions to produce a function $\mathcal{C}_\lambda(\cdot)$. The quality of the output of $\mathcal{C}_\lambda$ will be quantified by a loss function $\ell\left(\mathcal{C}_\lambda(x), y\right) \in (-\infty, B]$ for some $B < \infty$, that is a non-increasing as function of λ.

Goal: Choose $\hat{\lambda}$ based on the observed data $\{(X_i, Y_i)\}_{i=1}^n$ s.t. the risk control in (2) holds.

Consider an exchangeable collection of non-increasing, random functions $L_i : \Lambda \to (-\infty, B], i = 1, \ldots, n+1$. Assume $\lambda_{\max} \triangleq \sup \Lambda \in \Lambda$. Use the first n functions to choose a value of the parameter, $\hat{\lambda}$, s.t. the risk on the unseen function is controlled:

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \leq \alpha \tag{3}$$

Let $\widehat{R}_n(\lambda) = \frac{(L_1(\lambda)+\ldots+L_n(\lambda))}{n}$. Given any desired risk level upper bound $\alpha \in (-\infty, B)$, define

$$\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}. \tag{4}$$

When the set is empty, take $\hat{\lambda} = \lambda_{\max}$.

The proposed conformal risk control algorithm deploys $\hat{\lambda}$ on the new test point to achieve the guarantee in (3).

When the L_i are i.i.d. from a continuous distribution, the algorithm satisfies a tight lower bound illustrating that it is not too conservative,

$$\alpha - \frac{2B}{n+1} \leq \mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \leq \alpha.$$

2 Theory

2.1 Risk Control

Theorem 1. *Assume that $L_i(\lambda)$ is non-increasing in λ, right-continuous, and*

$$L_i(\lambda_{\max}) \leq \alpha, \quad \sup_\lambda L_i(\lambda) \leq B < \infty \text{ almost surely.} \tag{5}$$

Then

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \leq \alpha$$

Proof. Let $\widehat{R}_{n+1}(\lambda) = \frac{(L_1(\lambda)+\ldots+L_{n+1}(\lambda))}{n+1}$ and

$$\hat{\lambda}' = \inf\left\{\lambda \in \Lambda : \widehat{R}_{n+1}(\lambda) \leq \alpha\right\}.$$

Since $L_i(\lambda)$ is non-increasing in λ: $\inf_\lambda L_i(\lambda) = L_i(\lambda_{\max}) \leq \alpha$, thus $\hat{\lambda}'$ is well-defined almost surely.

By assumption $L_{n+1}(\lambda) \leq B$, we have $\widehat{R}_{n+1}(\lambda) = \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{L_{n+1}(\lambda)}{n+1} \leq \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1}$. Thus,

$$\left(\widehat{R}_{n+1}(\lambda) \leq\right) \underbrace{\frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha}_{\inf\{LHS\} \implies \hat{\lambda}} \implies \underbrace{\widehat{R}_{n+1}(\lambda) \leq \alpha}_{\inf\{RHS\} \implies \hat{\lambda}'}.$$

Since $\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}$, taking the infimum on both sides implies $\hat{\lambda}' \leq \hat{\lambda}$ when the LHS holds for some $\lambda \in \Lambda$.

When the LHS is above α for all $\lambda \in \Lambda$, the set, $\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}$ is empty, thus by definition, we take $\hat{\lambda} = \lambda_{\max} \implies \hat{\lambda} \geq \hat{\lambda}'$. Thus, $\hat{\lambda}' \leq \hat{\lambda}$ almost surely. Since $L_i(\lambda)$ is non-increasing in λ,

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \leq \mathbb{E}\left[L_{n+1}\left(\hat{\lambda}'\right)\right] \tag{6}$$

Let E be the multiset of loss functions $\{L_1, \ldots, L_{n+1}\}$. Then $\hat{\lambda}'$ is a function of $E \iff \hat{\lambda}'$ is a constant conditional on E. Additionally, $L_{n+1}(\lambda) \mid E \sim \text{Uniform}(\{L_1, \ldots, L_{n+1}\})$ by exchangeability. Combined with the right-continuity of L_i, it can be shown that

$$\mathbb{E}\left[L_{n+1}\left(\hat{\lambda}'\right) \mid E\right] = \frac{1}{n+1}\sum_{i=1}^{n+1} L_i\left(\hat{\lambda}'\right) \leq \alpha.$$

By the law of total expectation and (6), the proof is complete. ☐

2.2 A tight risk lower bound

Definition 2.1 (Jump Function)**.** Quantifies the size of the discontinuity in a right-continuous function $l(\lambda)$:

$$J(l,\lambda)=\lim_{\epsilon\to 0^+} l(\lambda-\epsilon)-l(\lambda)$$

Note:

$$J(l,\lambda)\begin{cases}>0 & \text{there is a discontinuity}\\ =0 & \text{there is no discontinuity}\end{cases}$$

Lemma 1 (Jump Lemma)**.** *Assume that $L_i(\lambda)$ is non-increasing in λ, right-continuous, and*

$$L_i\left(\lambda_{\max}\right)\leq\alpha,\ \sup_\lambda L_i(\lambda)\leq B<\infty \text{ almost surely,}$$

further assume that the L_i are i.i.d., $L_i\geq 0$, and for any λ, $\mathbb{P}\left(J\left(L_i,\lambda\right)>0\right)=0$, then any jumps in the empirical risk are bounded, i.e.,

$$\sup_\lambda J\left(\widehat{R}_n,\lambda\right)\overset{a.s.}{\leq}\frac{B}{n}.$$

Proof. By boundedness, the maximum contribution of any single point to the jump is $\frac{B}{n}$, so

$$\exists\lambda: J\left(\widehat{R}_n,\lambda\right)>\frac{B}{n}\Longrightarrow\exists\lambda: J\left(L_i,\lambda\right)>0 \text{ and } J\left(L_j,\lambda\right)>0 \text{ for some } i\neq j.$$

Let $\mathcal{D}_i=\left\{\lambda: J\left(L_i,\lambda\right)>0\right\}$ denote the sets of discontinuities in L_i. Since L_i is bounded monotone, $\mathcal{D}_i$ has countably many points.

Boole's Inequality/Union Bound states that:

For a countable set of events $A_1,A_2,A_3,\ldots$, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty}A_i\right)\leq\sum_{i=1}^{\infty}\mathbb{P}\left(A_i\right).$$

This means that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events.

This implies that

$$\mathbb{P}\left(\exists\lambda: J\left(\widehat{R}_n,\lambda\right)>\frac{B}{n}\right)\leq\sum_{i\neq j}\mathbb{P}\left(\mathcal{D}_i\cap\mathcal{D}_j\neq\emptyset\right)$$

Rewriting each term of the right-hand side using tower property and law of total probability gives

$$\begin{aligned}\mathbb{P}\left(\mathcal{D}_i\cap\mathcal{D}_j\neq\emptyset\right)&=\mathbb{E}\left[\mathbb{P}\left(\mathcal{D}_i\cap\mathcal{D}_j\neq\emptyset\mid\mathcal{D}_j\right)\right]\\&\leq\mathbb{E}\left[\sum_{\lambda\in\mathcal{D}_j}\mathbb{P}\left(\lambda\in\mathcal{D}_i\mid\mathcal{D}_j\right)\right]=\mathbb{E}\left[\sum_{\lambda\in\mathcal{D}_j}\mathbb{P}\left(\lambda\in\mathcal{D}_i\right)\right],\end{aligned}$$

Where the second inequality is because the union of the events $\lambda\in\mathcal{D}_j$ is the entire sample space, but they are not disjoint, and the third equality is due to the independence between $\mathcal{D}_i$ and $\mathcal{D}_j$. Applying the assumption $\mathbb{P}\left(J\left(L_i,\lambda\right)>0\right)=0$,

$$\mathbb{E}\left[\sum_{\lambda\in\mathcal{D}_j}\mathbb{P}\left(\lambda\in\mathcal{D}_i\right)\right]=\mathbb{E}\left[\sum_{\lambda\in\mathcal{D}_j}\mathbb{P}\left(J\left(L_i,\lambda\right)>0\right)\right]=0.$$

Thus we have that

$$\mathbb{P}\left(\exists\lambda: J\left(\widehat{R}_n,\lambda\right)>\frac{B}{n}\right)\leq 0\implies\mathbb{P}\left(\exists\lambda: J\left(\widehat{R}_n,\lambda\right)>\frac{B}{n}\right)=0$$

□

Theorem 2. *In the setting of Theorem 1, further assume that the L_i are i.i.d., $L_i \geq 0$, and for any λ, $\mathbb{P}\left(J\left(L_i, \lambda\right) > 0\right) = 0$. Then the conformal risk control procedure is not too conservative:*

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \geq \alpha - \frac{2B}{n+1}$$

Proof. If $L_i\left(\lambda_{\max}\right) \geq \alpha - \frac{2B}{n+1}$, then $\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \geq \alpha - \frac{2B}{n+1}$.

Assume that $L_i\left(\lambda_{\max}\right) < \alpha - \frac{2B}{n+1}$. Define

$$\hat{\lambda}'' = \inf\left\{\lambda : \widehat{R}_{n+1}(\lambda) + \frac{B}{n+1} \leq \alpha\right\}.$$

Since $L_i\left(\lambda_{\max}\right) < \alpha - \frac{2B}{n+1} < \alpha - \frac{B}{n+1}$, $\hat{\lambda}''$ exists almost surely.

Deterministically,

$$\frac{(L_1(\lambda) + \ldots + L_n(\lambda))}{n} \leq \frac{(L_1(\lambda) + \ldots + L_n(\lambda))}{n+1} \implies (n)\widehat{R}_n(\lambda) \leq (n+1)\widehat{R}_{n+1}(\lambda) \implies \frac{n}{n+1}\widehat{R}_n(\lambda) \leq \widehat{R}_{n+1}(\lambda)$$

Since,

$$\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\},$$

this yields that $\hat{\lambda} \leq \hat{\lambda}''$.

Since $L_i(\lambda)$ is non-increasing in λ,

$$\mathbb{E}\left[L_{n+1}\left(\hat{\lambda}''\right)\right] \leq \mathbb{E}\left[L_{n+1}(\hat{\lambda})\right]$$

By exchangeability and that $\hat{\lambda}''$ is a symmetric function of $L_1, \ldots, L_{n+1}$,

$$\mathbb{E}\left[L_{n+1}\left(\hat{\lambda}''\right)\right] = \mathbb{E}\left[\widehat{R}_{n+1}\left(\hat{\lambda}''\right)\right]$$

Now, find the lower-bound for $\widehat{R}_{n+1}\left(\hat{\lambda}''\right)$, we have that:

$$\alpha = \widehat{R}_{n+1}\left(\hat{\lambda}''\right) + \frac{B}{n+1} - \left(\widehat{R}_{n+1}\left(\hat{\lambda}''\right) + \frac{B}{n+1} - \alpha\right).$$

Equivalently:

$$\widehat{R}_{n+1}\left(\hat{\lambda}''\right) = \alpha - \frac{B}{n+1} + \left(\widehat{R}_{n+1}\left(\hat{\lambda}''\right) + \frac{B}{n+1} - \alpha\right).$$

By the Jump Lemma, bounding $\left(\widehat{R}_{n+1}\left(\hat{\lambda}''\right) + \frac{B}{n+1} - \alpha\right)$ below by $-\frac{B}{n+1}$ gives

$$\widehat{R}_{n+1}\left(\hat{\lambda}''\right) \geq \alpha - \frac{2B}{n+1}.$$

Finally,

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \geq \mathbb{E}\left[L_{n+1}\left(\hat{\lambda}''\right)\right] \geq \mathbb{E}\left[\widehat{R}_{n+1}\left(\hat{\lambda}''\right)\right] \geq \alpha - \frac{2B}{n+1}$$

□

Proposition 1. *In the setting of Theorem 2, for any $\epsilon > 0$, there exists a loss function and $\alpha \in (0,1)$ s.t.*

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \leq \alpha - \frac{2B-\epsilon}{n+1}$$

Proof. w.l.o.g., assume $B = 1$. Fix any $\epsilon' > 0$. Consider the following loss functions, which satisfy the conditions in Theorem 2 :

$$L_i(\lambda) \overset{\text{i.i.d.}}{\sim} \begin{cases} 1 & \lambda \in [0, Z_i) \\ \frac{k}{k+1} & \lambda \in [Z_i, W_i) \\ 0 & \text{else} \end{cases},$$

where $k \in \mathbb{N}$, the $Z_i \overset{i.i.d.}{\sim} \text{Uniform}(0, \frac{1}{2})$, the $W_i \overset{i.i.d.}{\sim} \text{Uniform}(\frac{1}{2}, 1)$ for $i \in \{1, \ldots, n+1\}$ and $\alpha = \frac{k+1-\epsilon'}{n+1}$.

Since $\hat{\lambda} = \inf\{\lambda : \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\}$, this implies that

$$\widehat{R}_n(\hat{\lambda}) \leq \frac{k-\epsilon'}{n}. \tag{7}$$

If $n > k+1$, and $\lambda \leq \frac{1}{2}$, $\widehat{R}(\lambda) \geq \frac{k}{k+1} > \frac{k}{n}$ by our definition of $L_i(\lambda)$. Thus, it must be the case that $\hat{\lambda} > \frac{1}{2}$.

Since $k \in \mathbb{N}$, and $\hat{R}_n(\lambda) = \frac{L_1(\lambda)+\cdots+L_n(\lambda)}{n}$, by (7), we know that $\left|\left\{i \in \{1, \ldots, n\} : L_i(\hat{\lambda}) > 0\right\}\right| \leq \left\lfloor \frac{(k+1)\left(k-\epsilon'\right)}{k} \right\rfloor \leq k$. This immediately implies that

$$\hat{\lambda} \geq W_{(n-k+1)}$$

where $W_{(j)}$ denotes the j-th order statistic. Notice that for all $\lambda > \frac{1}{2}$,

$$R(\lambda) = \mathbb{E}\left[L_i(\lambda)\right] = \frac{k}{k+1}\mathbb{P}\left(W_i > \lambda\right) = \frac{k}{k+1} \cdot 2(1-\lambda),$$

so $R(\hat{\lambda}) \leq \frac{k}{k+1} \cdot 2\left(1 - W_{(n-k+1)}\right)$. Let $U_{(k)}$ be the k-th smallest order statistic of n i.i.d. uniform random variables on $(0,1)$. Then, by symmetry and rescaling, $2\left(1 - W_{(n-k+1)}\right) \overset{d}{=} U_{(k)}$,

$$R(\hat{\lambda}) \preceq \frac{k}{k+1} U_{(k)},$$

where $\preceq$ denotes the stochastic dominance. It is well-known that $U_{(k)} \sim \text{Beta}(k, n+1-k)$ and hence

$$\mathbb{E}[R(\hat{\lambda})] \leq \frac{k}{k+1} \cdot \frac{k}{n+1}.$$

Thus,

$$\alpha - \mathbb{E}[R(\hat{\lambda})] \geq \frac{k+1-\epsilon}{n+1} - \frac{k^2}{(n+1)(k+1)} = \frac{1}{n+1} \cdot \frac{(2-\epsilon')\,k + 1 - \epsilon'}{k+1}.$$

For any given $\epsilon > 0$, let $\epsilon' = \epsilon/2$ and $k = \left\lceil \frac{2}{\epsilon} - 1 \right\rceil$. Then

$$\frac{(2-\epsilon')\,k + 1 - \epsilon'}{k+1} \geq 2 - \epsilon,$$

implying that

$$\alpha - \mathbb{E}[R(\hat{\lambda})] \geq \frac{2-\epsilon}{n+1}$$

□

2.3 Conformal prediction reduces to risk control

The risk lower bround in Theorem 2 has a slightly worse constant than the usual conformal guarantee.

Consider conformal scores $s\left(X_i, Y_i\right)$ for some score function $s: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$.

Rrediction sets for the test point X_{n+1} are constructed as

$$\mathcal{C}_{\hat{\lambda}}\left(X_{n+1}\right)=\left\{y: s\left(X_{n+1}, y\right) \leq \hat{\lambda}\right\},$$

where $\hat{\lambda}$ is the conformal quantile, i.e. the $\frac{\lceil(n+1)(1-\alpha)\rceil}{n}$ sample quantile of $\left\{s\left(X_i, Y_i\right)\right\}_{i=1}^n$.

Define the miscoverage loss $L_i^{\mathrm{Cvg}}(\lambda)=\mathbb{1}\left\{Y_i \notin \widehat{\mathcal{C}}_\lambda\left(X_i\right)\right\}=\mathbb{1}\left\{s\left(X_i, Y_i\right)>\lambda\right\}$

Recall:

$$\hat{\lambda}=\inf \left\{\lambda: \frac{n}{n+1} \widehat{R}_n(\lambda)+\frac{B}{n+1} \leq \alpha\right\} ;\ \widehat{R}_n(\lambda)=\frac{\left(L_1(\lambda)+\ldots+L_n(\lambda)\right)}{n}$$

$$\hat{\lambda}=\inf \left\{\lambda: \frac{1}{n+1} \sum_{i=1}^n \underbrace{\mathbb{1}\left\{s\left(X_i, Y_i\right)>\lambda\right\}}_{L_i(\lambda)}+\frac{1}{n+1} \leq \alpha\right\}=\underbrace{\inf \left\{\lambda: \frac{1}{n} \sum_{i=1}^n \mathbb{1}\left\{s\left(X_i, Y_i\right) \leq \lambda\right\} \geq \frac{\lceil(n+1)(1-\alpha)\rceil}{n}\right\}}_{\text{conformal prediction algorithm}} .$$

Remark: $B=1$ here since we are using a binary loss function, and $\ell\left(\mathcal{C}_\lambda(x), y\right) \in(-\infty, B]$ for some $B<\infty$.

From Theorem 2, we have that with the defined miscoverage loss function,

$$\mathbb{P}\left(Y_{n+1} \notin \mathcal{C}_{\hat{\lambda}}\left(X_{n+1}\right)\right) \geq \alpha-\frac{2}{n+1}$$

Note: Distribution of discontinuities specializes to the continuity of the score function when the miscoverage loss is used:

$$\mathbb{P}\left(J\left(L_i^{\mathrm{Cvg}}, \lambda\right)>0\right)=0 \Longleftrightarrow \mathbb{P}\left(s\left(X_i, Y_i\right)=\lambda\right)=0 .$$

2.4 Controlling general loss function

Conformal risk control does *not* control the risk if the L_i are not monotone.

Proposition 2. *For any ϵ, there exists a non-monotone loss function such that*

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \geq B - \epsilon$$

Consider an exchangeable collection of non-increasing, random functions $L_i : \Lambda \to (-\infty, B]$. For any desired level α,

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] \leq \alpha$$

can be arbitrarily close to B. Thus, further assumptions must be placed on the L_i to control the risk.

Proof. w.l.o.g., assume $B = 1$. Assume $\hat{\lambda}$ takes values in $[0, 1]$ and $\alpha \in (1/(n+1), 1)$. Let $p \in (0, 1), N$ be any positive integer, and $L_i(\lambda)$ be i.i.d. right-continuous piecewise constant (random) functions with

$$L_i(N/N) = 0, \quad (L_i(0/N), L_i(1/N), \ldots, L_i((N-1)/N)) \overset{i.i.d.}{\sim} \text{Ber}(p).$$

By definition, $\hat{\lambda}$ is independent of L_{n+1}. Thus, for any $j = 0, 1, \ldots, N-1$,

$$\left\{L_{n+1}(\hat{\lambda}) \mid \hat{\lambda} = j/N\right\} \sim \text{Ber}(p), \quad \left\{L_{n+1}(\hat{\lambda}) \mid \hat{\lambda} = 1\right\} \sim \delta_0.$$

Then,

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] = p \cdot \mathbb{P}(\hat{\lambda} \neq 1)$$

Note that

$$\hat{\lambda} \neq 1 \iff \min_{j \in \{0,\ldots,N-1\}} \frac{1}{n+1} \sum_{i=1}^{n} L_i(j/N) \leq \alpha - \frac{1}{n+1}.$$

Since $\alpha > 1/(n+1)$,

$$\begin{aligned}
\mathbb{P}(\hat{\lambda} \neq 1) = 1 - \mathbb{P}(\hat{\lambda} = 1) &= 1 - \mathbb{P}\left(\text{ for all } j, \text{ we have } \frac{1}{n+1} \sum_{i=1}^{n} L_i(j/N) > \alpha - \frac{1}{n+1}\right) \\
&= 1 - \left(\sum_{k=\lceil (n+1)\alpha \rceil}^{n} \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{(n-k)}\right)^N \\
&= 1 - (1 - \text{BinoCDF}(n, p, \lceil (n+1)\alpha \rceil - 1))^N
\end{aligned}$$

As a result,

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] = p\left(1 - (1 - \text{BinoCDF}(n, p, \lceil (n+1)\alpha \rceil - 1))^N\right).$$

Now let N be sufficiently large such that

$$\left(1 - (1 - \text{BinoCDF}(n, p, \lceil (n+1)\alpha \rceil - 1))^N\right) > p.$$

Then

$$\mathbb{E}\left[L_{n+1}(\hat{\lambda})\right] > p^2$$

For any $\alpha > 0$, we can take p close enough to 1 to render the claim false. □

Corollary 2.1. *Allow $L_i(\lambda)$ to be any (possibly non-monotone) function of λ satisfying:*

$$L_i\left(\lambda_{\max}\right) \leq \alpha, \quad \sup_\lambda L_i(\lambda) \leq B<\infty$$

almost surely. Take

$$\tilde{L}_i(\lambda)=\sup_{\lambda^{\prime} \geq \lambda} L_i\left(\lambda^{\prime}\right), \quad \tilde{R}_n(\lambda)=\frac{1}{n} \sum_{i=1}^n \tilde{L}_i(\lambda) \quad \text{and } \tilde{\lambda}=\inf \left\{\lambda: \frac{n}{n+1} \tilde{R}_n(\lambda)+\frac{B}{n+1} \leq \alpha\right\}.$$

Then,

$$\mathbb{E}\left[L_{n+1}(\tilde{\lambda})\right] \leq \alpha.$$

If the loss function is already monotone, then $\tilde{\lambda}$ reduces to $\hat{\lambda}$.

3 Monotonizing non-monotone risks

Present an algorithm for picking λ that provides an asymptotic risk-control guarantee for non-monotone loss functions when applied to a monotonized version of the empirical risk.

Let the monotonized empirical risk be

$$\widehat{R}_n^{\uparrow}(\lambda)=\sup_{t \geq \lambda} \widehat{R}_n(t),$$

then define

$$\hat{\lambda}_n^{\uparrow}=\inf \left\{\lambda: \widehat{R}_n^{\uparrow}(\lambda) \leq \alpha\right\}.$$

Theorem 3. *Let the $L_i(\lambda)$ be right-continuous, i.i.d., bounded (both above and below) functions satisfying*

$$L_i\left(\lambda_{\max}\right) \leq \alpha, \quad \sup_\lambda L_i(\lambda) \leq B<\infty$$

almost surely. Then,

$$\lim_{n \rightarrow \infty} \mathbb{E}\left[L_{n+1}\left(\hat{\lambda}_n^{\uparrow}\right)\right] \leq \alpha.$$

Proof. Define the monotonized population risk as

$$R^{\uparrow}(\lambda)=\sup_{t \geq \lambda} \mathbb{E}\left[L_{n+1}(t)\right]$$

Note that the independence of L_{n+1} and $\hat{\lambda}_n^{\uparrow}$ implies that for all n,

$$\mathbb{E}\left[L_{n+1}\left(\hat{\lambda}_n^{\uparrow}\right)\right] \leq \mathbb{E}\left[R^{\uparrow}\left(\hat{\lambda}_n^{\uparrow}\right)\right].$$

Since $R^{\uparrow}$ is bounded, monotone, and one-dimensional, a generalization of the Glivenko-Cantelli Theorem given in Theorem 1 of [40] gives that uniformly over λ,

$$\lim_{n \rightarrow \infty} \sup_\lambda\left|\widehat{R}_n(\lambda)-R(\lambda)\right| \stackrel{\text{a.s.}}{\rightarrow} 0.$$

As a result,

$$\lim_{n \rightarrow \infty} \sup_\lambda\left|\widehat{R}_n^{\uparrow}(\lambda)-R^{\uparrow}(\lambda)\right| \stackrel{a.s.}{\rightarrow} 0,$$

which implies that

$$\lim_{n \rightarrow \infty}\left|\widehat{R}_n^{\uparrow}\left(\hat{\lambda}^{\uparrow}\right)-R^{\uparrow}\left(\hat{\lambda}^{\uparrow}\right)\right| \stackrel{a.s.}{\rightarrow} 0.$$

By definition, $\widehat{R}^{\uparrow}\left(\hat{\lambda}^{\uparrow}\right) \leq \alpha$ almost surely and thus this directly implies

$$\limsup_{n\to\infty} R^{\dagger}\left(\hat{\lambda}^{\uparrow}_{n}\right) \leq \alpha \quad \text{a.s.}$$

Finally, since for all n, $R^{\uparrow}\left(\hat{\lambda}^{\dagger}_{n}\right) \leq B$, by Fatou's lemma,

$$\lim_{n\to\infty} \mathbb{E}\left[L_{n+1}\left(\hat{\lambda}^{\uparrow}_{n}\right)\right] \leq \limsup_{n\to\infty} \mathbb{E}\left[R^{\uparrow}\left(\hat{\lambda}^{\uparrow}_{n}\right)\right] \leq \mathbb{E}\left[\limsup_{n\to\infty} R^{\uparrow}\left(\hat{\lambda}^{\uparrow}_{n}\right)\right] \leq \alpha$$

□

This implies that an analogous procedure to $\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}$ also controls the risk asymptotically. In particular, taking

$$\tilde{\lambda}^{\uparrow} = \inf\left\{\lambda : \widehat{R}^{\uparrow}_{n}(\lambda) + \frac{B}{n+1} \leq \alpha\right\}$$

also results in asymptotic risk control.

Note: In the case of a monotone loss function, $\tilde{\lambda}^{\uparrow} = \hat{\lambda}$.