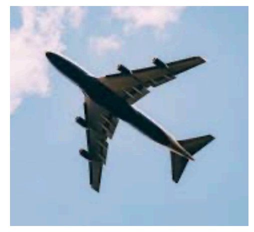
Robust Conformal Prediction With a Single Binary Certificate

Soroush H. Zargarbashi, Aleksandar Bojchevski

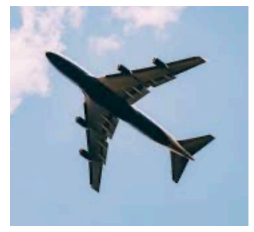
CISPA Helmholtz Center for Information Security
University of Cologne

What is Conformal Prediction?

Instead of a single prediction (unknown accuracy), predict a set of labels, guaranteed to include the true answer.



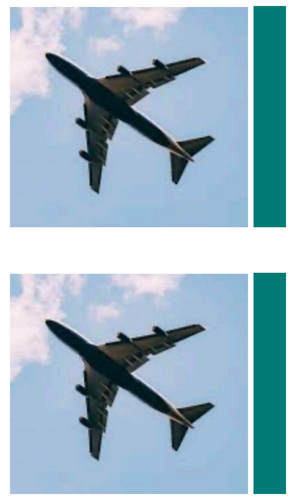
Airplane Unknown accuracy



Airplane, or car 90% Coverage

What is Conformal Prediction?

Instead of a single prediction (unknown accuracy), predict a set of labels, guaranteed to include the true answer.

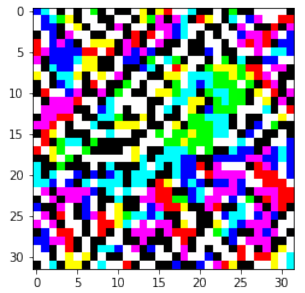


	Airplane	Unknown accuracy
	Airplane, or car	90% Coverage

Conformal Guarantee can break!

Small noise can decrease the coverage drastically!

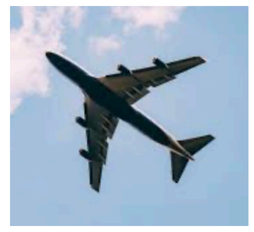
$\Pr[\text{airplane} \in \mathcal{C}($ $)] \geq 1 - \alpha$

$\Pr[\text{airplane} \in \mathcal{C}($ $+\, 0.001$  $)] \not\geq 1 - \alpha$

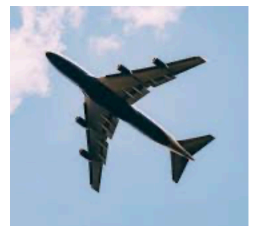
TL;DR

What is Conformal Prediction?

Instead of a single prediction (unknown accuracy), predict a set of labels, guaranteed to include the true answer.



Airplane Unknown accuracy

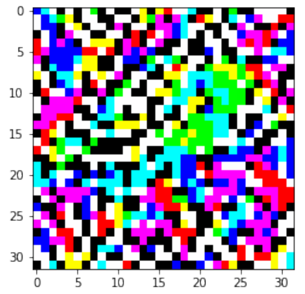


Airplane, or car 90% Coverage

Conformal Guarantee can break!

Small noise can decrease the coverage drastically!

$\Pr[\text{airplane} \in \mathcal{C}($ $)] \geq 1 - \alpha$

$\Pr[\text{airplane} \in \mathcal{C}($ $+ 0.001$  $)] \not\geq 1 - \alpha$

Robust Conformal Prediction

Same guarantee, extended to the worst case noise.

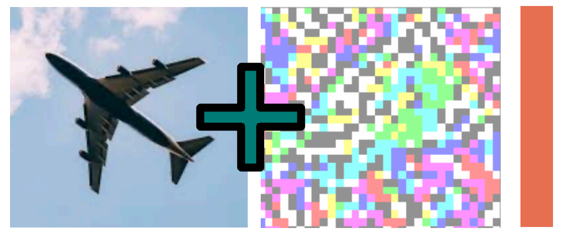
$$\Pr[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{\text{noisy}}), \boldsymbol{x}_{\text{noisy}} \in \mathcal{B}(\boldsymbol{x}_{\text{clean}})] \geq 90\%$$

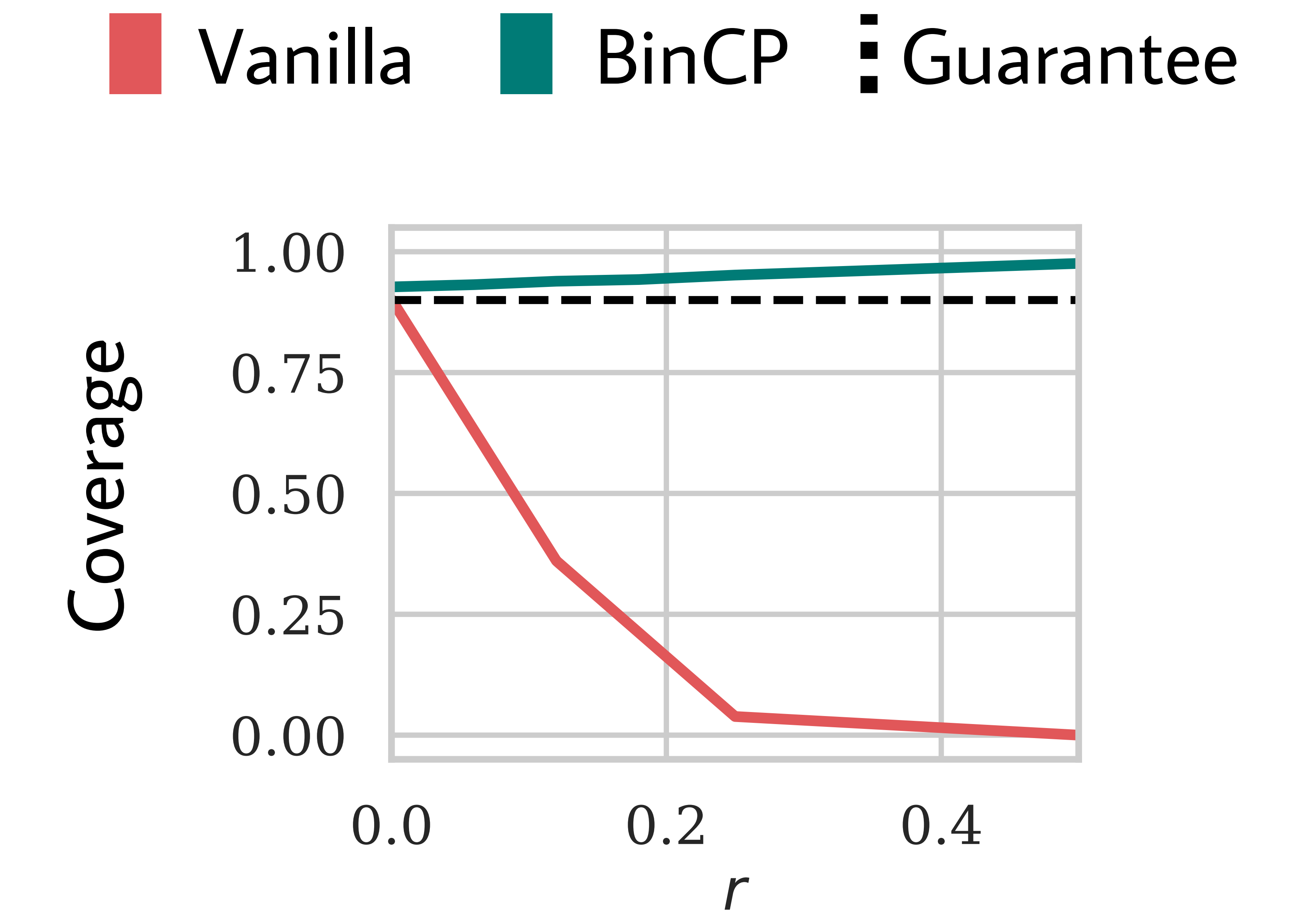
TL;DR

Our Results in One Glance

Prediction sets guaranteed to include the true label even with the worst case noise!

{Airplane, car} 90% Coverage

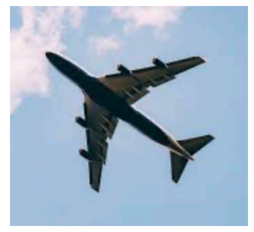

{Plane, car, cat} Still 90% Coverage


CIFAR-10 under adversarial perturbation

TL;DR

Our Results in One Glance

Prediction sets guaranteed to include the true label even with the worst case noise!

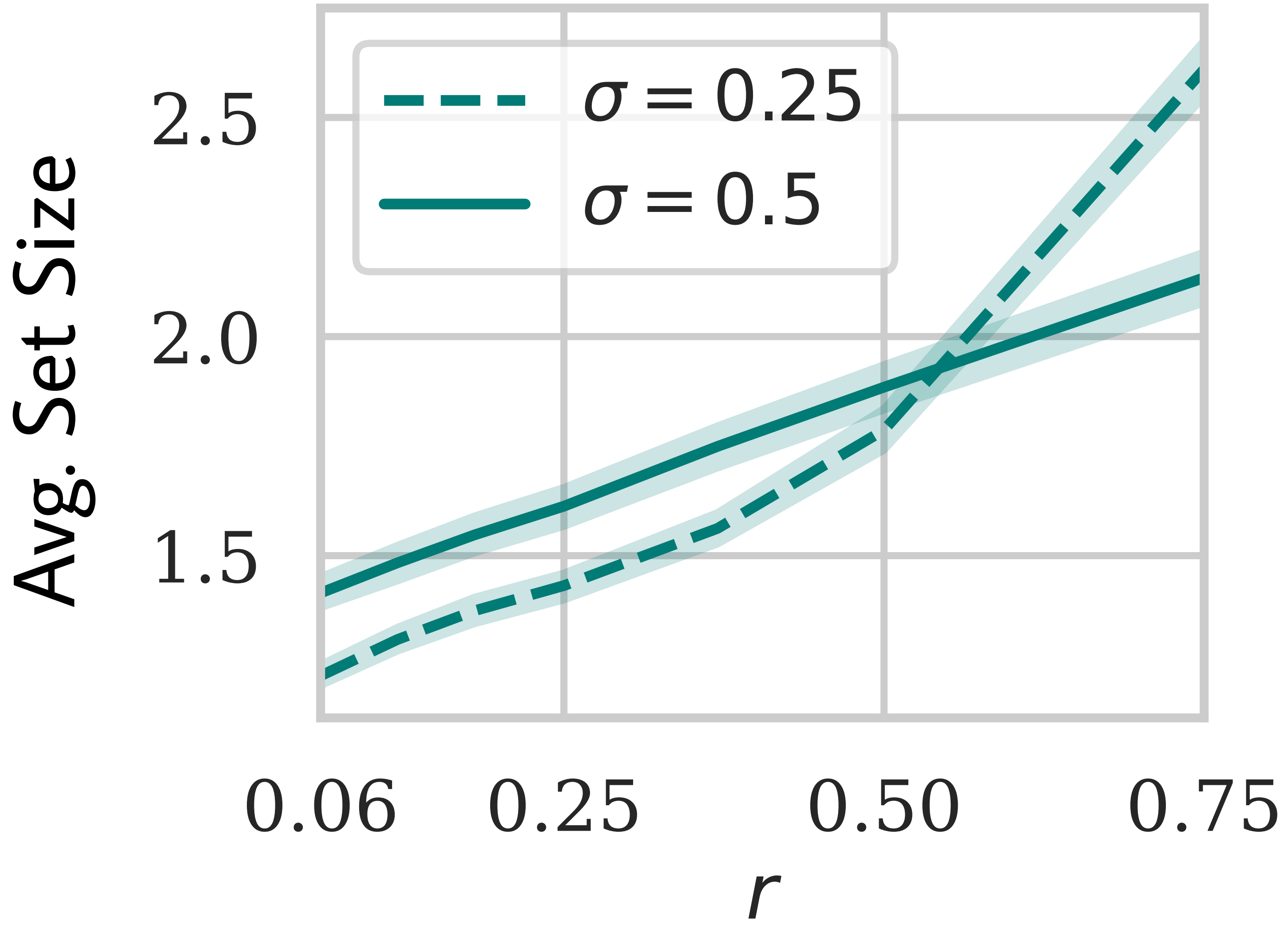


Airplane, or car 90% Coverage

Airplane, car, or cat Still 90% Coverage

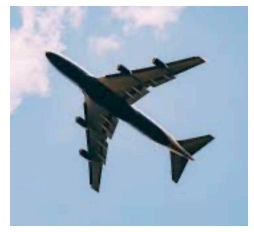
Works with many existing smoothing certificates.

Using L1 De-randomized Certificate



Our Results in One Glance

Prediction sets guaranteed to include the true label even with the worst case noise!

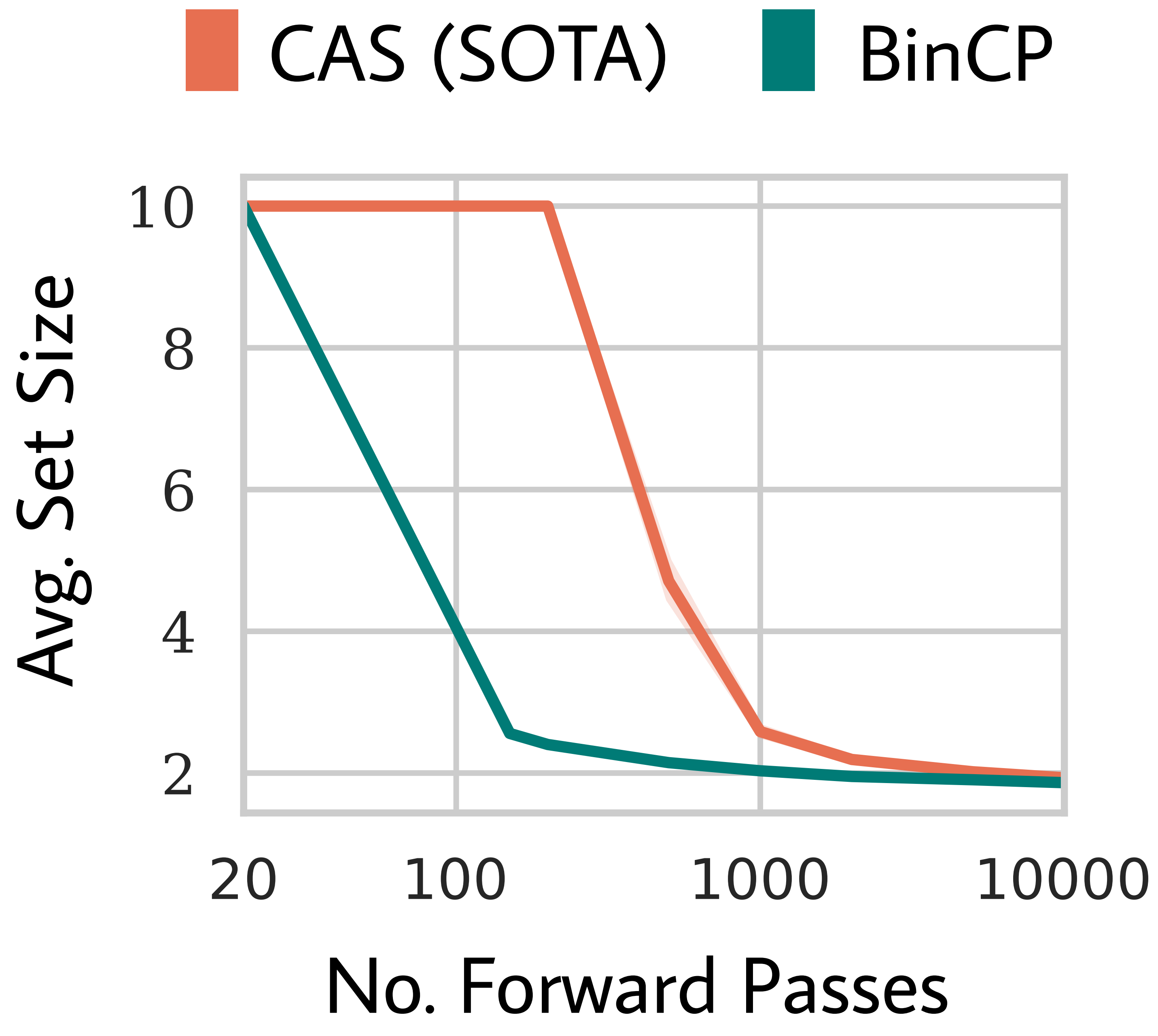


Airplane, or car 90% Coverage

Airplane, car, or cat Still 90% Coverage

Works with many existing smoothing certificates.

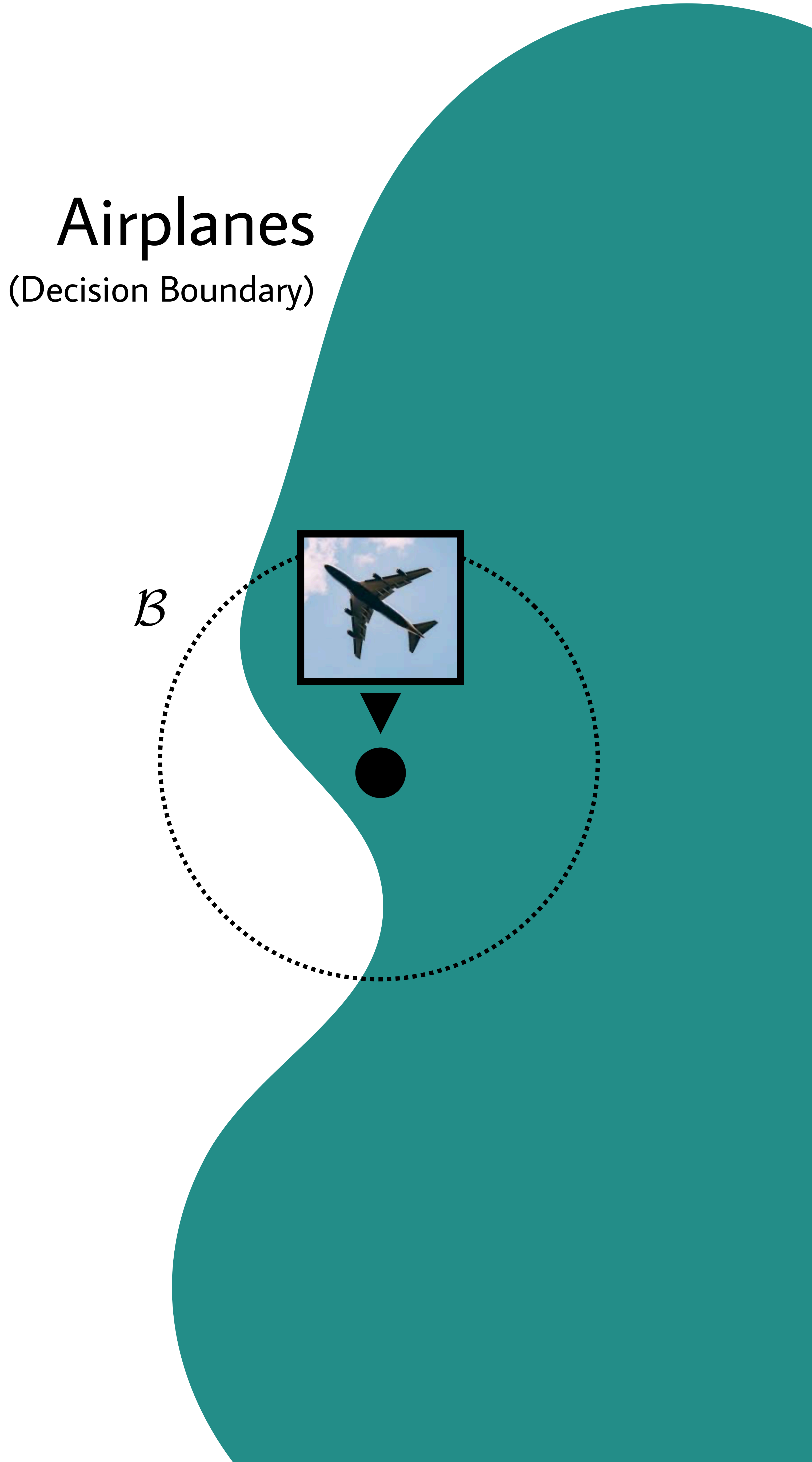
Returns smaller sets with fewer forward passes.



Robust Conformal Prediction

Same guarantee, extended to the worst case noise.

$$\Pr[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{\text{noisy}}), \boldsymbol{x}_{\text{noisy}} \in \mathcal{B}(\boldsymbol{x}_{\text{clean}})] \geq 90\%$$

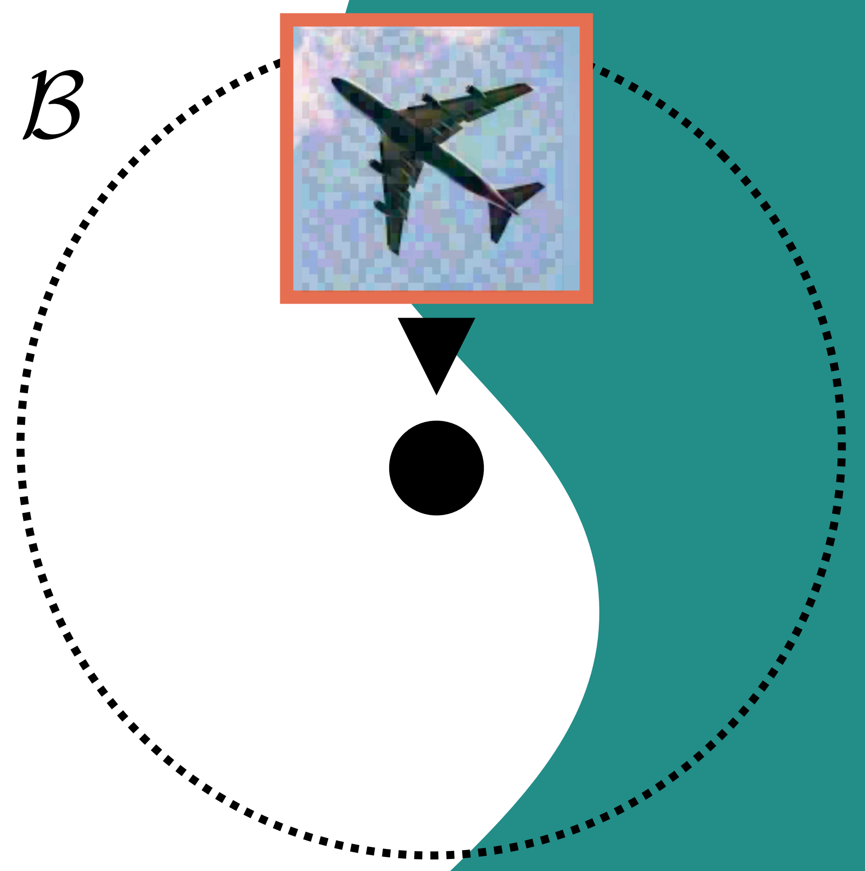


Robust Conformal Prediction

Same guarantee, extended to the worst case noise.

$$\Pr[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{\text{noisy}}), \boldsymbol{x}_{\text{noisy}} \in \mathcal{B}(\boldsymbol{x}_{\text{clean}})] \geq 90\%$$

Worst case noise: Small perturbation, same semantics, different prediction!



Robust Conformal Prediction

Same guarantee, extended to the worst case noise.

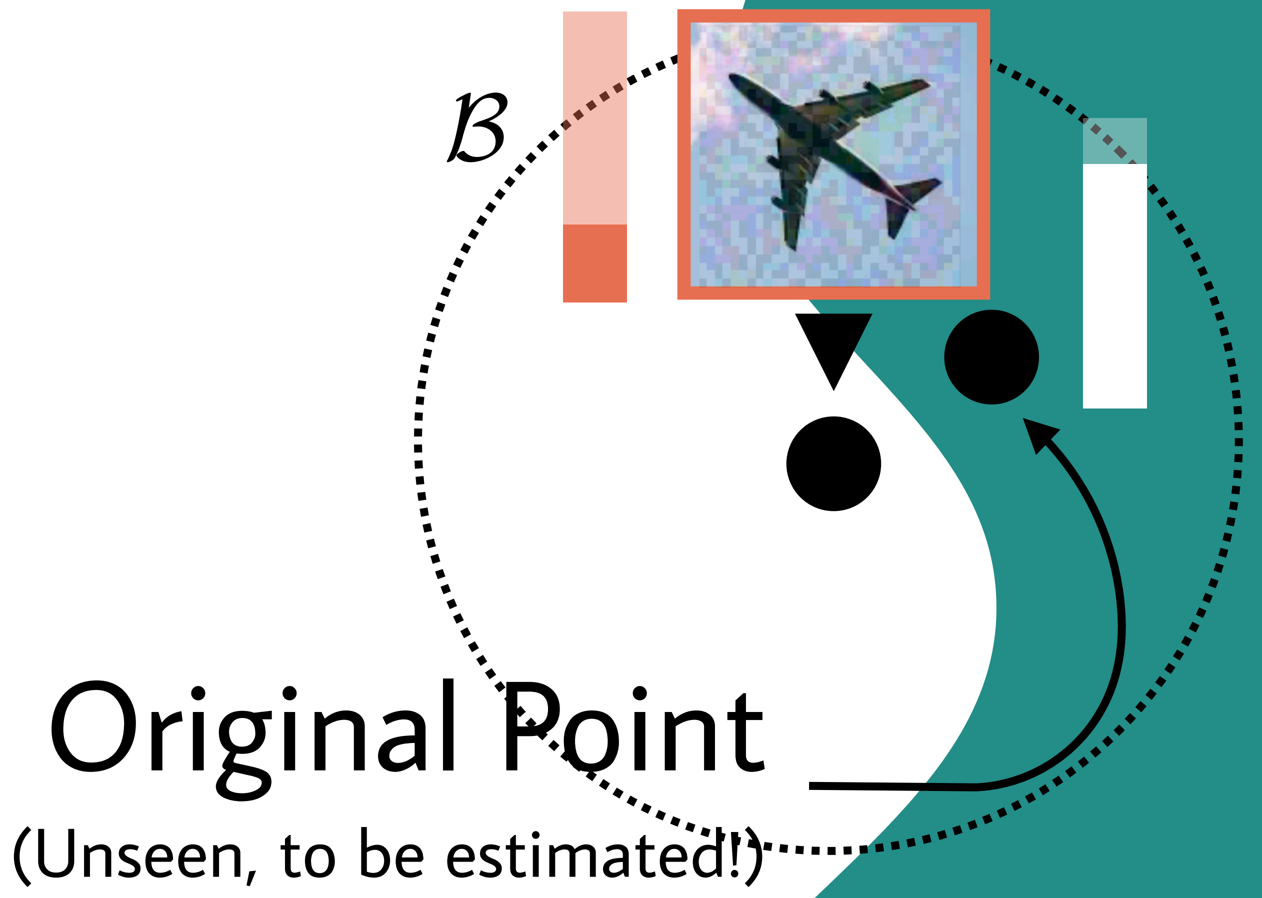
$$\Pr[y_{n+1} \in \mathcal{C}(\boldsymbol{x}_{\text{noisy}}), \boldsymbol{x}_{\text{noisy}} \in \mathcal{B}(\boldsymbol{x}_{\text{clean}})] \geq 90\%$$

Worst case noise: Small perturbation, same semantics, different prediction!

How CP Works?

A score function captures agreement between input, and labels: e.g. softmax.

We accept any label with score above some computed threshold τ.

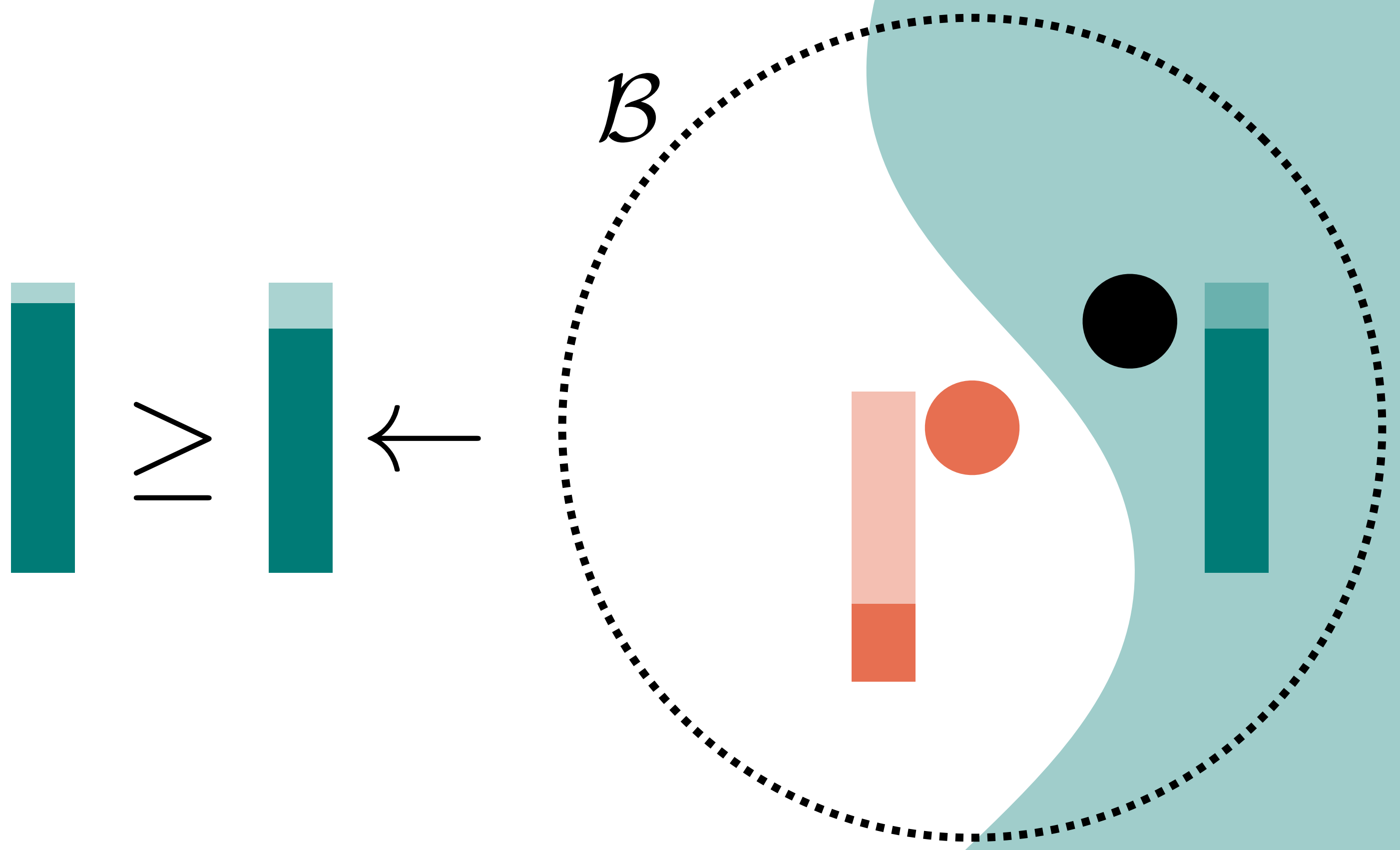


How Robust CP Works?

Upper bound the maximum score in the perturbation ball! How?

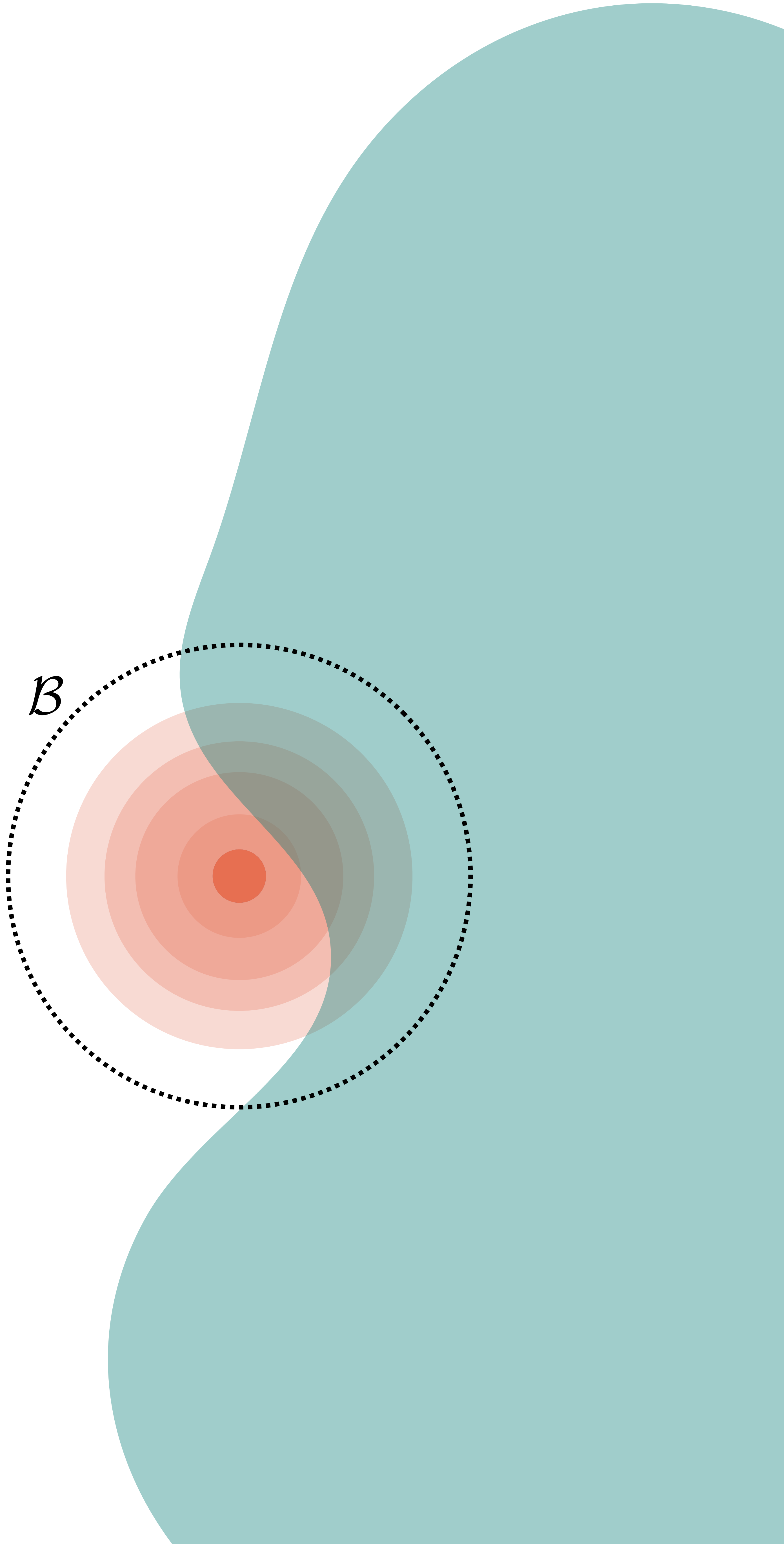
Neural network verifiers? Limited!

What else? Smoothing!



What else? Smoothing!

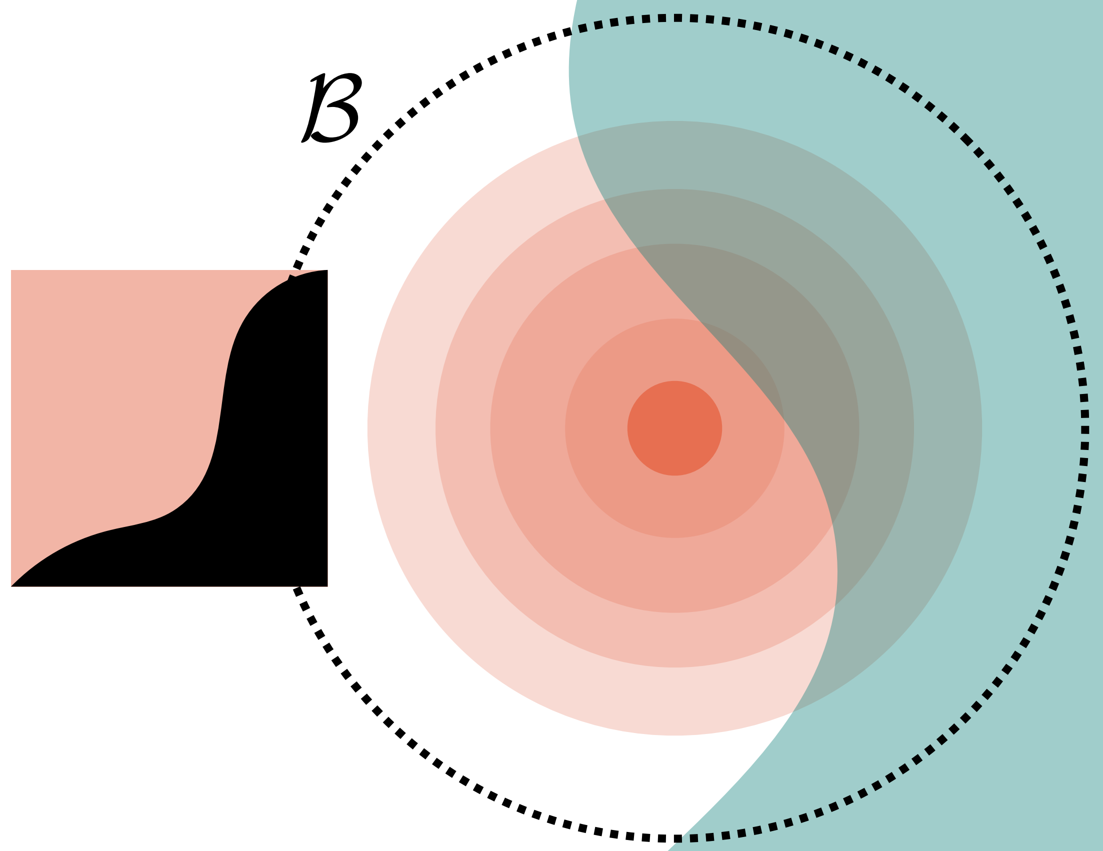
Instead one point look at the distribution of nearby points!



What else? Smoothing!

Instead one point look at the distribution of nearby points!

$$\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$$

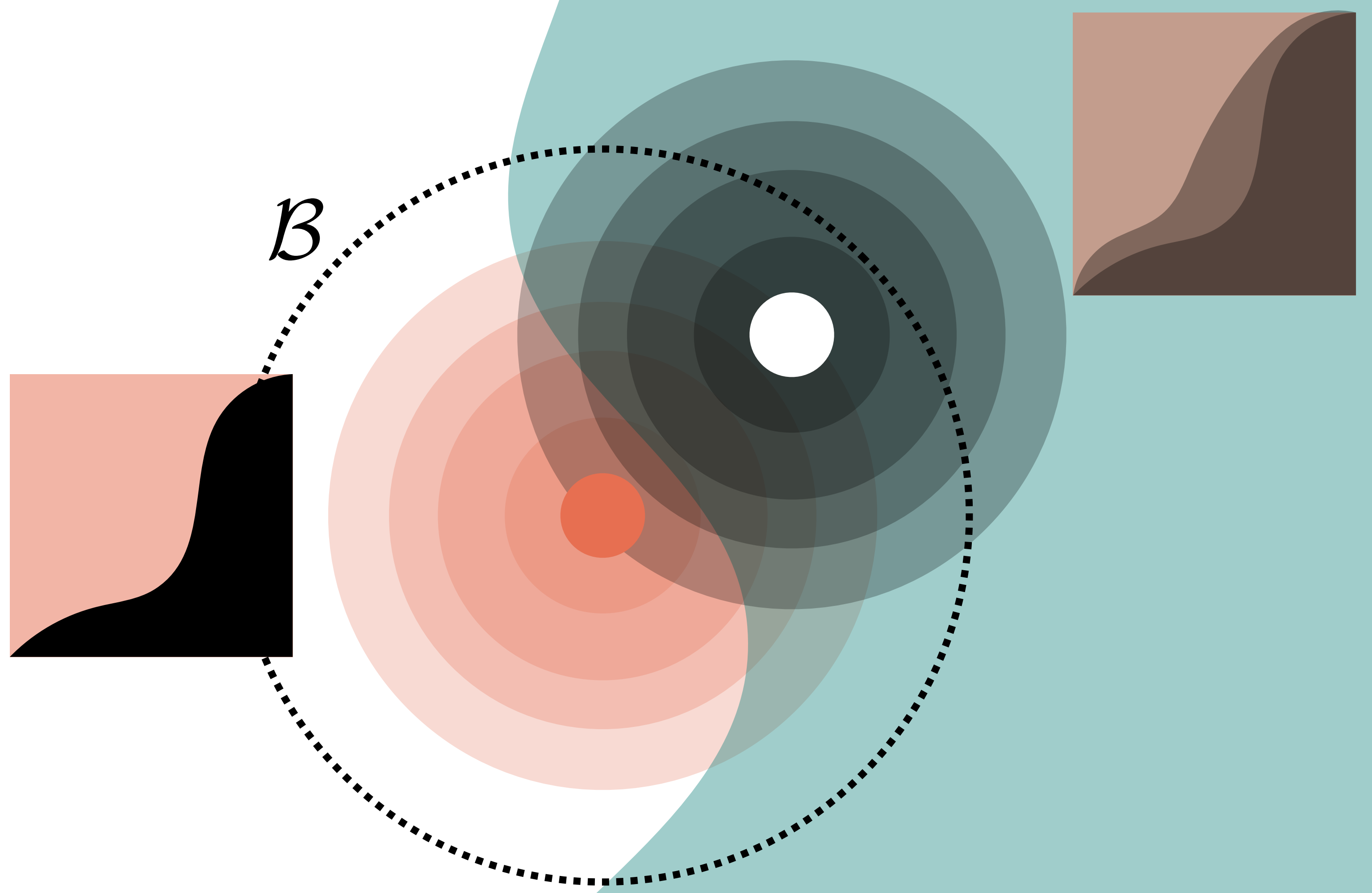


What else? Smoothing!

Instead one point look at the distribution of nearby points!

$$\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$$

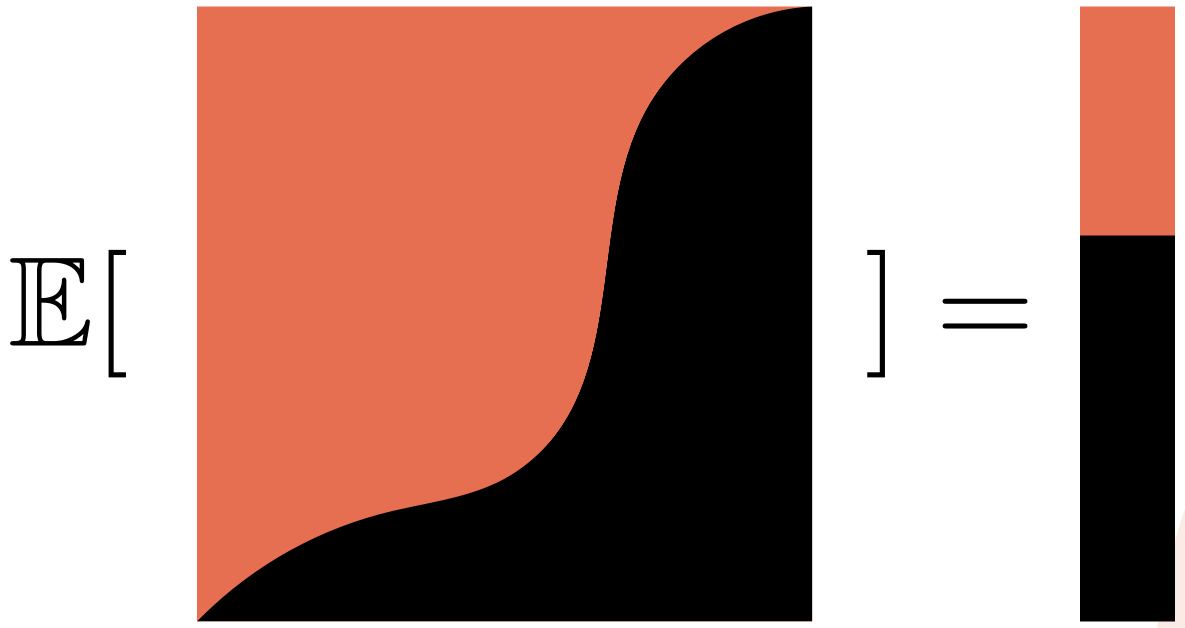
This distribution changes slowly around the input $\boldsymbol{x}_i$.



What score to choose?

How to summarize a distribution $\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$

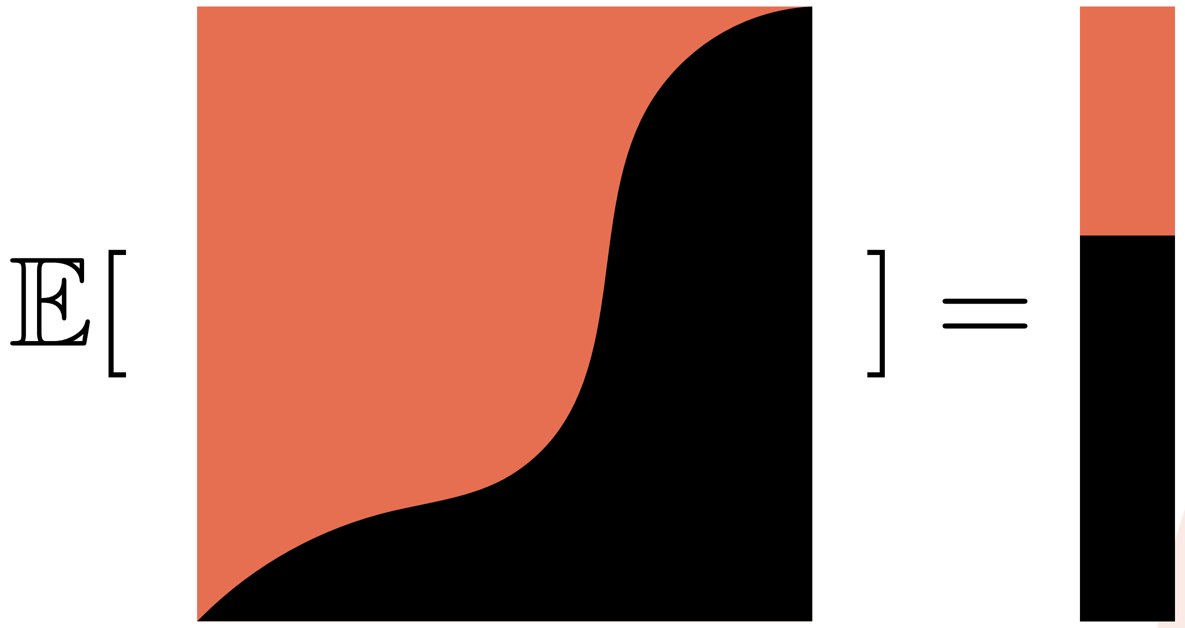
Into a single number? Baseline: Take the mean! $\mathbb{E}[\mathcal{S}_i]$



What score to choose?

How to summarize a distribution $\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$

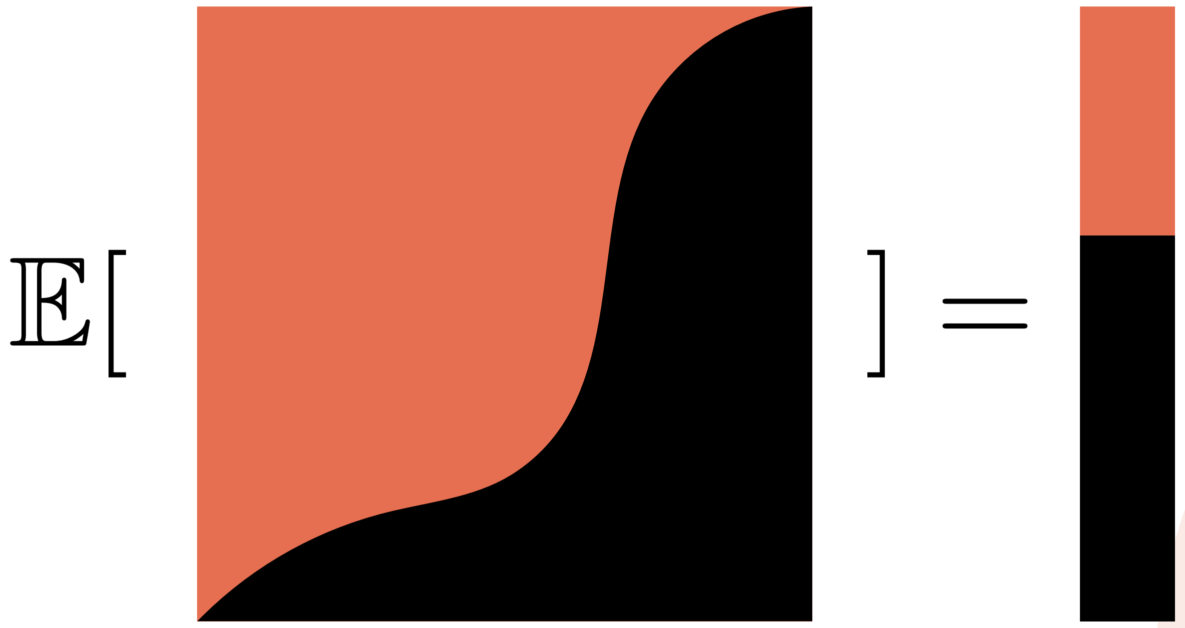
Into a single number? Baseline: Take the mean! $\mathbb{E}[\mathcal{S}_i]$

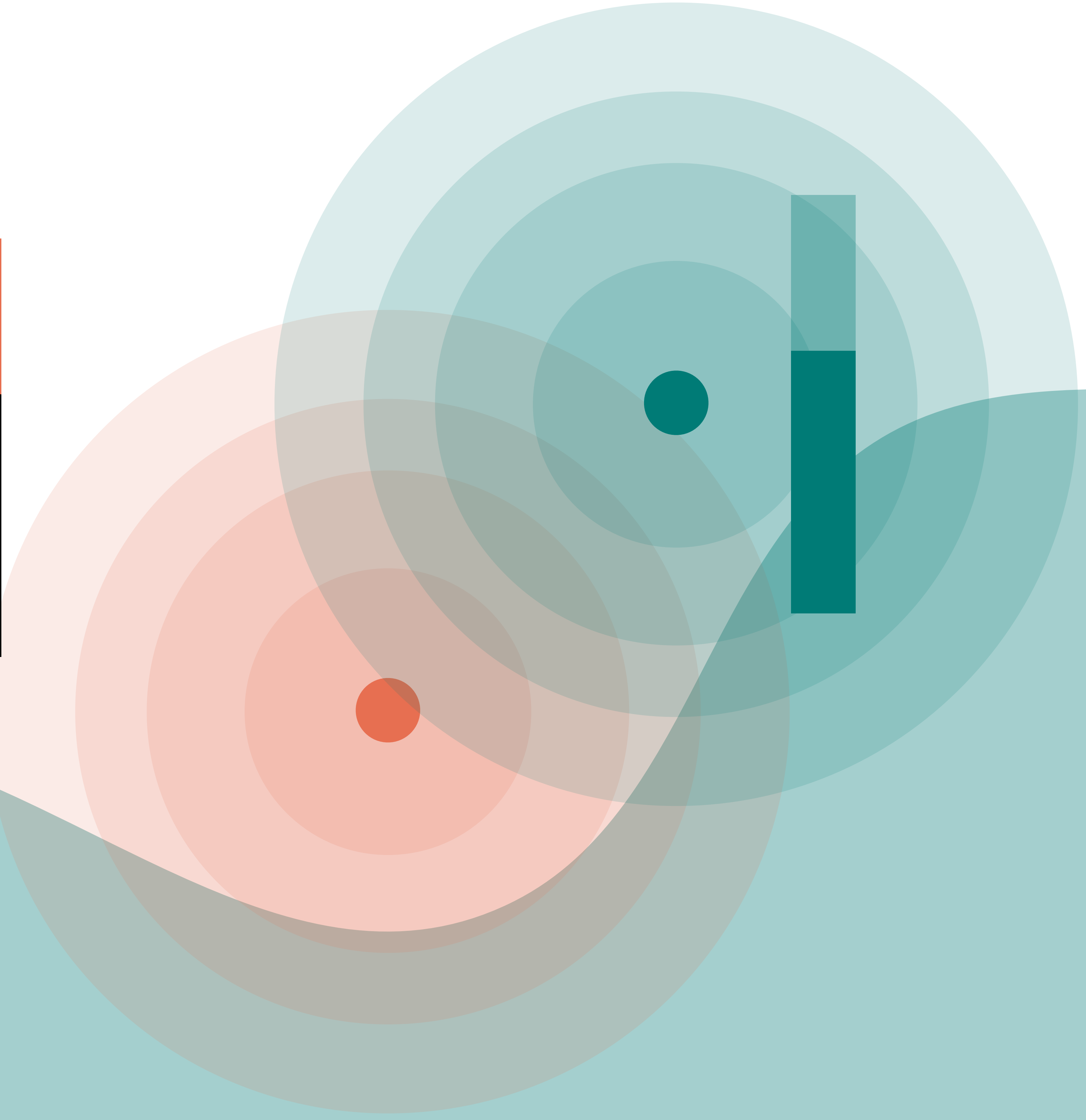


What score to choose?

How to summarize a distribution $\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$

Into a single number? Baseline: Take the mean! $\mathbb{E}[\mathcal{S}_i]$

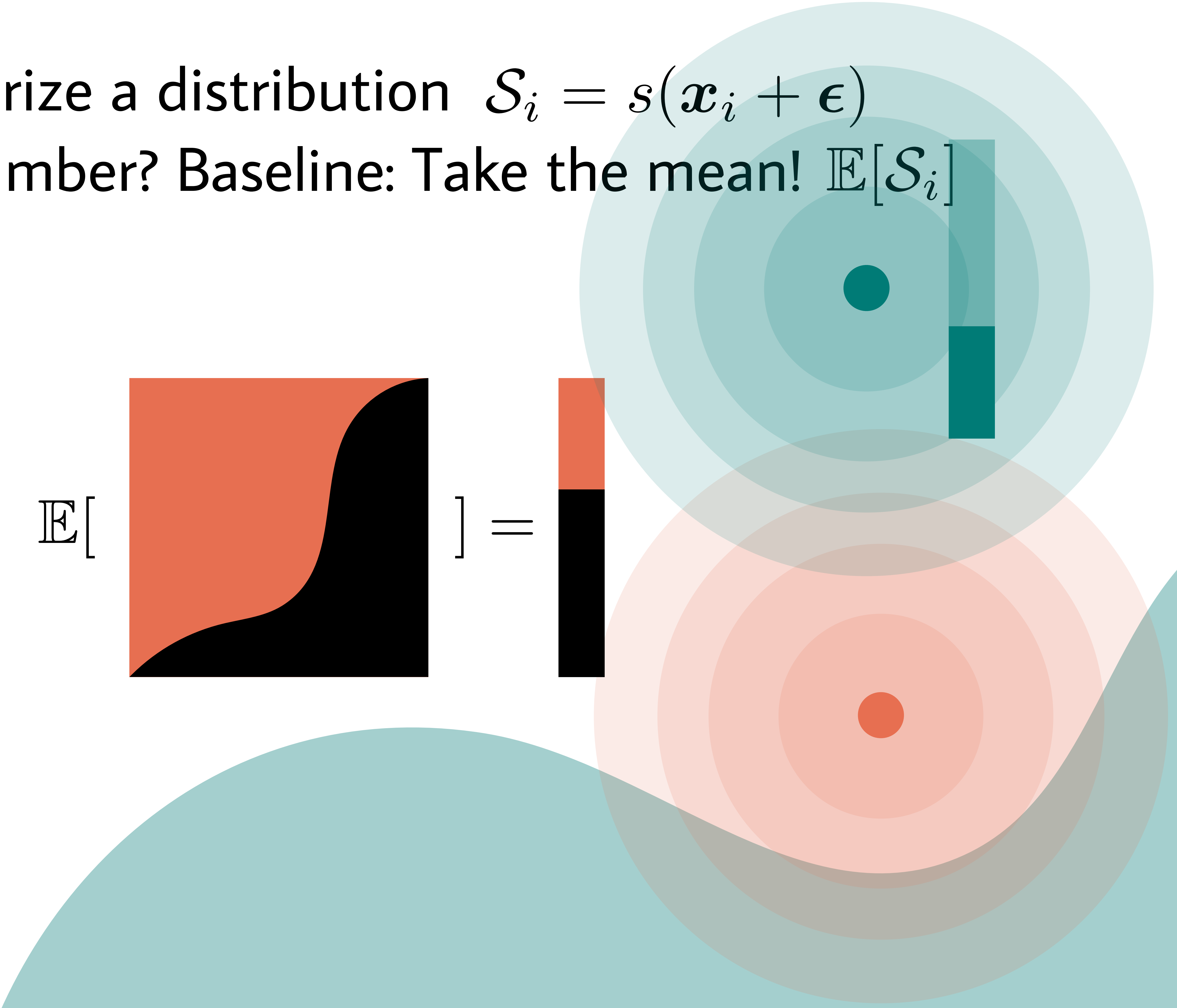




What score to choose?

How to summarize a distribution $\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$

Into a single number? Baseline: Take the mean! $\mathbb{E}[\mathcal{S}_i]$

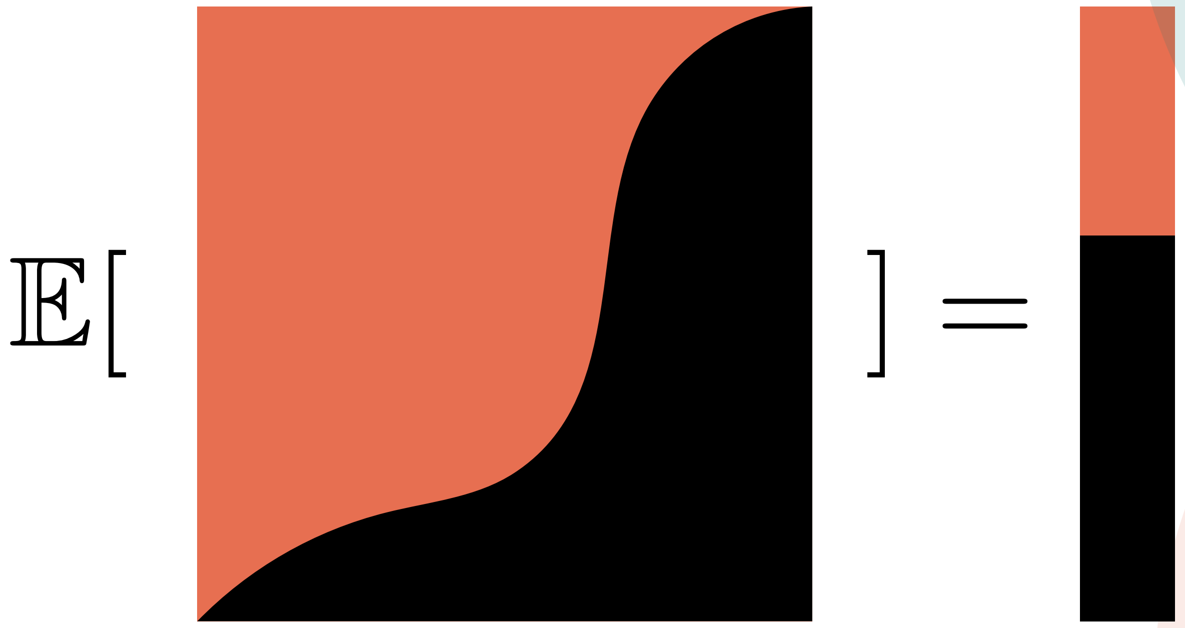


What score to choose?

How to summarize a distribution $\mathcal{S}_i = s(\boldsymbol{x}_i + \boldsymbol{\epsilon})$

Into a single number? Baseline: Take the mean! $\mathbb{E}[\mathcal{S}_i]$

Many Monte-Carlo samples are needed!

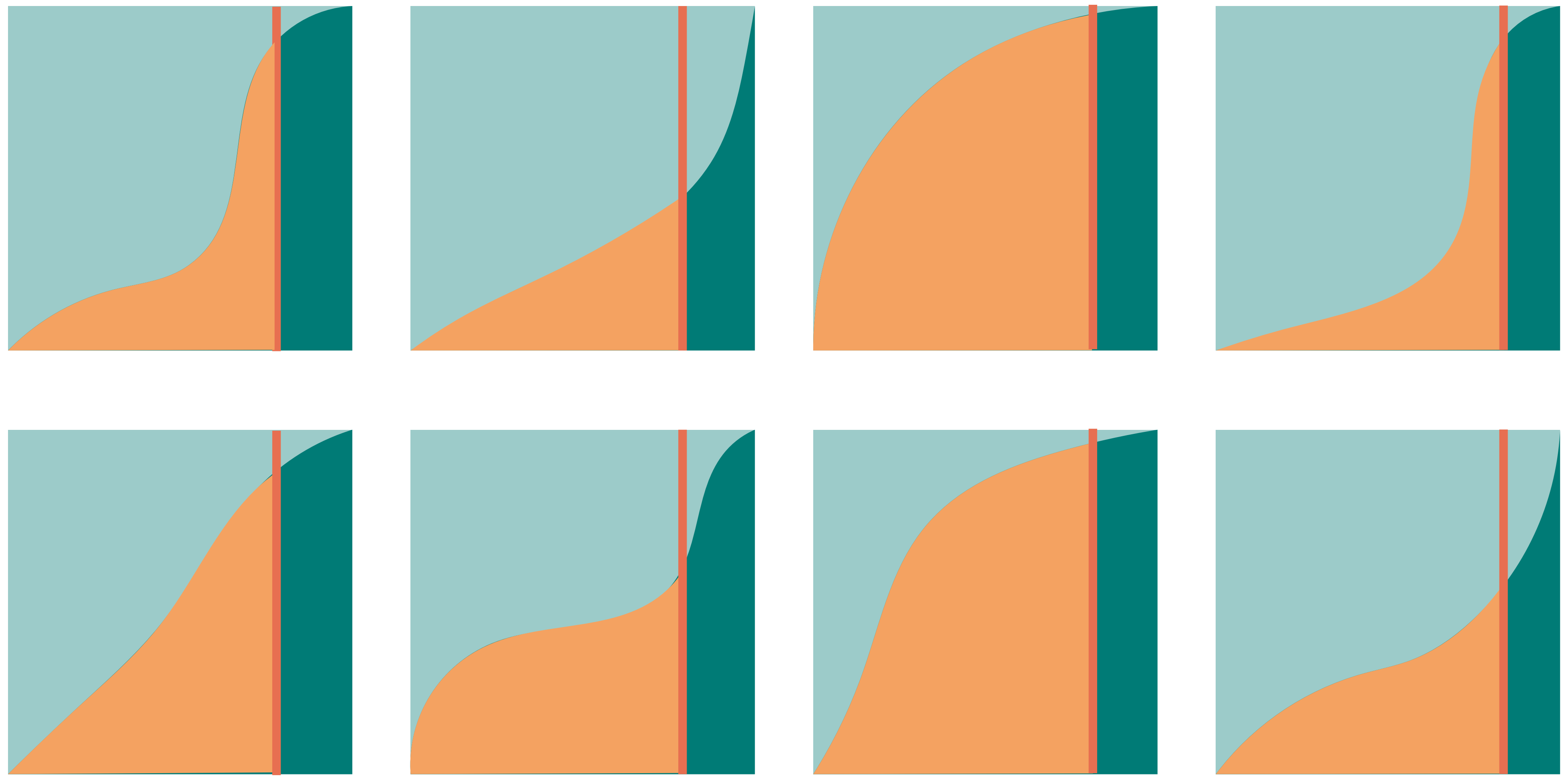


BinCP: Quantile of Quantiles

We take the quantile of this distribution as the conformity score.

BinCP: Quantile of Quantiles

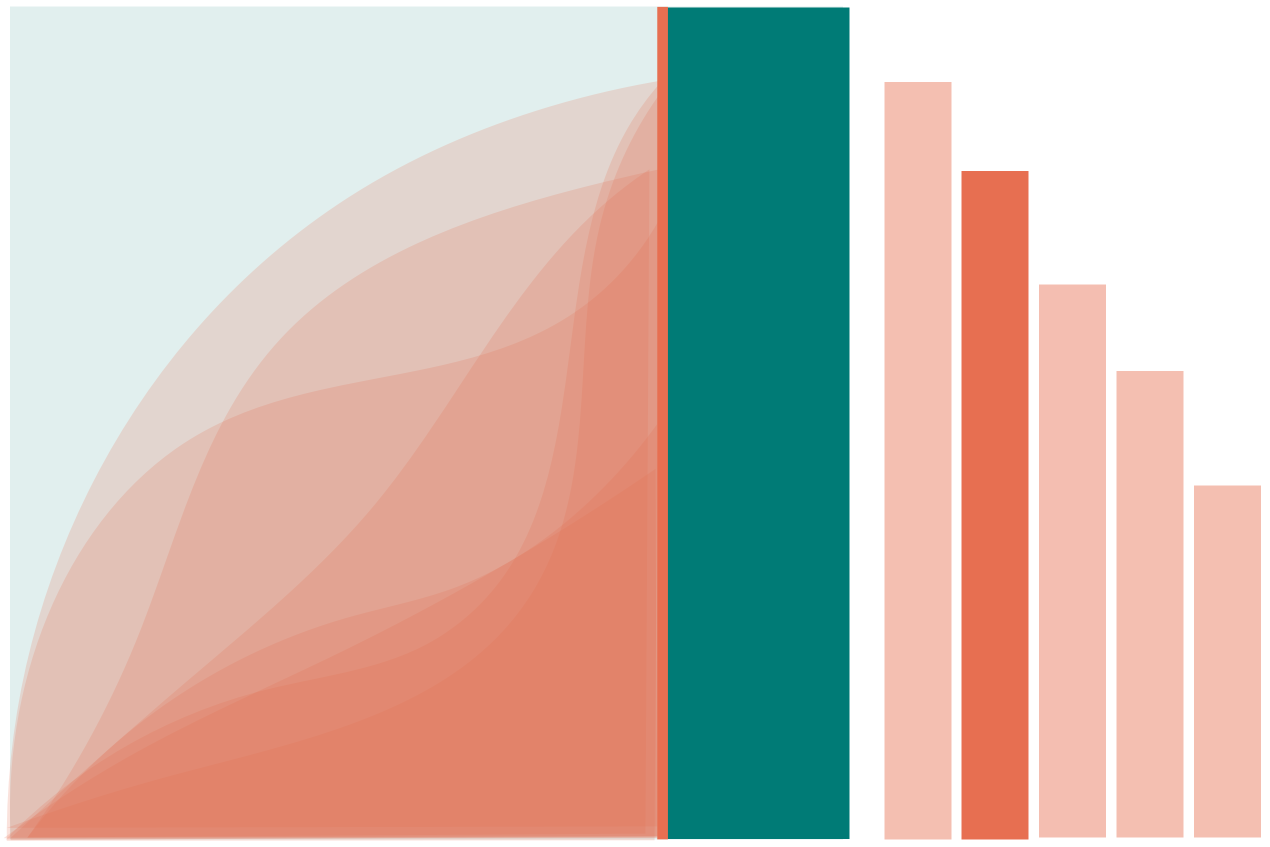
We take the quantile of this distribution as the conformity score.



BinCP: Quantile of Quantiles

We take the quantile of this distribution as the conformity score.

We compute the conformal quantile of quantiles.



BinCP: Quantile of Quantiles

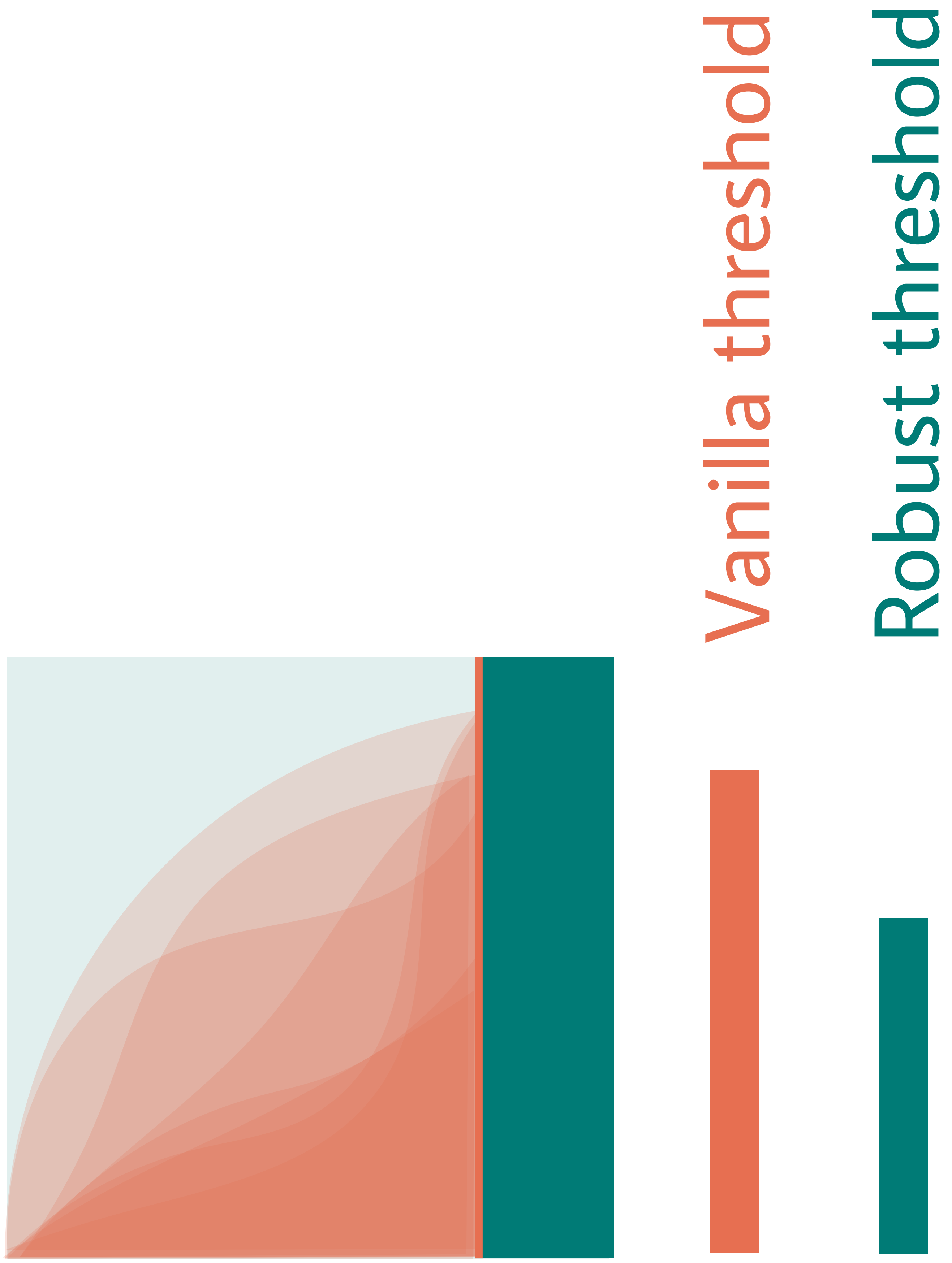
We take the quantile of this distribution as the conformity score.

We compute the conformal quantile of quantiles.

Robust BinCP

Turns out we only need one binary certificate. Read our paper to see why!

Due to the binary variable we can use tighter confidence intervals.

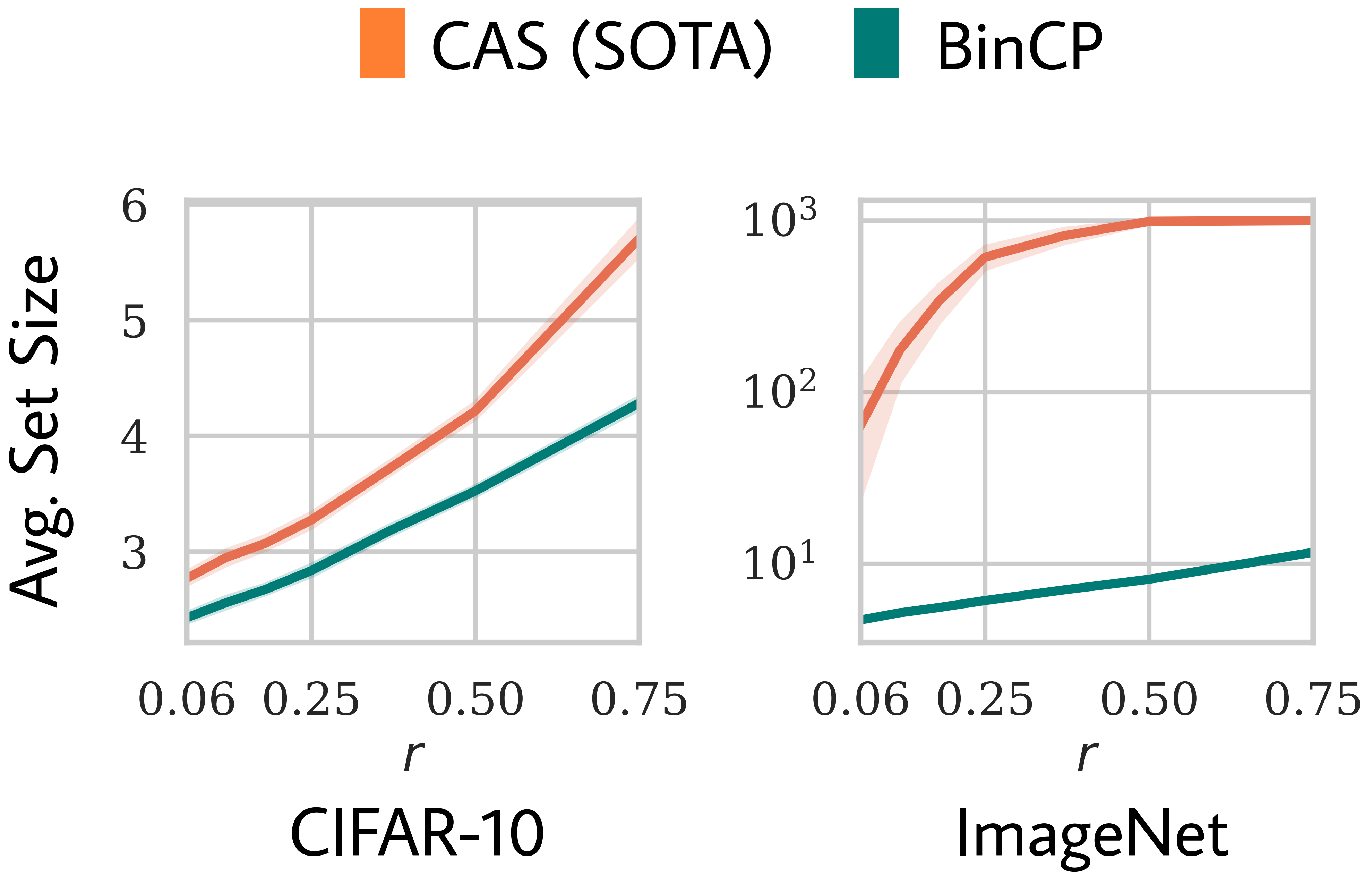


Results

Works on any binary certificate providing probability lower bounds.

Uses tighter confidence intervals

> Smaller prediction set for fewer forward passes.
>
> Even more improvements on datasets with many classes.

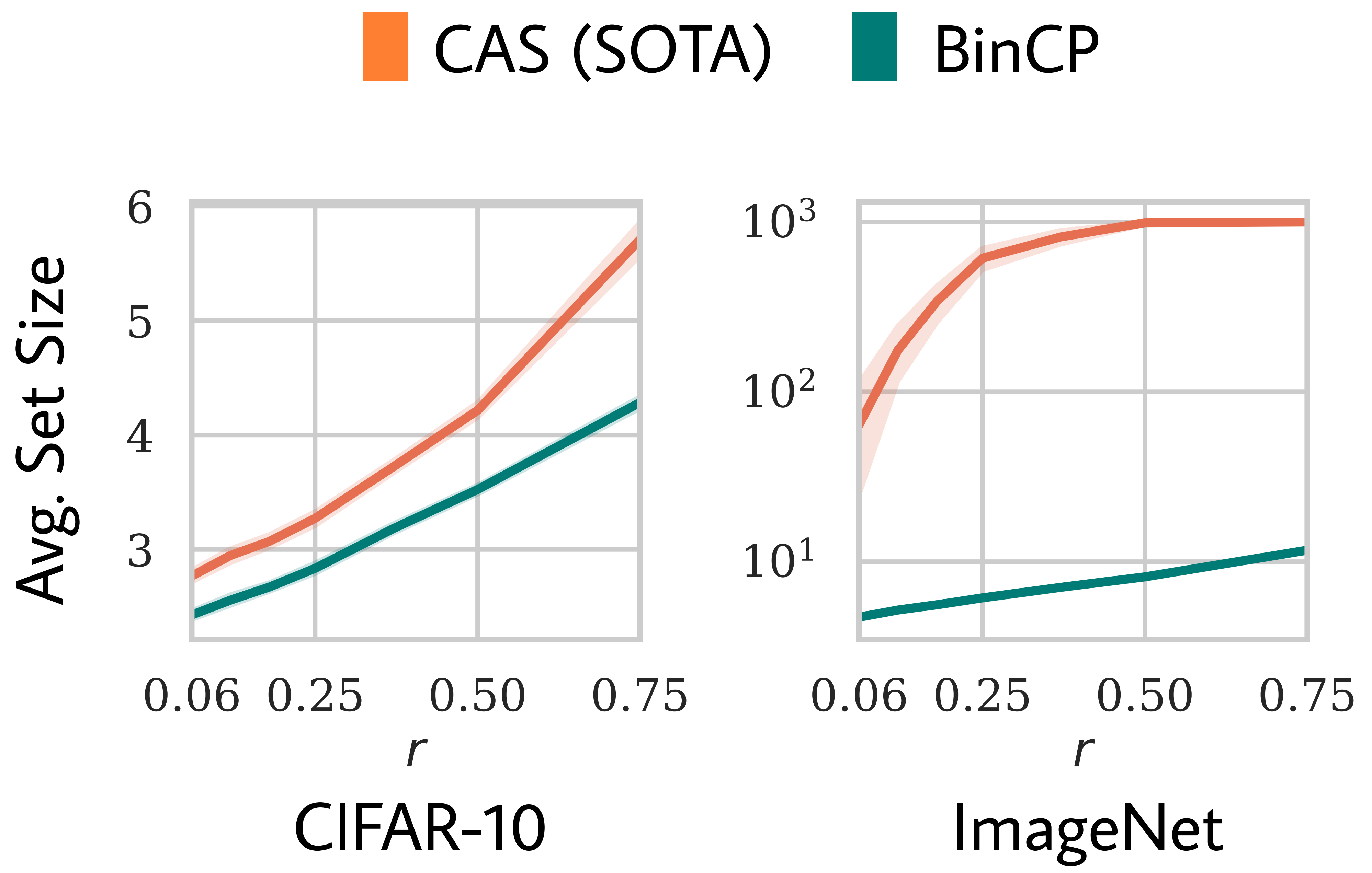


Results

Works on any binary certificate providing probability lower bounds.

Uses tighter confidence intervals

> Smaller prediction set for fewer forward passes.
>
> Even more improvements on datasets with many classes.



CIFAR-10

ImageNet

Paper, Code = ,