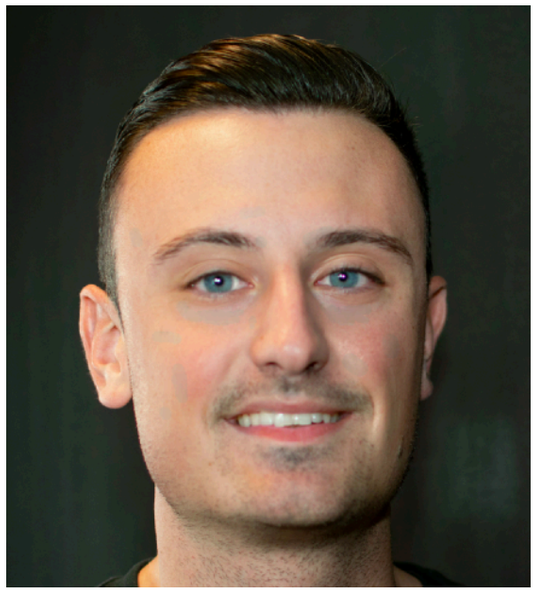
Conformal online model aggregation (COMA)

Aaditya Ramdas

Associate Professor
Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University



Matteo Gasparin

Dept. of Statistical Sciences
University of Padova

arXiv:2401.09379 and arXiv:2403.15527

Outline

- Motivating the problem of combining uncertainty sets
- Majority vote and all its variants
- Applications to *offline* conformal prediction (iid)
- Extensions to *online* conformal prediction (non-iid, COMA)

Conformal prediction

Conformal prediction is a "wrapper" around regression or classification algorithms to provide uncertainty quantification.

Score function

$s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Pre-trained prediction model

$\mu : \mathcal{X} \to \mathcal{Y}$

Held out data

$\{(X_i, Y_i)\}_{i=1}^{n}$

$q_{1-\alpha} :=$ $1-\alpha$ quantile of scores $\{s(\mu(X_i), Y_i)\}_{i=1}^{n}$

For test point X_{n+1}, we output

$C(\alpha) := \{y : s(\mu(X_{n+1}), y) \leq q_{1-\alpha}\}$

Theorem: If $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are iid from P, then

$P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha\,.$

But, which model should we choose?!
If we have K candidate models, we get K prediction sets.

Abstract problem statement

Input: Sets $C_1, \ldots, C_K$ that are (arbitrarily) dependent, and satisfy $P(Y \in C_k) \geq 1 - \alpha$.

Output: A single set C that combines them in a black-box manner and has (nearly) the same coverage guarantee.

Eg: $\bigcup_{k=1}^{K} C_k$ has coverage $1 - \alpha$, but it is too conservative

$\bigcap_{k=1}^{K} C_k$ has coverage $1 - K\alpha$, but it is too anti-conservative

Outline

Motivating the problem of combining uncertainty sets

- Majority vote and all its variants
- Applications to *offline* conformal prediction (iid)
- Extensions to *online* conformal prediction (non-iid, COMA)

First solution: majority vote

Define $C^M := \left\{ y : \frac{1}{K} \sum_{k=1}^{K} 1(y \in C_k) > 1/2 \right\}.$

Theorem: $P(Y \in C^M) \geq 1 - 2\alpha\,.$

Kuncheva et al.'03
Cherubin'19
Solari+Djordjilović'22

Proof: Let $\phi_k := 1(Y \notin C_k)$ and note $\mathbb{E}[\phi_k] \leq \alpha\,.$

By Markov, $\mathbb{P}(Y \notin C^M) = \mathbb{P}\left(\sum_{k=1}^{K} \phi_k \geq K/2 \right) \leq \frac{2}{K} \mathbb{E}\left[\sum_{k=1}^{K} \phi_k \right] \leq 2\alpha\,.$

Define $C^\tau := \left\{ y : \frac{1}{K} \sum_{k=1}^{K} 1(y \in C_k) > \tau \right\}$

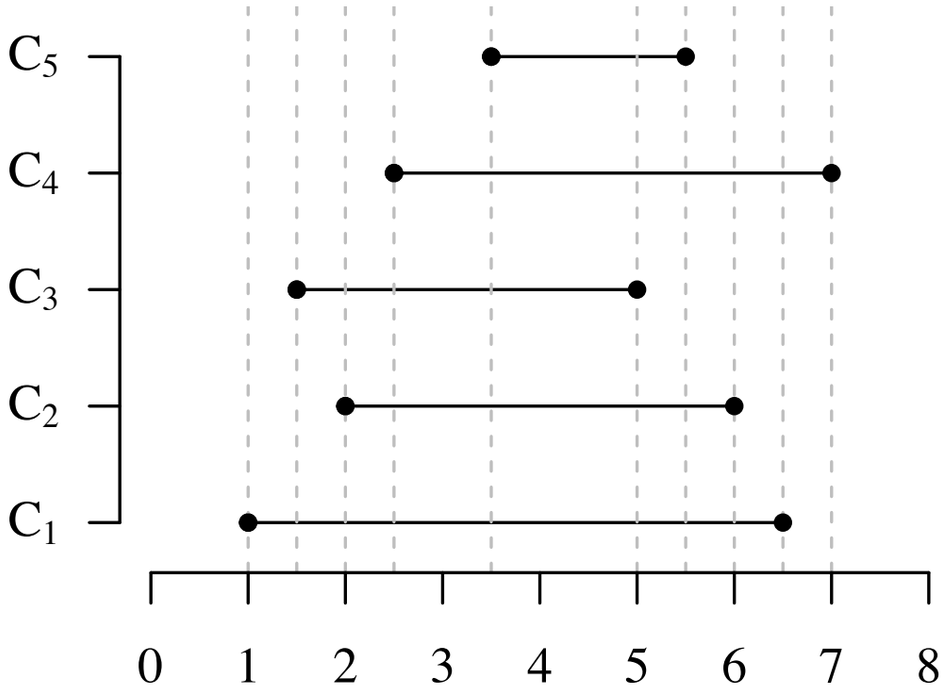
Theorem: $P(Y \in C^\tau) \geq 1 - \frac{\alpha}{1-\tau}\,.$

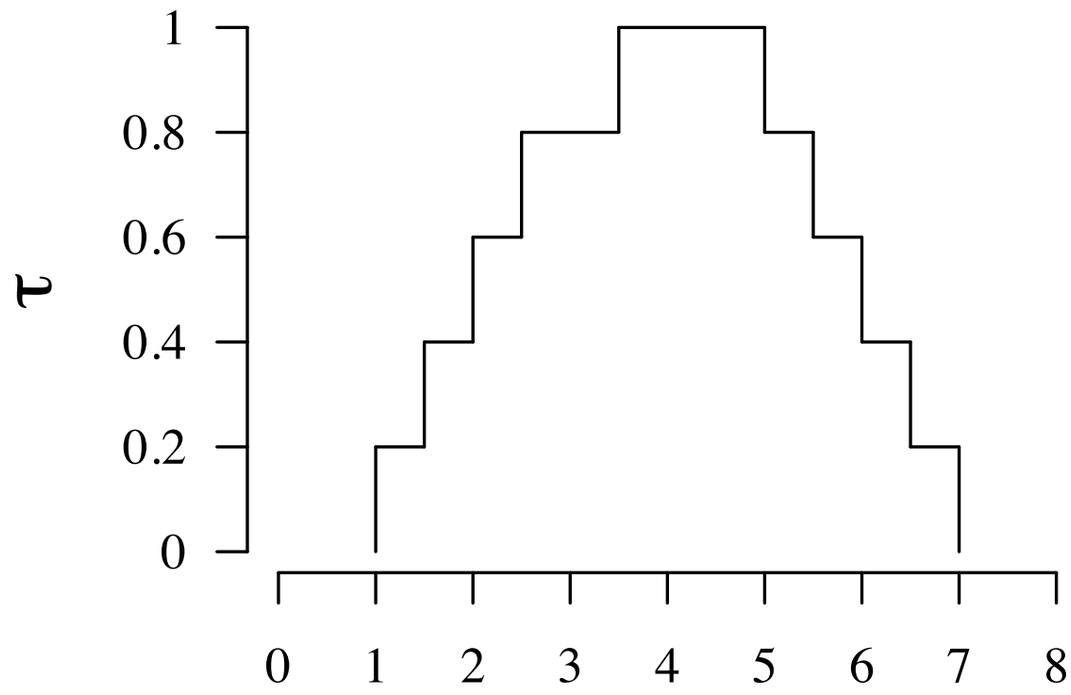
$\tau = 0$ is the union, and $\tau = 1 - 1/K$ is the intersection.

Sometimes, the majority vote yields an interval

Lemma: If $\bigcap_k C_k \neq \emptyset$,

then C^τ is an interval for any τ.





Otherwise, it is at most a union of K intervals.

How large is the set?

Theorem: $\text{size}(C^M) \leq 2 \sum_{k=1}^{K} \text{size}(C_k)/K$.

Also, $\text{size}(C^M) \leq \max_{k=1}^{K} \text{size}(C_k)$.

Here, size = length (regression) or cardinality (classification)

Special case: equal intervals "median of midpoints"

Suppose all intervals have equal size (assume odd K for simplicity).
Sort them by their midpoints.

Report the "median interval"
(interval corresponding to the median of midpoints).

This has coverage $1 - \alpha$ and is larger than C^M,
and has the same size as input intervals.

Using prior information

Suppose we have some idea that certain models
are likely to be more accurate (small sets) than others.

We can incorporate this knowledge using a (data-independent)
prior distribution w over the models.

Define $C^W := \left\{ y : \frac{1}{K} \sum_{k=1}^{K} w_k 1(y \in C_k) > 1/2 \right\}$
as the "weighted majority vote" set.

Theorem: $P(Y \in C^W) \geq 1 - 2\alpha$.

Randomized voting

Let u be the realization of a uniform $[0,1]$ random variable, Independent of all the data.

Define $C^R := \left\{ y : \frac{1}{K} \sum_{k=1}^{K} w_k 1(y \in C_k) > 1/2 + u/2 \right\}$

Theorem: $C^R \subseteq C^W$ and $P(Y \in C^R) \geq 1 - 2\alpha$.

Proof uses "randomized Markov's inequality" (R+Manole'23):
$P(X \geq Ua) \leq \mathbb{E}[X]/a$ for any nonnegative X, a.

Alternately $C^U := \left\{ y : \frac{1}{K} \sum_{k=1}^{K} w_k 1(y \in C_k) > u \right\}$

Theorem: $P(Y \in C^U) \geq 1 - \alpha$.

Outline

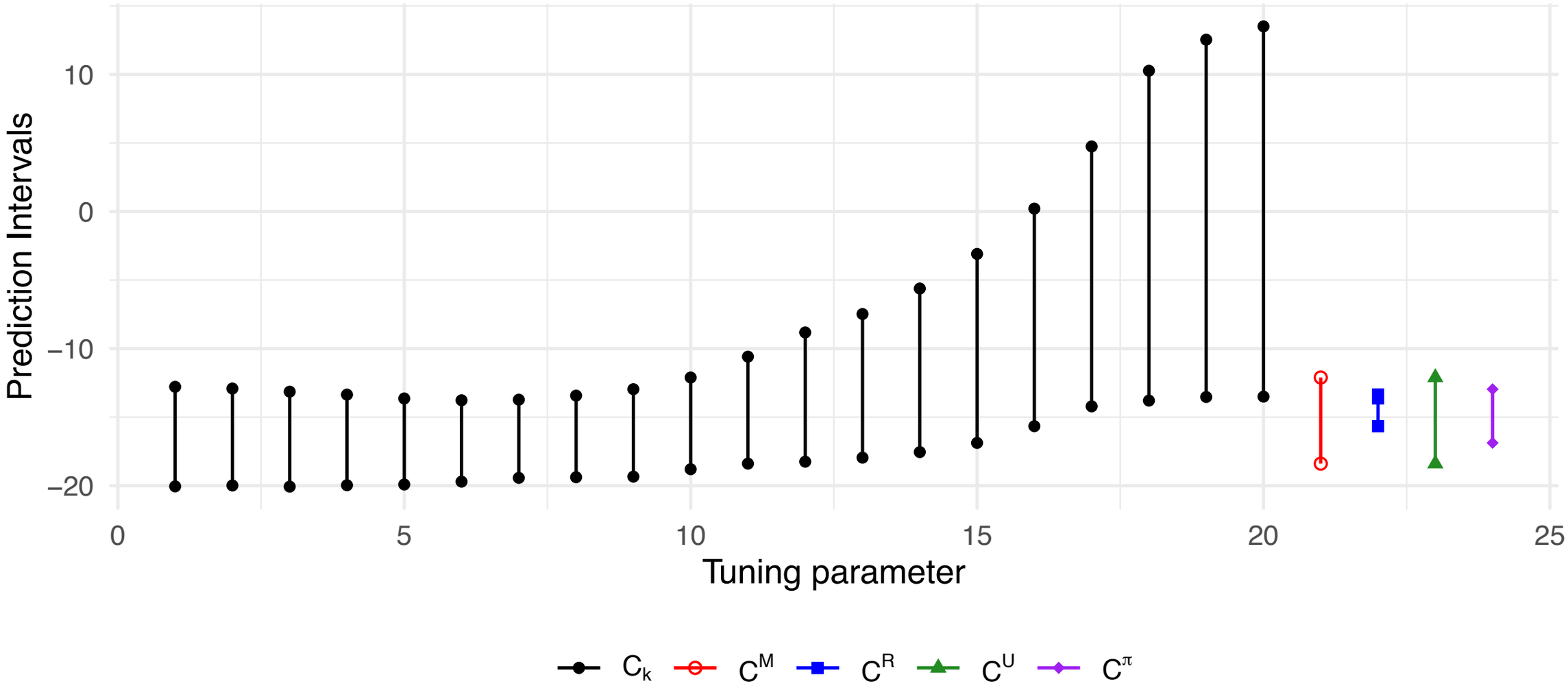

Motivating the problem of combining uncertainty sets


Majority vote and its variants

- Applications to *offline* conformal prediction (iid)
- Extensions to *online* conformal prediction (non-iid, COMA)

Eg: conformal prediction with Lasso

$K = 20$ models: Fit lasso to training data, with ℓ_1 penalty $\lambda_1, \ldots, \lambda_{20}$.
Combine the 20 conformal prediction sets at level $\alpha = 0.05$.
The plot below is for a particular random test point X.



Randomized sets used $u = 1/2$ for visualization.
Coverage: 97% (C^M), 92% (C^R), 96% (C^U)
>99% of the time, these are intervals.

Combining *exchangeable* sets

Let $C^M(1:k)$ denote the majority vote of sets $C_1, \ldots, C_k$.

Define $C^E(1:K) := \bigcap_{k=1}^{K} C^M(1:k)$.

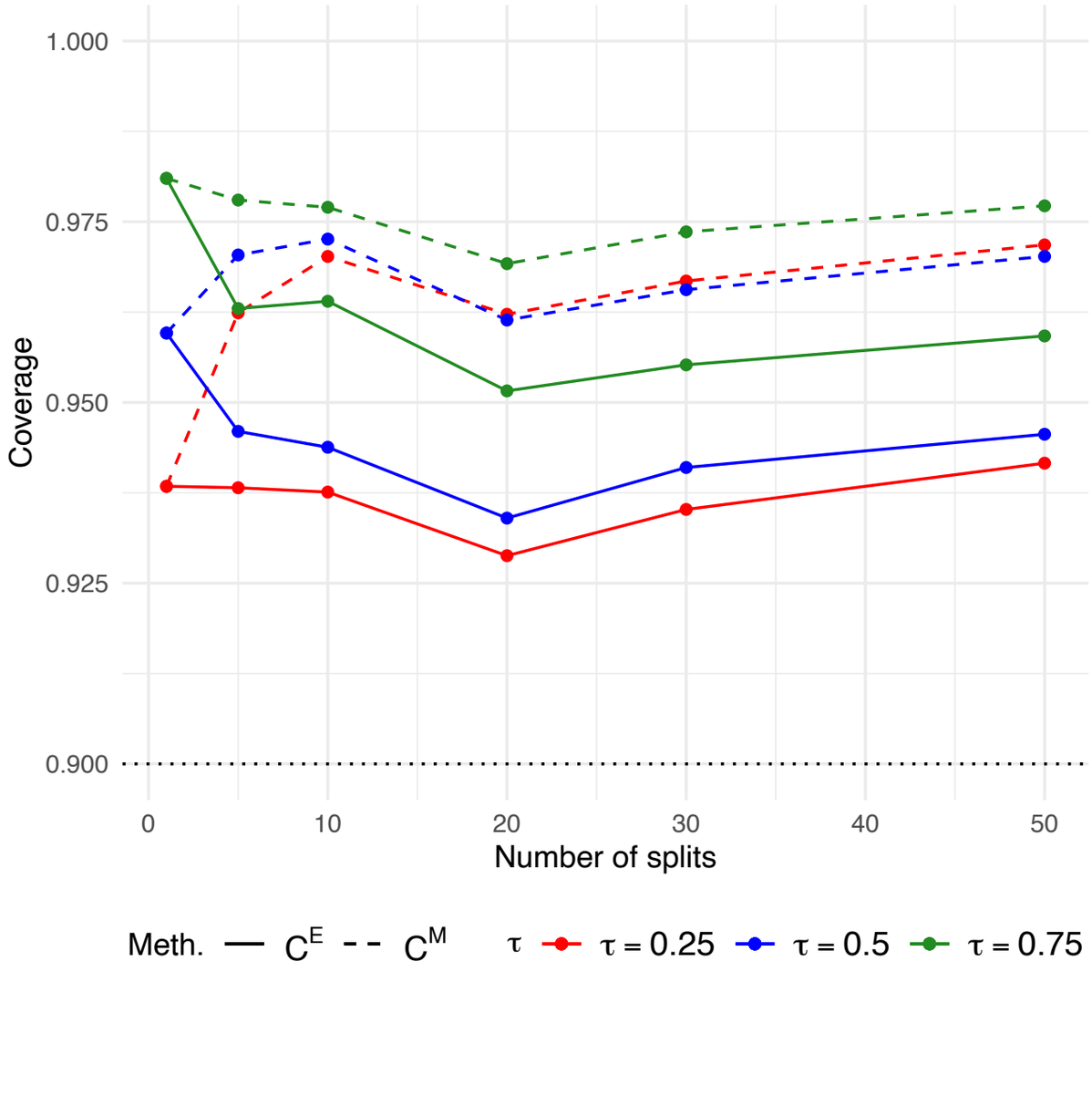
Theorem: If the input sets are exchangeable and have coverage $1-\alpha$, then $C^E(1:K)$ has coverage $1-2\alpha$. In fact, for an infinite sequence of exchangeable sets, $P(\exists t \geq 1 : Y \notin C^E(1:t)) \leq 2\alpha$.

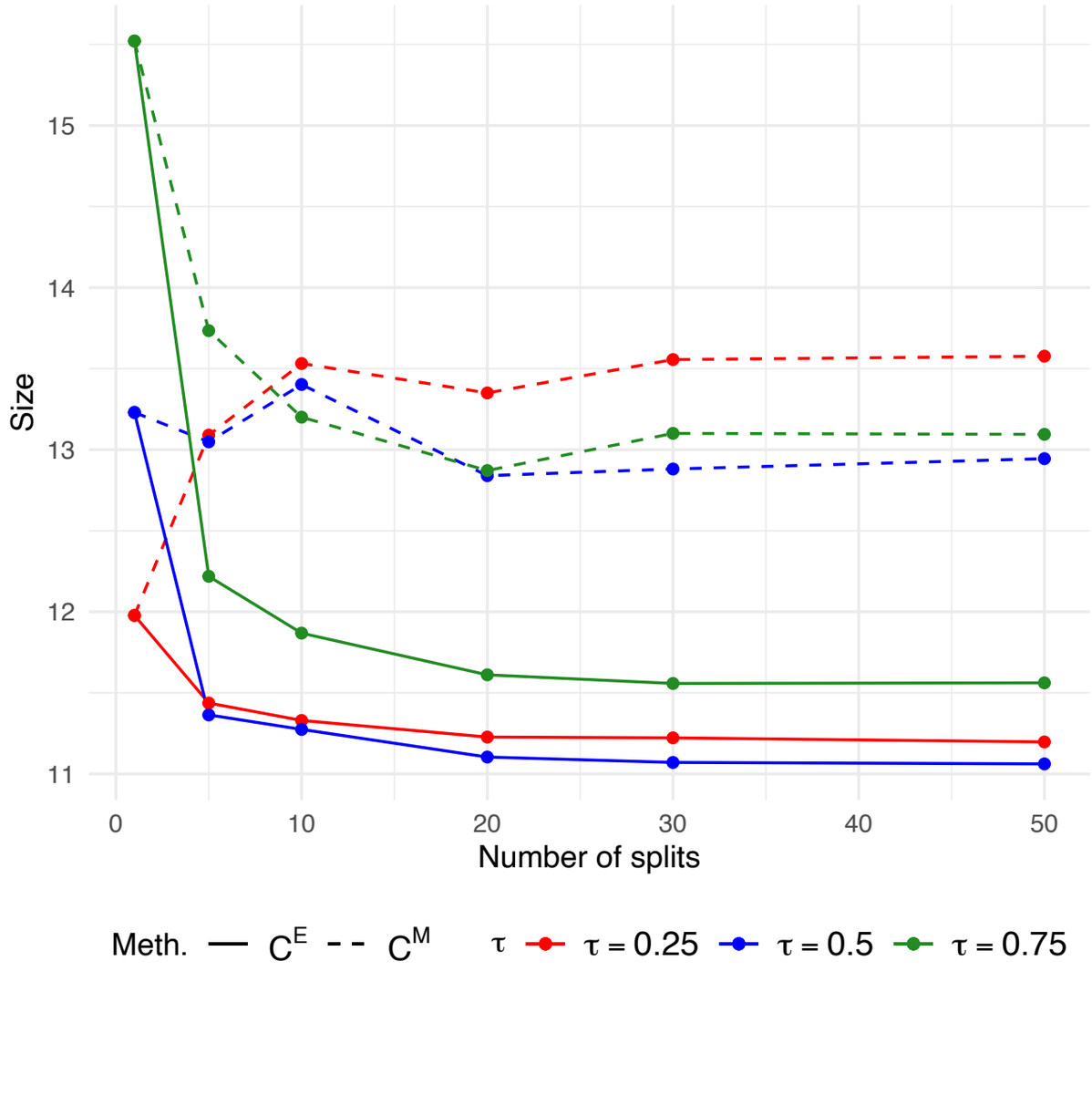
Proof: Use "exchangeable Markov inequality" (Manole+Ramdas'23) which states that for any sequence $X_1, X_2, \ldots$ of exchangeable nonnegative random variables, $P(\exists k \geq 1 : \bar{X}_k \geq 1/\alpha) \leq \mathbb{E}[X]\alpha$.

Note: Arbitrarily dependent sets can always be made exchangeable by randomizing their order.

Eg: multi-split conformal prediction

Split-conformal prediction is based on sample splitting.
(Use part of the data to train model, held-out data for conformal prediction)
We suggest: Repeat many times and take *exchangeable* majority vote.

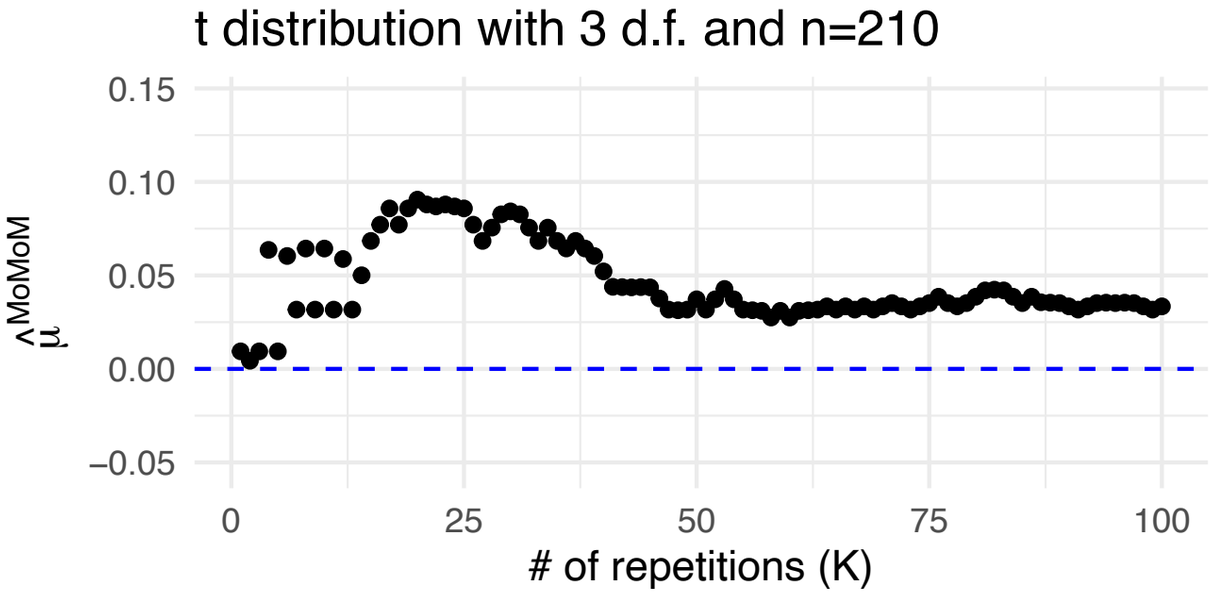


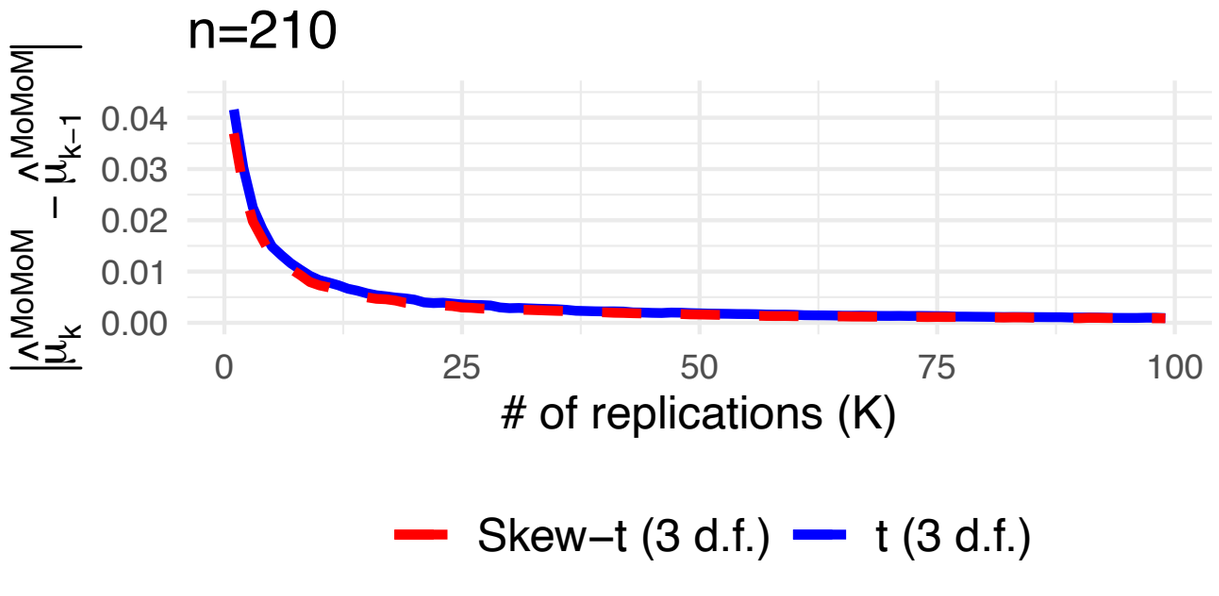


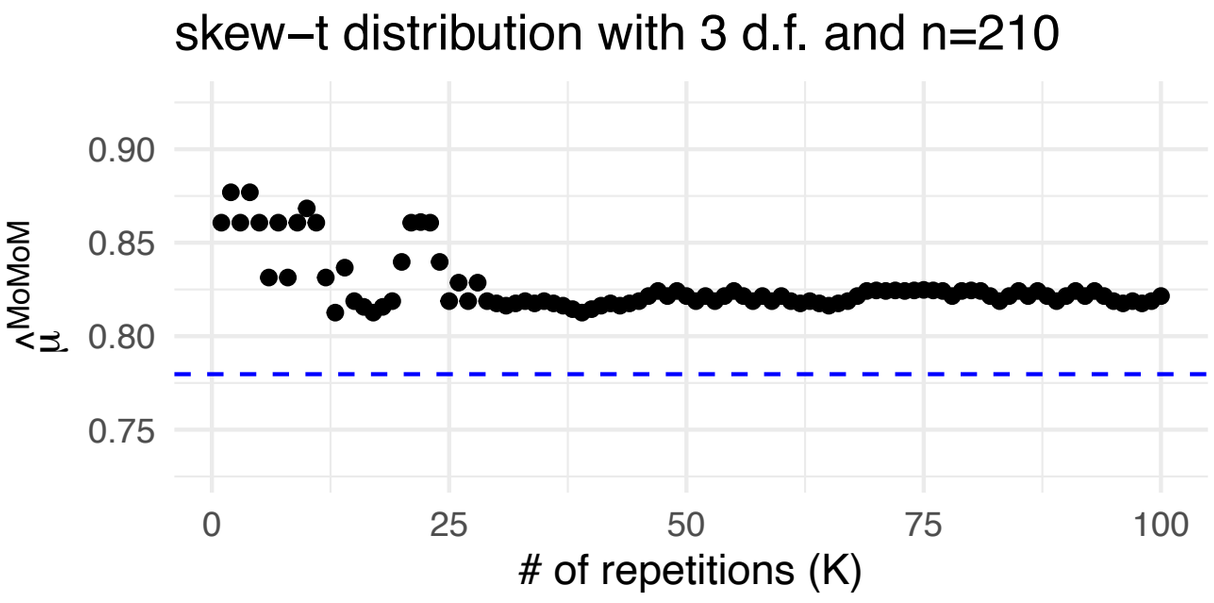
Eg: derandomizing Median-of-means (MoM)

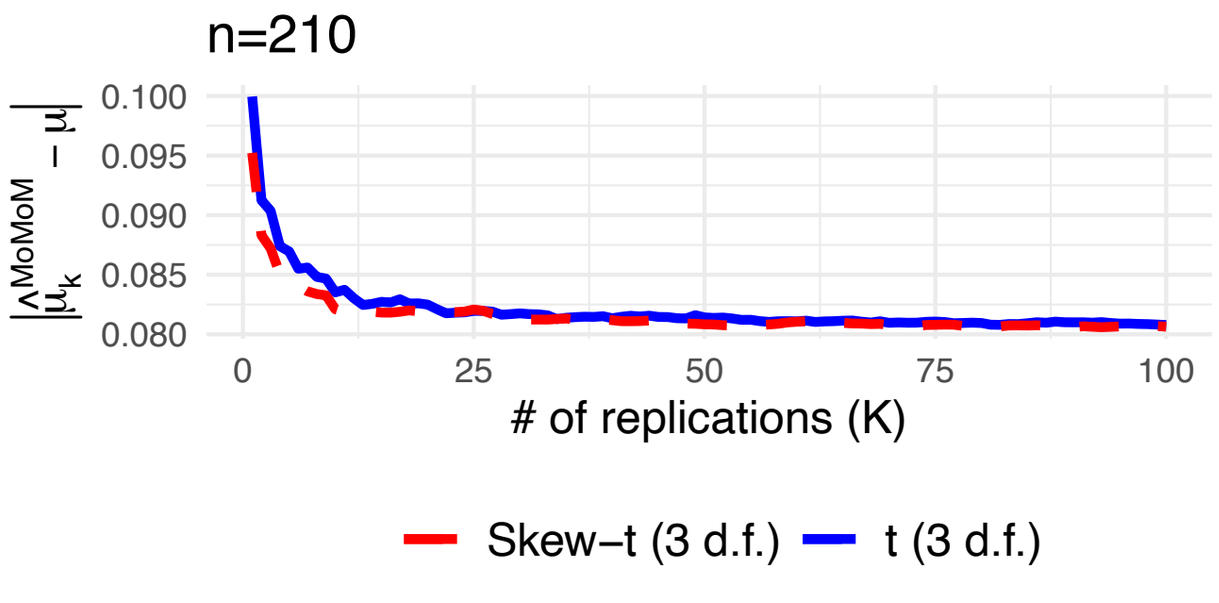
MoM is a method to estimate the mean of a (heavy-tailed) distribution with two moments. It splits the data into B buckets, calculates mean within each bucket, and then takes the median across buckets.

We suggest: repeat many times, take "median of median of means".
(Essentially the same theoretical guarantee, much better practically.)









What if the sets have different coverages?

If the data are not actually iid or exchangeable, then each model may achieve a different coverage level.

Theorem: If set C_k has coverage $1 - \alpha_k$, then

$P(Y \in C^R) \geq 1 - \frac{2}{K} \sum_{k=1}^{K} w_k \alpha_k$. Same guarantee for C^M, C^W.

Corollary: if the input sets have asymptotic coverage, so does the majority vote set.

Outline

- Motivating the problem of combining uncertainty sets
- Majority vote and all its variants
- Applications to *offline* conformal prediction (iid)
- Extensions to *online* conformal prediction (COMA)
    A. IID settings
    B. Non-IID settings

A. If we use K models to make repeated predictions **on iid data**, how can we get **small sets with valid coverage?**

Improving with experience: case 1 (iid data)

Start with uniform weights over the K models.
As predictions are made, outcomes are observed,
update the weights via the "exponential weights" algorithm.

Pick loss function $\ell^{(t)} = \text{size}(C^{(t)})$
eg: count (classification) or length (regression).
Some subtleties: not bounded, but nonnegative.

Algorithm 1: Exponentially Weighted Majority Vote

Data: $\mathcal{C}_1^{(t)}, \ldots, \mathcal{C}_K^{(t)}$ at each round t, initial learning rate $\eta^{(0)} \geq 0$
$w_k^{(1)} \leftarrow 1, L_k^{(0)} \leftarrow 0,\ k = 1, \ldots, K;$
for *rounds* $t = 1, \ldots, T$ *do* **do**
  $\mathcal{C}^{(t)} \leftarrow \left\{ s \in \mathcal{S} : \sum_{k=1}^{K} w_k^{(t)} \mathbb{1}\{s \in \mathcal{C}_k^{(t)}\} > \frac{\sum_{k=1}^{K} w_k^{(t)}}{2} \right\};$
  Receive loss $\ell_k^{(t)}$, update $L_k^{(t)} := L_k^{(t-1)} + \ell_k^{(t)};$
  Update learning rate $\eta^{(t)}$;
  $w_k^{(t+1)} \leftarrow \exp\{-\eta^{(t)} L_k^{(t)}\};$
end

AdaHedge step-size

So how should we set learning rate $\eta^{(t)}$?

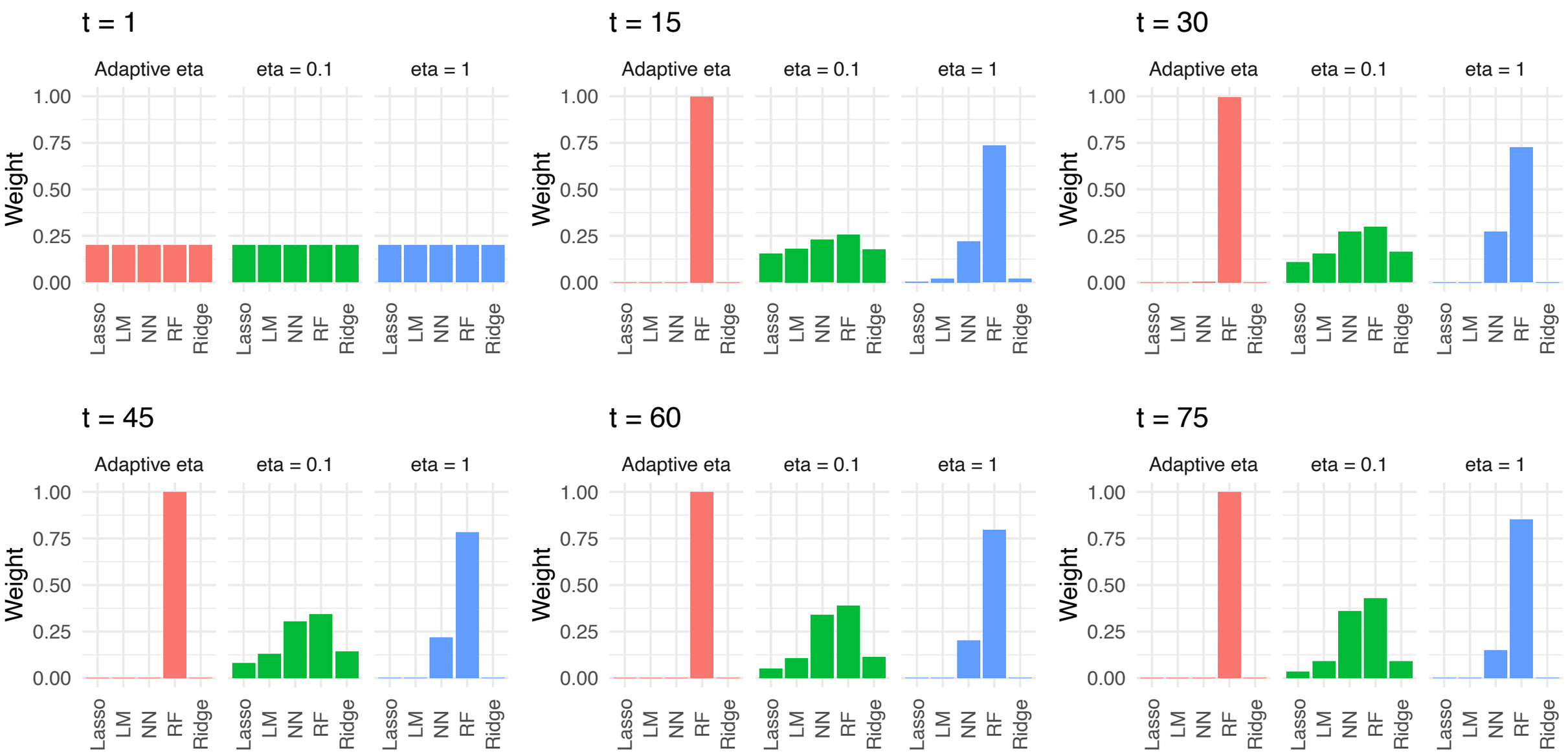
AdaHedge (de Rooij et al. 2014)

$$\eta^{(t)} := \frac{\ln K}{\delta^{(1)} + \ldots + \delta^{(t-1)}}$$

where $\delta^{(i)} := h^{(i)} - m^{(i)}$,

$h^{(i)} := w^{(i)} \cdot \ell^{(i)}$ is the hedge loss,

$m^{(i)} := -\frac{1}{\eta^{(i)}} \ln(w^{(i)} \cdot \exp(-\eta^{(i)} \ell^{(i)}))$ is the "mix loss"


t = 1
t = 15
t = 30
t = 45
t = 60
t = 75
Adaptive eta
eta = 0.1
eta = 1
Weight
1.00
0.75
0.50
0.25
0.00
Lasso
LM
NN
RF
Ridge

Improving with experience: case 2 (dist. shift)

In the previous algorithm, we implicitly assumed that each expert was providing level $(1-\alpha)$ sets as requested, so our loss function only charged the size of the set.

This is reasonable in iid settings, because conformal prediction does yield such a guarantee.

But what if there is distribution shift/drift, due to which different models may be (unknowingly) over/underconfident?

Can we still combine them to yield small sets, at the desired level?

Next, we wrap around an existing algorithm by Gibbs and Candes ('21-23) called "adaptive conformal inference" (ACI), and a recent algorithm called "quantile tracking" (Angelopoulos et al.' 23), built to adapt to distribution shifts.

Adaptive conformal inference (ACI) — for one model

For a single model μ, ACI outputs $(1-\alpha^{(t)})$ prediction sets, where

$$\alpha^{(1)} = \alpha \text{ and}$$

$$\alpha^{(t)} := \alpha^{(t-1)} + \gamma_t(\alpha - \phi^{(t-1)}),$$

where $\phi^{(t)} = 1\{Y_t \notin C^{(t)}(\alpha^{(t)})\}$ indicates miscoverage,

$\gamma_t > 0$ is a stepsize.

(Gibbs+Candes'21)

A "better" algorithm: quantile tracking

To avoid the issue of possibly negative $\alpha^{(t)}$, switch to quantile space.

$$q^{(1)} = q \text{ and}$$

$$q^{(t)} := q^{(t-1)} - \gamma_t(\alpha - \phi^{(t-1)})$$

$$C^{(t)}(q^{(t)}) := \{y : s^{(t)}(\mu(x^{(t)}), y) \le q^{(t)}\}$$

(Angelopoulos, Candes, Tibshirani'24)

Both methods satisfy $\frac{1}{T}\sum_{t=1}^{T} \phi^{(t)} = \alpha + o(1)$ deterministically

for appropriate stepsize γ choices.

Conformal online model aggregation (COMA)

We propose two ways to combine ACI with our dynamic merging algorithm.

Method 1 (uncoordinated): we have K uncoordinated ACI algorithms running, each having a different base prediction method (random forest or lasso), and we simply wrap our exponential weighted majority vote on top.

Since each ACI takes care of ensuring that its own algorithm has well calibrated error level, our wrapper can take care of focusing on length.

We will get asymptotic coverage $1 - 2\alpha$.

Method 2 (coordinated): we have K different ACI algorithms running, but the feedback provided to them is not about *their own* miscoverage, but the miscoverage of the combined majority vote set.

Now, our overall algorithm will be level $1 - \alpha$ asymptotically.

Predict Amazon daily opening stock price (2006-14)

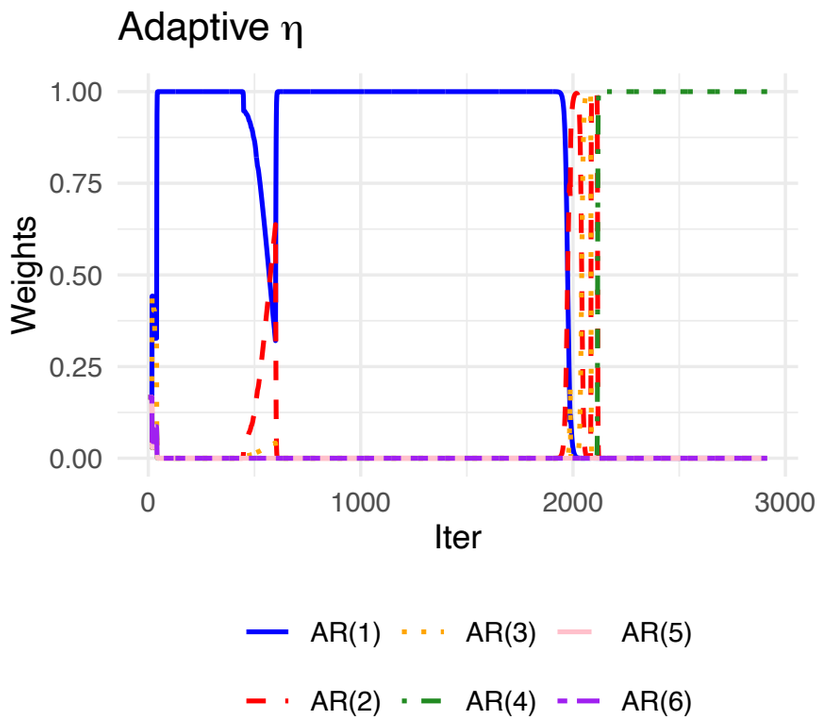
Combining 6 models: AR(1) to AR(6)

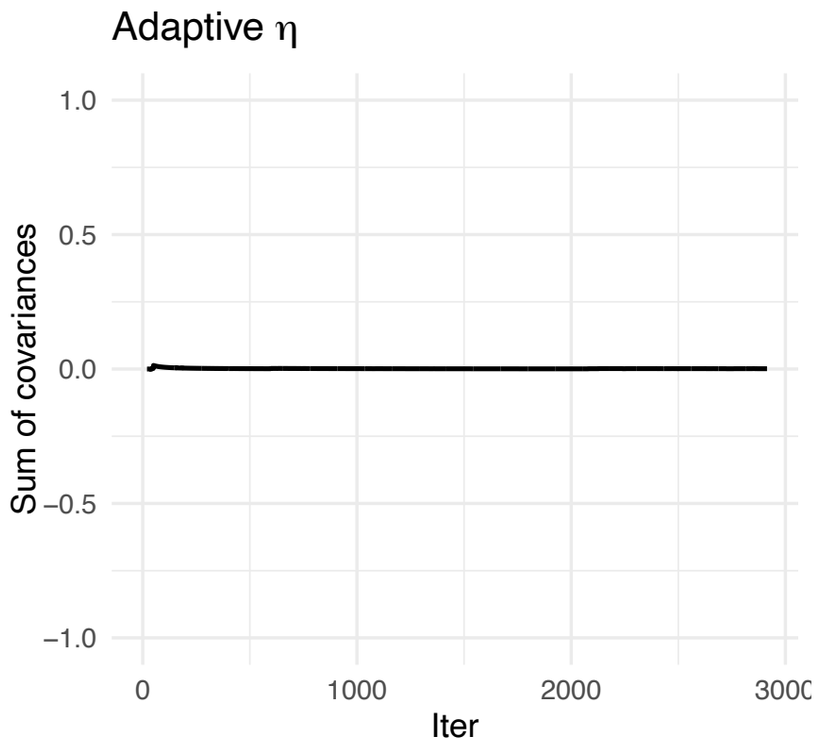
Predict Amazon daily opening stock price (2006-14)

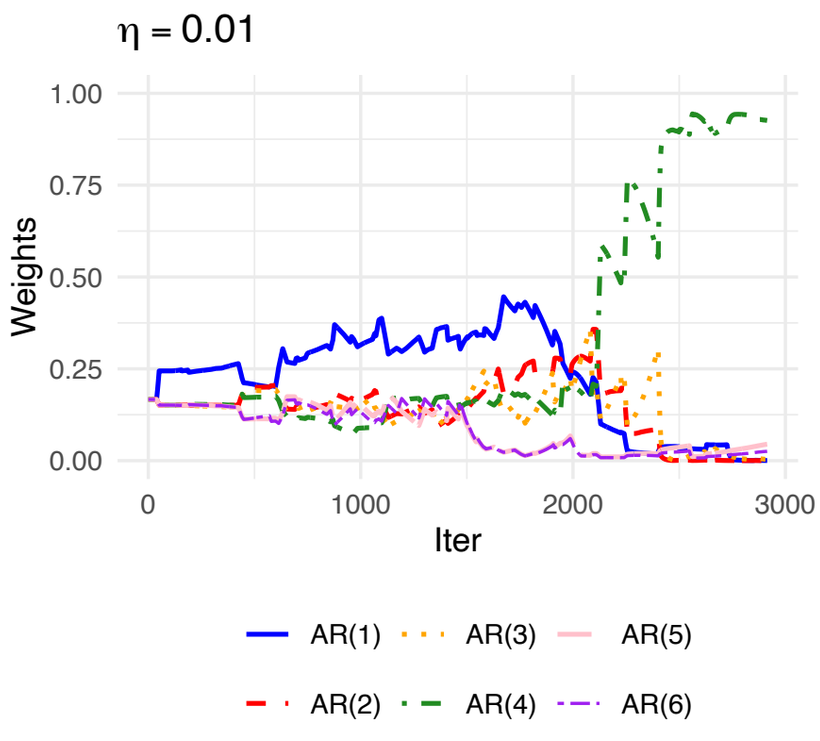
Combining 6 models: AR(1) to AR(6)

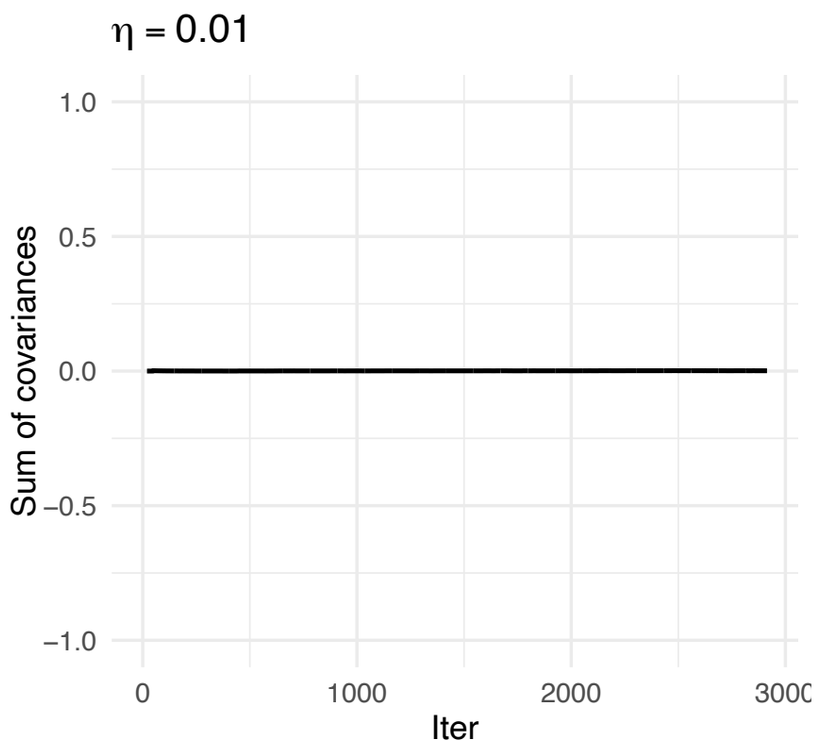
Predict Amazon daily opening stock price (2006-14)

Combining 6 models: AR(1) to AR(6)









Outline

- ✓ Motivating the problem of combining uncertainty sets
- ✓ Majority vote and its variants
- ✓ Applications to *offline* conformal prediction (iid)
- ✓ Extensions to *online* conformal prediction (non-iid, COMA)

Summary of talk

- Proposed randomized, weighted variants of majority vote that can efficiently combine dependent uncertainty sets

- Demonstrated applications to derandomization of statistical procedures (like split conformal prediction)

- Showed how to extend these to online settings, in particular performing "conformal online model aggregation"

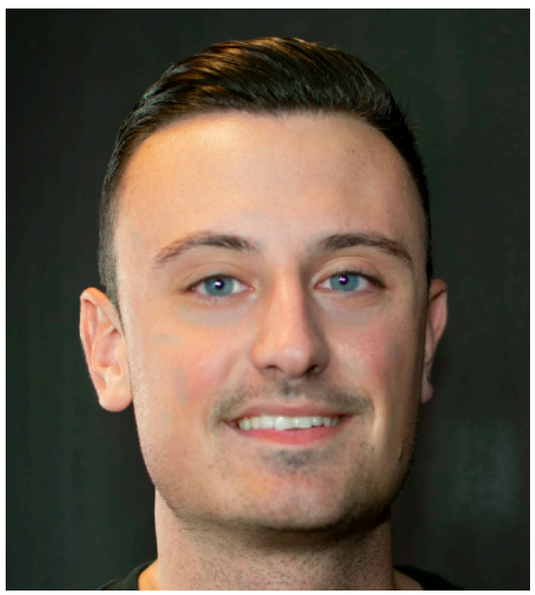
- The paper also has an extension "beyond coverage" to bounded loss functions ("conformal risk control")

arXiv:2401.09379 and arXiv:2403.15527

Conformal online model aggregation (COMA)

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University



Matteo Gasparin

Dept. of Statistical Sciences
University of Padova

arXiv:2401.09379 and arXiv:2403.15527