Discussion of "Prediction inference is free with the jackknife+-after-bootstrap"

Yachong Elsa Yang

University of Pennsylvania, Department of Statistics and Data Science

International Selective Inference Seminar Oct.14,2021

Conformal Prediction

Given i.i.d. pairs $(X_i, Y_i) \sim P, i = 1, \ldots, N$, for a distribution P on

$$\mathcal{X} \times \mathbb{R} \quad \left(\text{e.g., } \mathcal{X} = \mathbb{R}^d\right)$$

Goal. Build a prediction set $\widehat{C}_N$, such that for new i.i.d. pair (X_{N+1}, Y_{N+1}) :

$$\mathbb{P}\left((X_{N+1}, Y_{N+1}) \in \widehat{C}_N\right) \geq 1 - \alpha,$$

(where the probability is over all $N + 1$ pairs).

Conformal prediction provides a simple and finite-sample valid solution to this problem without any assumptions on P.

Jackknife+ after Bootstrap

Theorem 1

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}^{J+aB}_{\alpha,n,B}\left(X_{n+1}\right)\right] \geq 1 - 2\alpha$$

for any data distribution, any base regression method $\mathcal{R}$, and any aggregation procedure φ, provided that B is chosen as

- $B \sim \text{Binomial}\left(\tilde{B}, \left(1 - \frac{1}{n+1}\right)^m\right)$ *if sampling with replacement, or*
- $B \sim \text{Binomial}\left(\tilde{B}, 1 - \frac{m}{n+1}\right)$ *if sampling without replacement. $m \geq 1$ is the size of each resampled / subsampled data set, and $\tilde{B} \geq 1$.*

Jackknife+ after Bootstrap

Theorem 1

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}^{J+aB}_{\alpha,n,B}\left(X_{n+1}\right)\right] \geq 1 - 2\alpha$$

for any data distribution, any base regression method $\mathcal{R}$, and any aggregation procedure φ, provided that B is chosen as

- $B \sim \text{Binomial}\left(\tilde{B}, \left(1 - \frac{1}{n+1}\right)^m\right)$ *if sampling with replacement, or*
- $B \sim \text{Binomial}\left(\tilde{B}, 1 - \frac{m}{n+1}\right)$ *if sampling without replacement. $m \geq 1$ is the size of each resampled / subsampled data set, and $\tilde{B} \geq 1$.*

In most settings where a large number of models are being aggregated, we would not expect a big difference from random vs fixed B.

Questions I have

- Coverage tightness: in practice should we use the α level or $\alpha/2$ level in order to guarantee the $1 - \alpha$ coverage?

Questions I have

- Coverage tightness: in practice should we use the α level or $\alpha/2$ level in order to guarantee the $1-\alpha$ coverage?
- In split conformal prediction, the coverage is bounded between $1-\alpha$ and $1-\alpha+1/n$. In comparison, the coverage of J+aB can get to 1. Under what settings can it get too conservative?

Questions I have

- Coverage tightness: in practice should we use the α level or $\alpha/2$ level in order to guarantee the $1-\alpha$ coverage?
- In split conformal prediction, the coverage is bounded between $1-\alpha$ and $1-\alpha+1/n$. In comparison, the coverage of J+aB can get to 1. Under what settings can it get too conservative?
- Does higher values of B (number of bootstrap samples) and m (the number of samples in each bootstrap) give better theoretical results? How to choose tuning parameters?

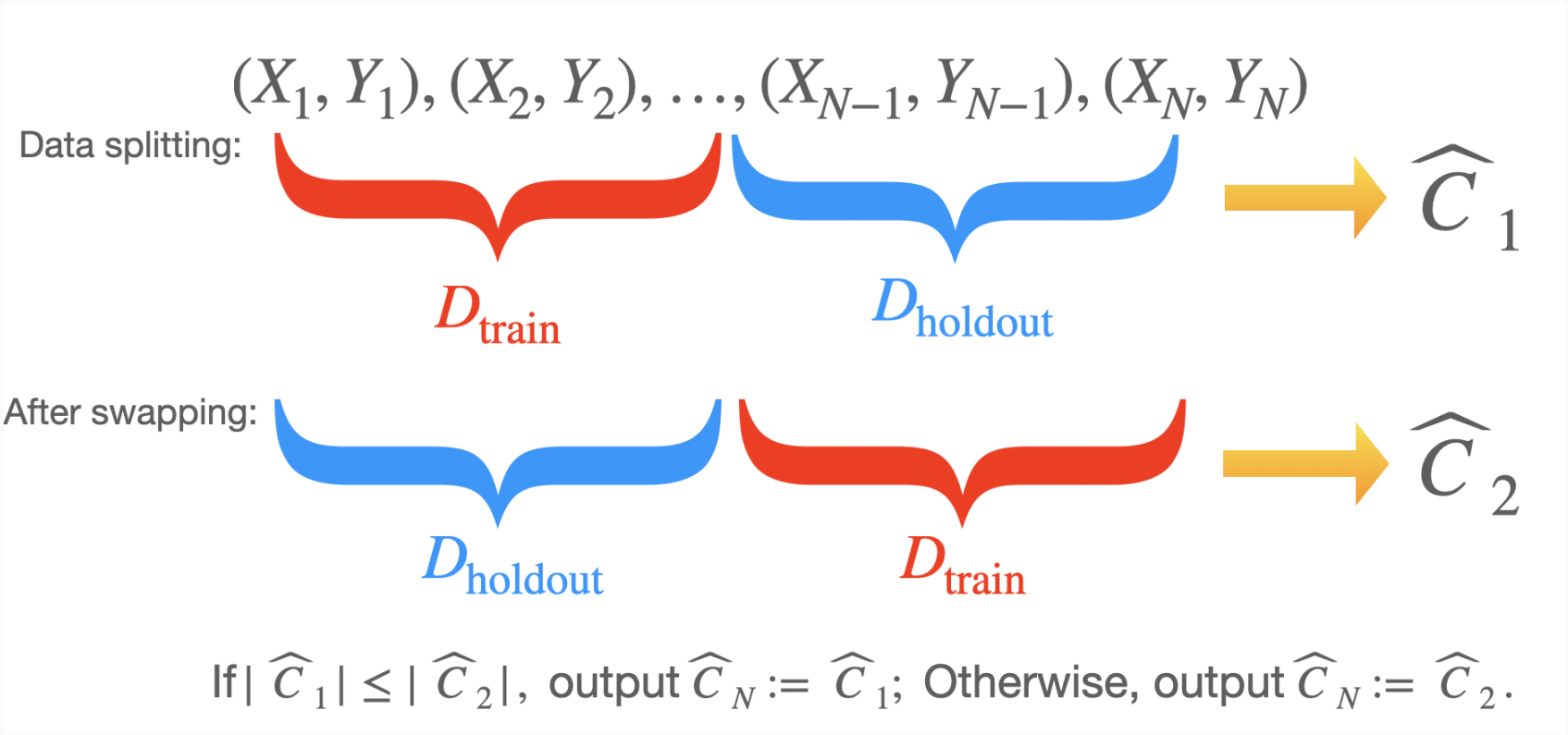
When do we need to ensemble?

- Split conformal prediction is computationally very efficient but a major drawback is that it does training and validating on separate parts of the data, therefore not using the entire data efficiently.

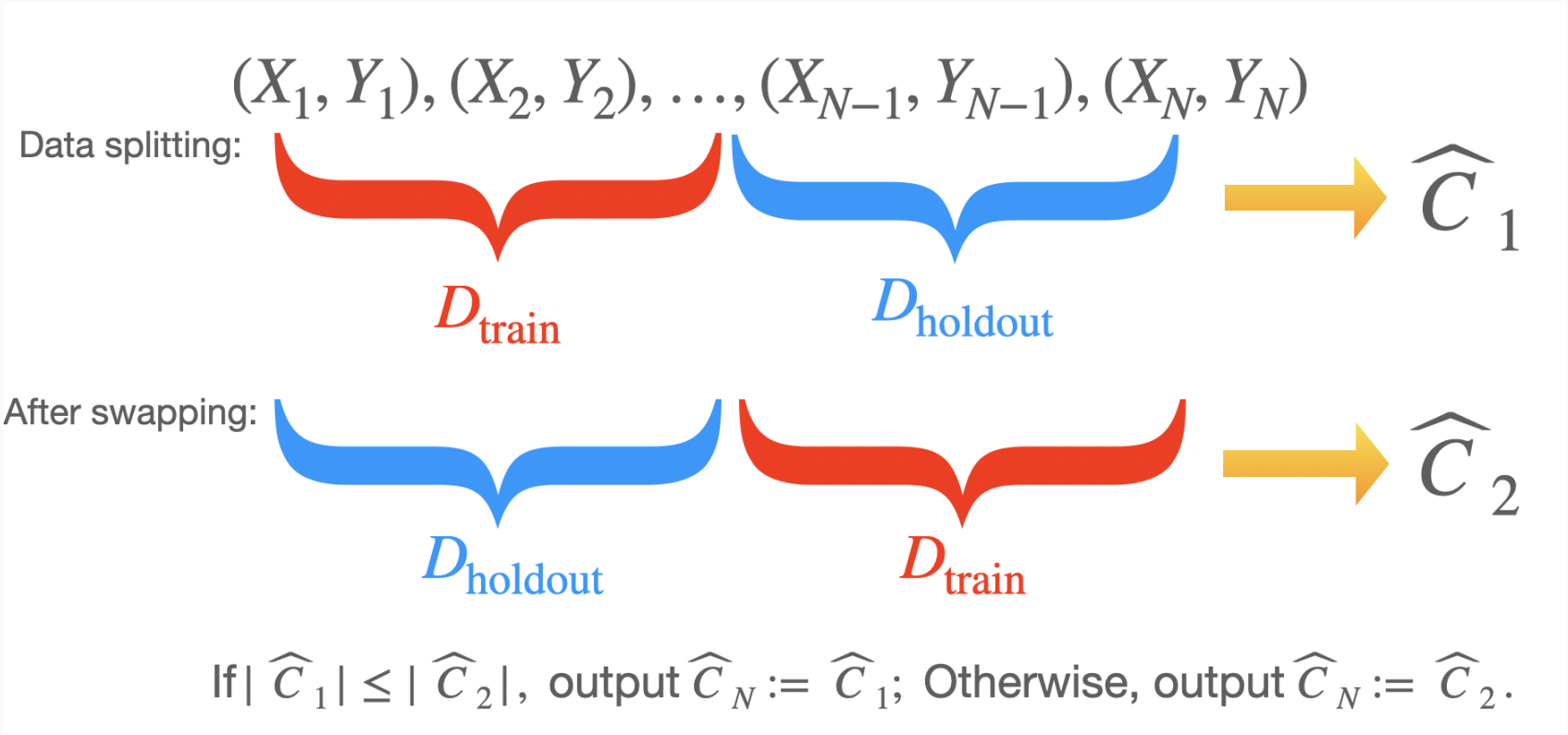
When do we need to ensemble?

- Split conformal prediction is computationally very efficient but a major drawback is that it does training and validating on separate parts of the data, therefore not using the entire data efficiently.
- One way to overcome this is to swap the training data and the holdout data and among the two predictions sets obtained by swapping, choose the one with the smaller size.

An improvement upon split conformal prediction



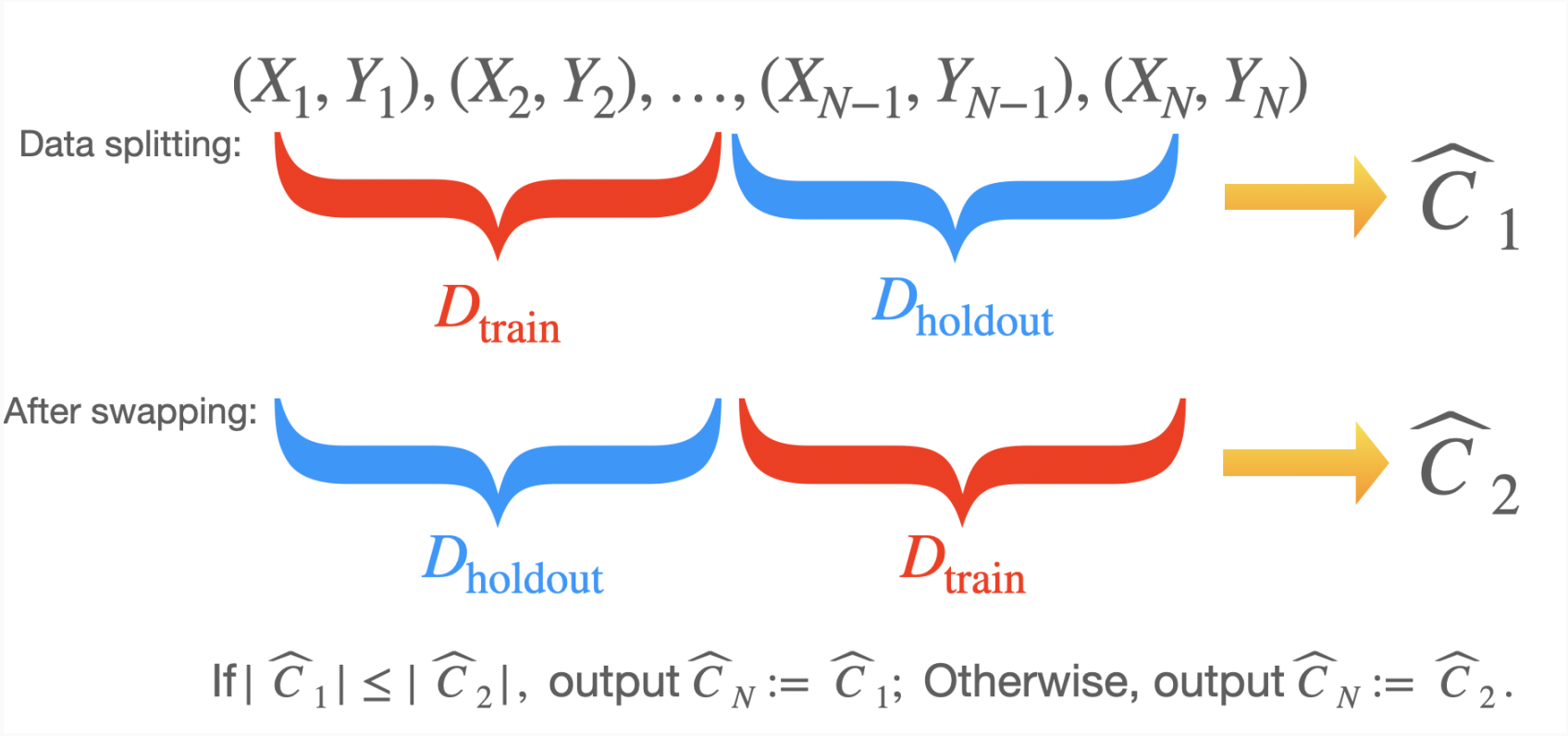
An improvement upon split conformal prediction



Validity:

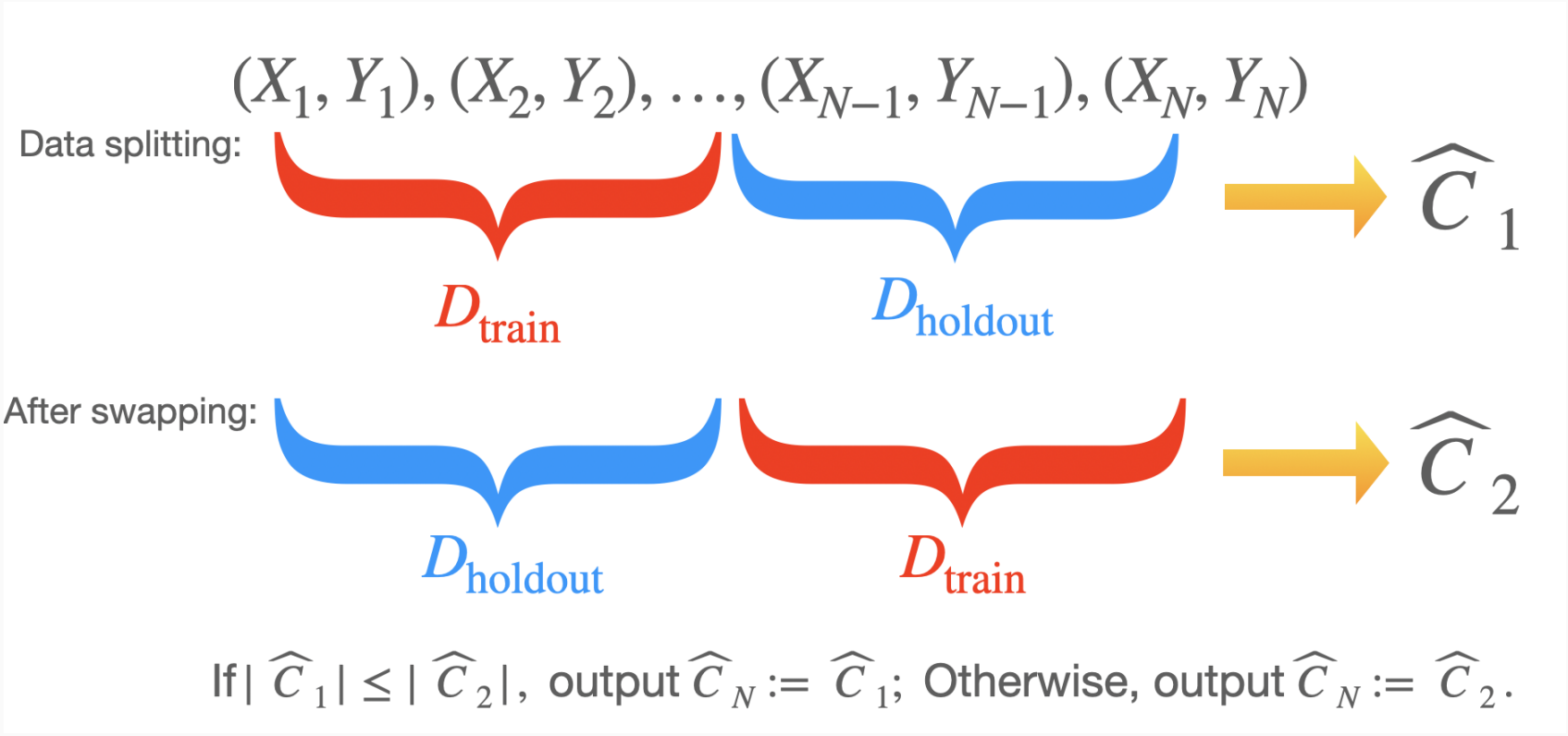
$$\left| \mathbb{P}\left((X_{N+1}, Y_{N+1}) \in \widehat{C}_N\right) - \frac{\lceil (1-\alpha)(1+\mathcal{D}_{\text{holdout}}) \rceil}{|\mathcal{D}_{\text{holdout}}|} \right| \leq \frac{\sqrt{\log(4)/2} + 1/3}{\sqrt{|\mathcal{D}_{\text{holdout}}|}}.$$

An improvement upon split conformal prediction



- It's mentioned in the original Jackknife+ paper that empirically Jackknife+ produces shorter predictions sets compared to split conformal prediction.

An improvement upon split conformal prediction



- It's mentioned in the original Jackknife+ paper that empirically Jackknife+ produces shorter predictions sets compared to split conformal prediction.
  - How does J+aB compare to the method proposed above?

An alternative way of doing ensembling

The end goal is the prediction set, so instead of aggregating estimators, would it be better to aggregate the prediction sets obtained from different splits of the data?

[1] Finite-sample Efficient Conformal Prediction, Yachong Yang, Arun Kumar Kuchibhotla, arxiv 2021.

An alternative way of doing ensembling

The end goal is the prediction set, so instead of aggregating estimators, would it be better to aggregate the prediction sets obtained from different splits of the data?

- One way is to do different splits of the data where we perform split conformal prediction on each split and choose the prediction set with the smallest size.

[1] Finite-sample Efficient Conformal Prediction, Yachong Yang, Arun Kumar Kuchibhotla, arxiv 2021.

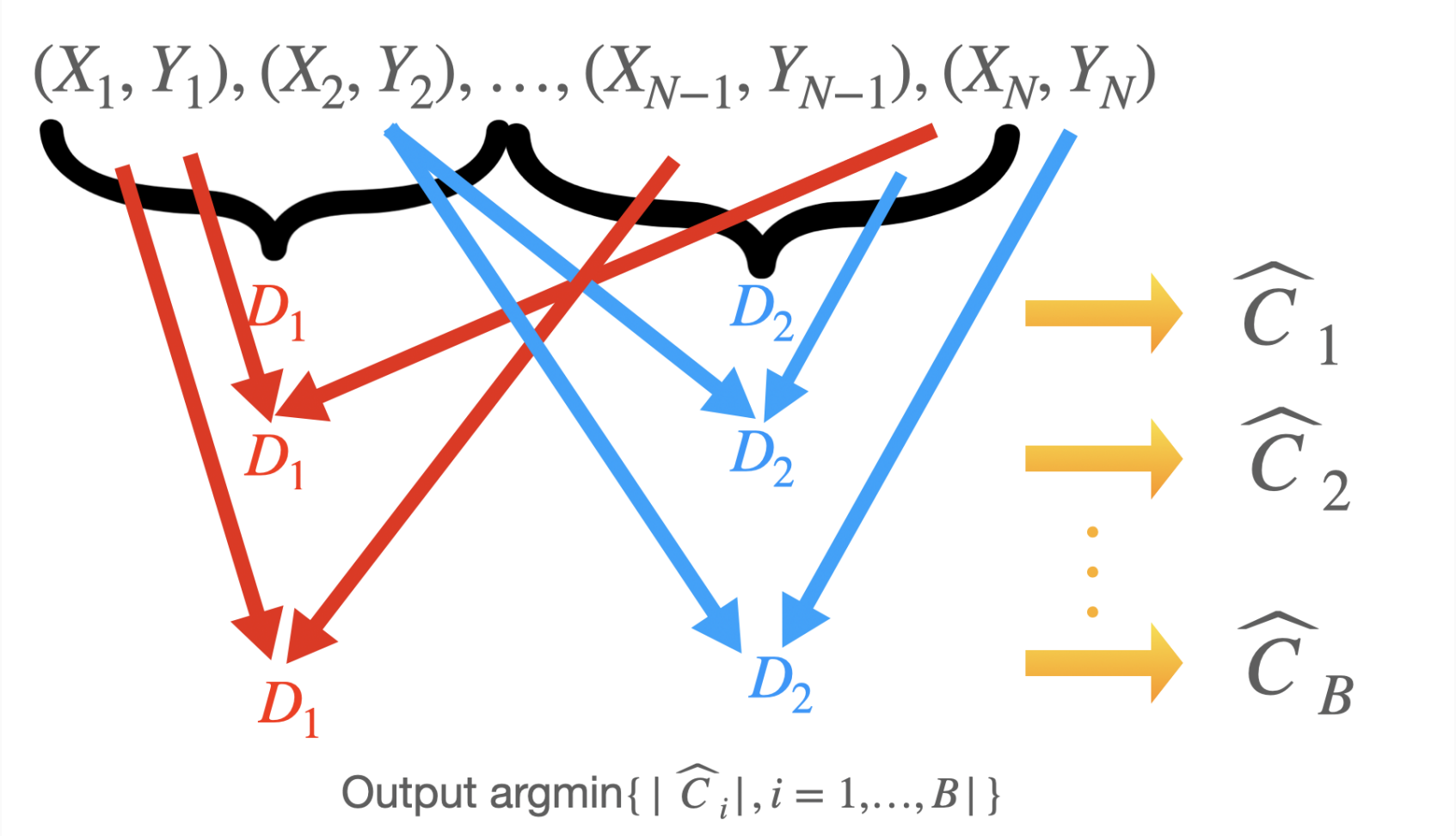
An alternative way of doing ensembling

The end goal is the prediction set, so instead of aggregating estimators, would it be better to aggregate the prediction sets obtained from different splits of the data?

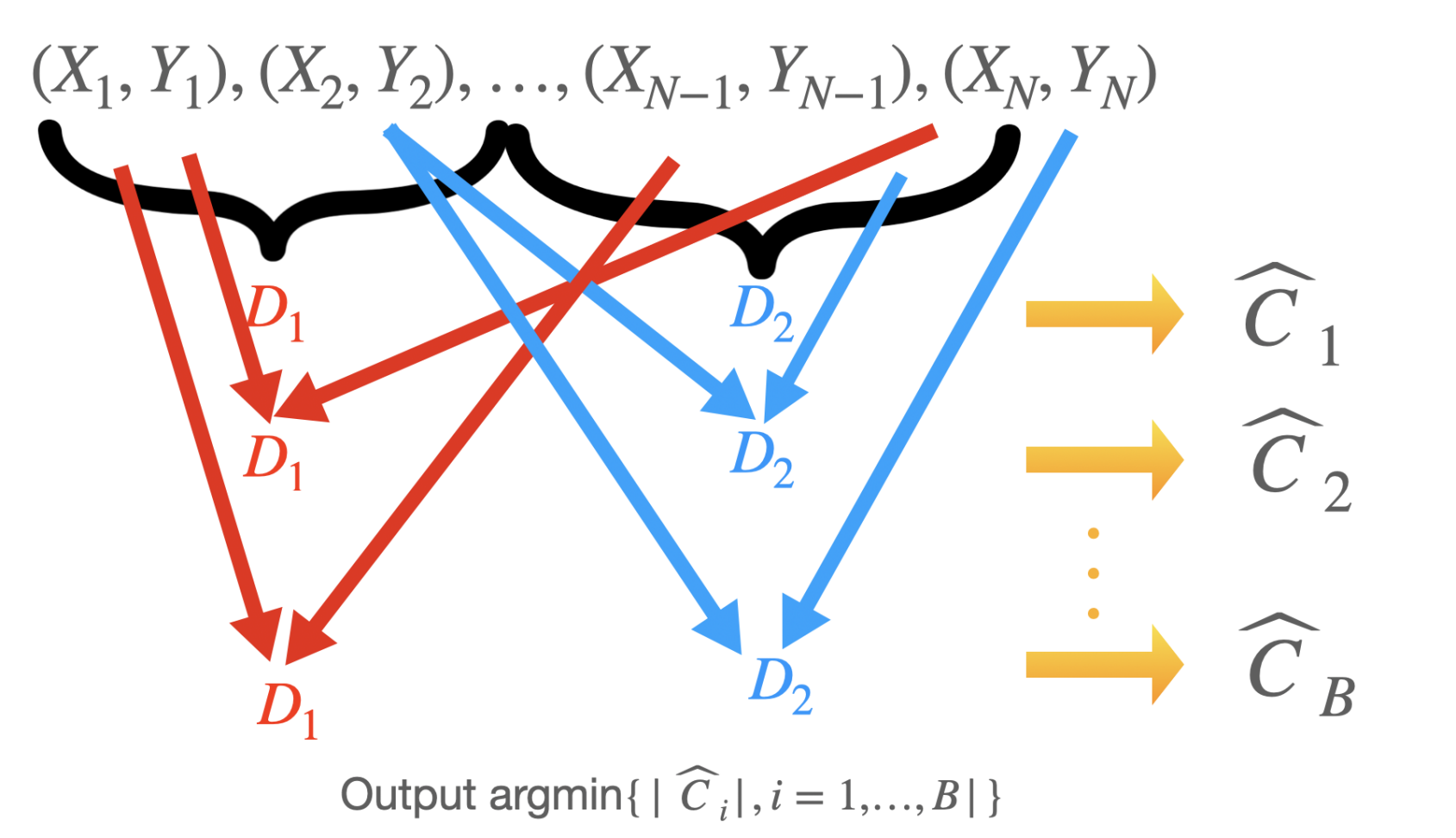
- One way is to do different splits of the data where we perform split conformal prediction on each split and choose the prediction set with the smallest size.
  - This goes into the class of efficiency first comformal prediction(EFCP) that we proposed in *Finite-sample Efficient Conformal Prediction*.[1].

[1] Finite-sample Efficient Conformal Prediction, Yachong Yang, Arun Kumar Kuchibhotla, arxiv 2021.

Aggregate prediction sets



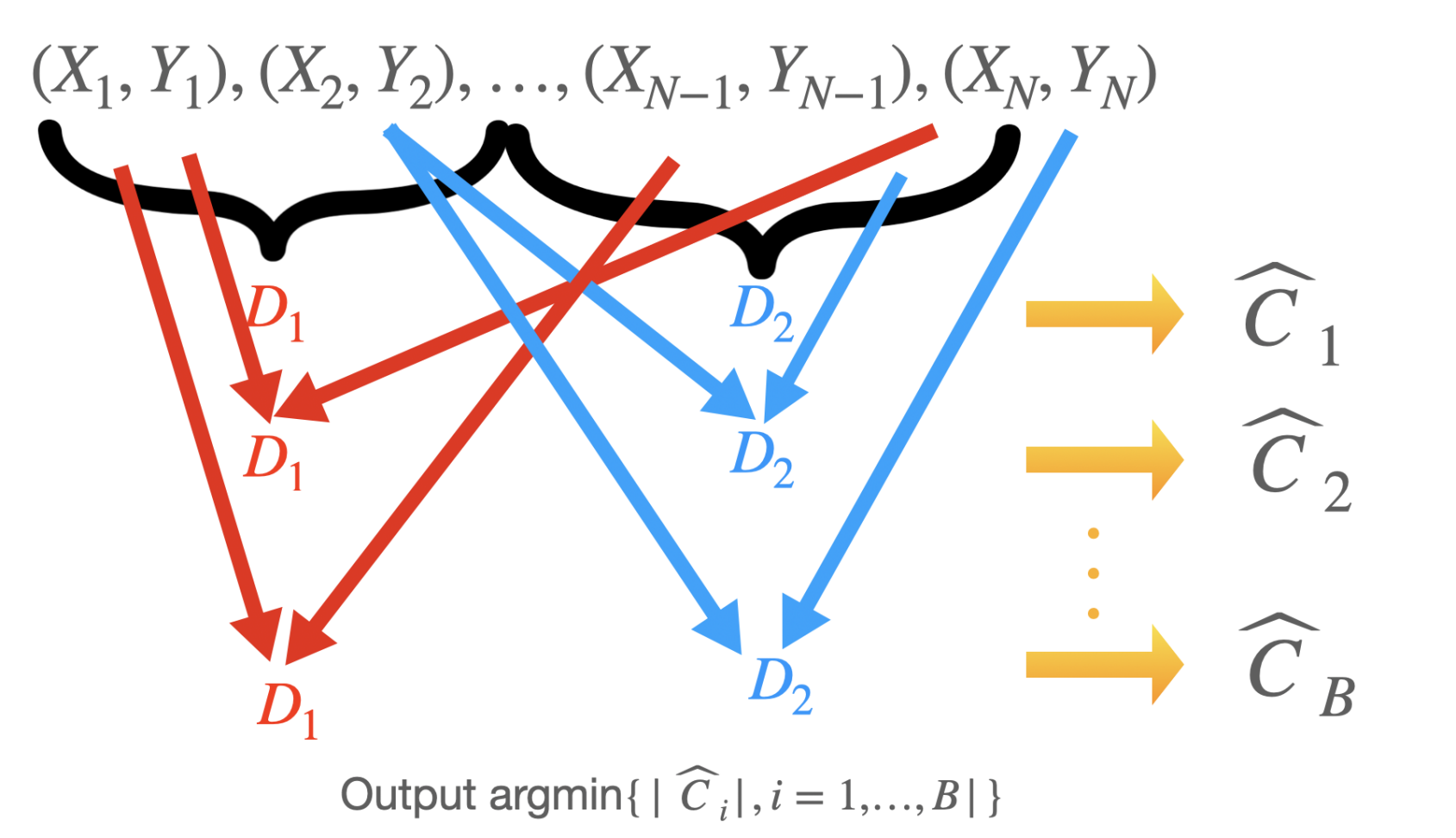
Aggregate prediction sets



Validity:

$$\left| \mathbb{P}\left((X_{N+1}, Y_{N+1}) \in \widehat{C}_N \right) - \frac{\lceil (1-\alpha)(1+|\mathcal{D}_2|) \rceil}{|\mathcal{D}_2|} \right| \leq \frac{\sqrt{\log(2B)/2} + 1/3}{\sqrt{|\mathcal{D}_2|}}.$$

Aggregate prediction sets



- Upon doing simulations we find that setting $B = 10$ in J+aB has the same computation time as doing 10 differents splits in the proposed method. But this new method has an upper bound with respect to coverage so it won't get conservative.

In conclusion

- J+aB has great computational efficiency in the setting of ensemble learning and has assumption-free coverage guarantee.
- It opens up the question of when practitioners would need to do ensembling and how does it compare to other methods?

Congratulations on this elegant and thought-provoking paper!