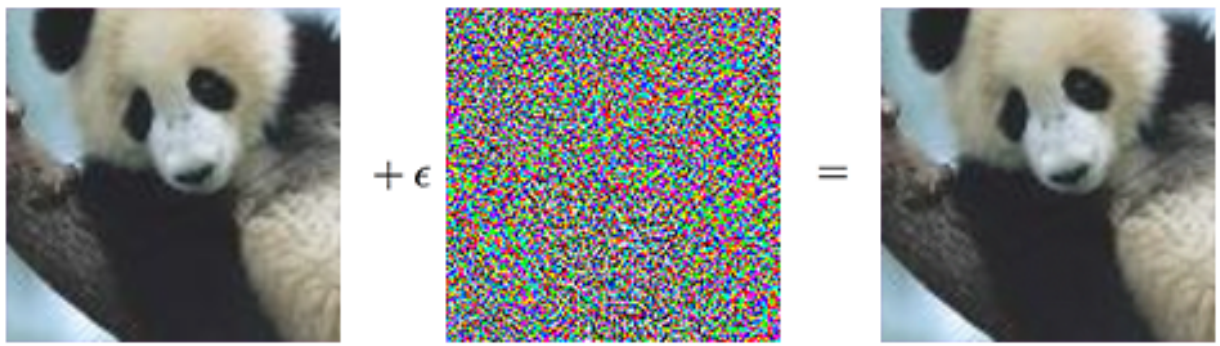
Stable Conformal Prediction Sets[1]

Eugene Ndiaye
Georgia Institute of Technology

[1] International Conference on Machine Learning, Baltimore, 2022

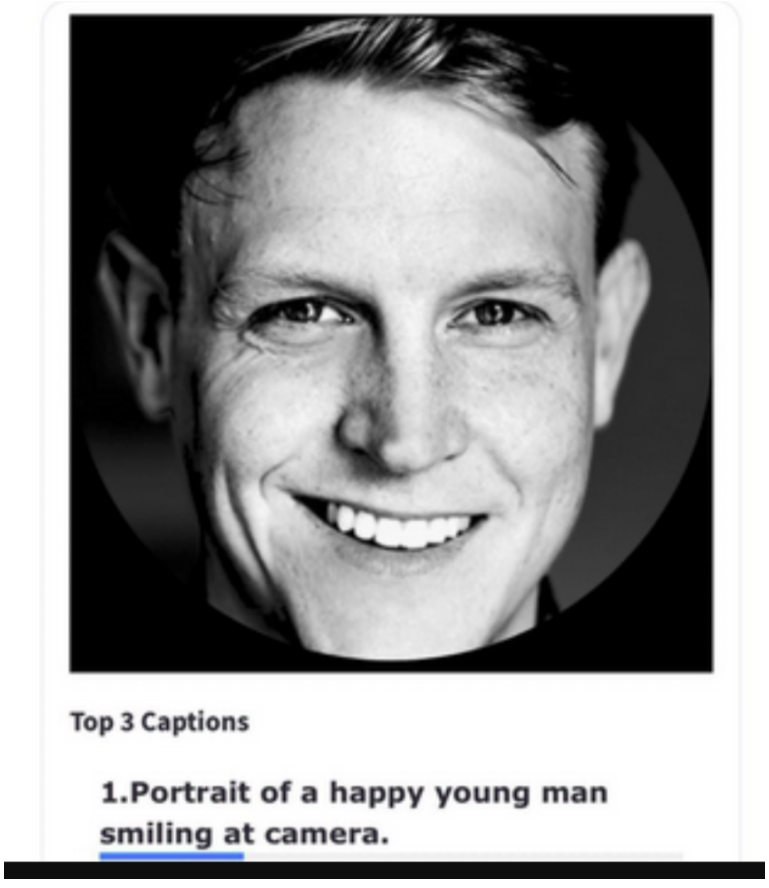
"Confident" prediction?[2]



An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

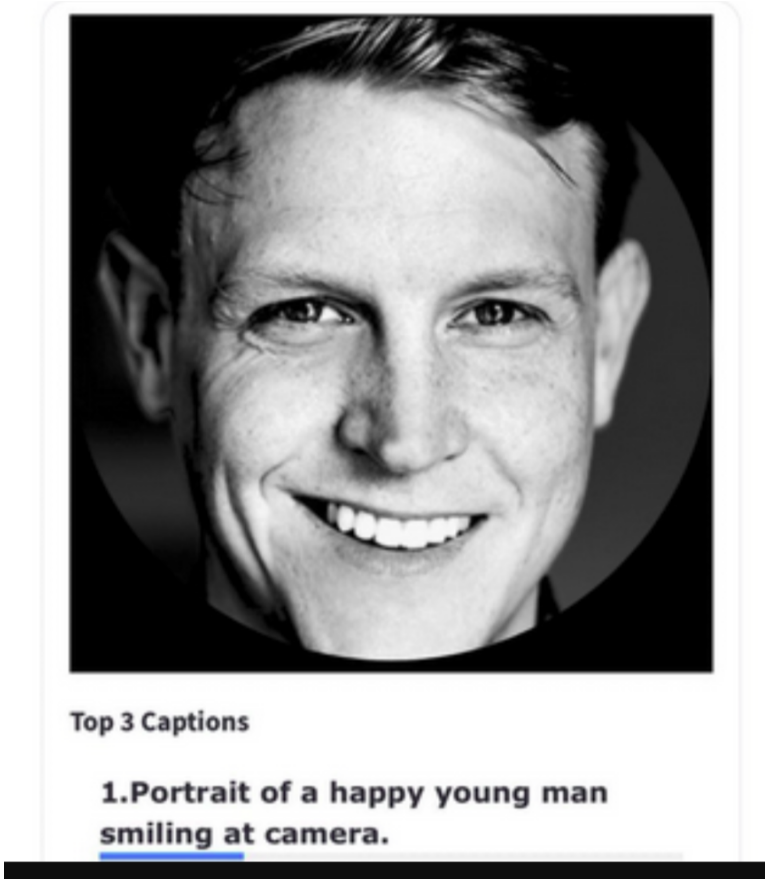
[2](Goodfellow, Shlens, and Szegedy, 2014)

Adversarial?



[3](Shafer and Vovk, 2008)

Adversarial?



What if applied to Job recommendation? Justice? Health?

Question: [3] If you predict a label y of a new object with $\hat{y}$, how `confident` are you that "$y = \hat{y}$"?

[3] (Shafer and Vovk, 2008)

Observations: $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ iid $\sim \mathbb{P}$

New input data: $(x_{n+1}, \not{y}_{\not{n}\not{+}\not{1}})$

Goal: build a set that contains y_{n+1} with high probability.

Observations: $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ iid $\sim \mathbb{P}$

New input data: $(x_{n+1}, \cancel{y_{n+1}})$

Goal: build a set that contains y_{n+1} with high probability.

Desirable property

- The coverage guarantee must hold for any sample size n.
- The set must be as small as possible.
- No assumptions on the distribution $\mathbb{P}$.

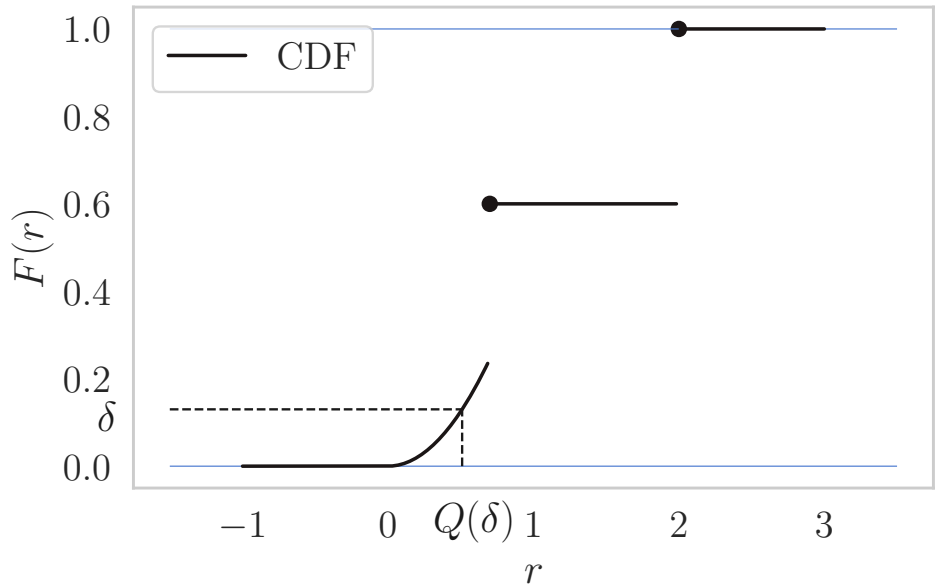
- Evaluate[4] the prediction error of your model $\mu(X) \approx Y$

$$E := |Y - \mu(X)| \leq r \quad \Longleftrightarrow \quad Y \in [\mu(X) \pm r]$$

[4] CDF $F(r) := \mathbb{P}(E \leq r)$ and Quantile $Q(\delta) = \inf\{r : \delta \leq F(r)\}$

- Evaluate[4] the prediction error of your model $\mu(X) \approx Y$

$$E := |Y - \mu(X)| \leq r \quad \Longleftrightarrow \quad Y \in [\mu(X) \pm r]$$



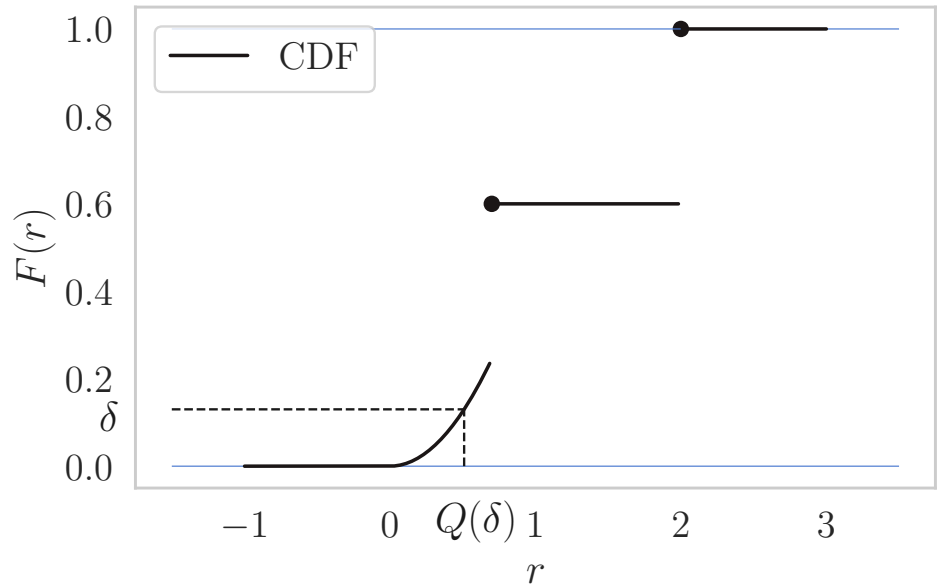
- Coverage

$$\delta \leq F(Q(\delta)) = \mathbb{P}(|Y - \mu(X)| \leq Q(\delta))$$

[4] CDF $F(r) := \mathbb{P}(E \leq r)$ and Quantile $Q(\delta) = \inf\{r : \delta \leq F(r)\}$

- Evaluate[4] the prediction error of your model $\mu(X) \approx Y$

$$E := |Y - \mu(X)| \leq r \quad \Longleftrightarrow \quad Y \in [\mu(X) \pm r]$$



- Coverage

$$\delta \leq F(Q(\delta)) = \mathbb{P}(|Y - \mu(X)| \leq Q(\delta))$$

- $100(1-\alpha)\%$ Confidence Set

$$\Gamma(X) = \{z : |z - \mu(X)| \leq Q(1-\alpha)\}$$

[4] CDF $F(r) := \mathbb{P}(E \leq r)$ and Quantile $Q(\delta) = \inf\{r : \delta \leq F(r)\}$

In practice, the **distribution** F (and so Q) **is unknown** and we only have access to a sample of the data $\{E_1, \ldots, E_n, E_{n+1}\}$

[5] Empirical estimation of the CDF and Quantile
$F_{n+1}(r) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}_{E_i \leq r}$ and $Q_{n+1}(\delta) = \inf\{r : \delta \leq F_{n+1}(r)\}$

In practice, the **distribution** F (and so Q) **is unknown** and we only have access to a sample of the data $\{E_1, \ldots, E_n, E_{n+1}\}$

- where we fit and evaluate scores on the training dataset

$$E_i = |y_i - \mu(x_i)|, \qquad \forall i \in \{1, \ldots, n, n+1\}$$

- Coverage[5]

$$\delta \leq \mathbb{P}\left(|y_{n+1} - \mu_{y_{n+1}}(x_{n+1})| \leq Q_{n+1}(\delta, y_{n+1}) \right)$$

[5] Empirical estimation of the CDF and Quantile
$F_{n+1}(r) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}_{E_i \leq r}$ and $Q_{n+1}(\delta) = \inf\{r : \delta \leq F_{n+1}(r)\}$

In practice, the **distribution** F (and so Q) **is unknown** and we only have access to a sample of the data $\{E_1, \ldots, E_n, E_{n+1}\}$

- where we fit and evaluate scores on the training dataset

$$E_i = |y_i - \mu(x_i)|, \qquad \forall i \in \{1, \ldots, n, n+1\}$$

- Coverage[5]

$$\delta \leq \mathbb{P}\left(|y_{n+1} - \mu_{y_{n+1}}(x_{n+1})| \leq Q_{n+1}(\delta, y_{n+1}) \right)$$

- $100(1-\alpha)\%$ **Conformal Prediction Set**

$$\Gamma(x_{n+1}) = \{z : |z - \mu_z(x_{n+1})| \leq Q_{n+1}(1-\alpha, z)\}$$

[5] Empirical estimation of the CDF and Quantile $F_{n+1}(r) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}_{E_i \leq r}$ and $Q_{n+1}(\delta) = \inf\{r : \delta \leq F_{n+1}(r)\}$

Proof [6] [7] [8] [9]

$$\delta \leq F_{n+1}(Q_{n+1}(\delta)) = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbf{1}_{E_i \leq Q_{n+1}(\delta)}$$

(by taking expectation on both side)

$$\delta \leq \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{P}(E_i \leq Q_{n+1}(\delta)) \underset{\textbf{iid}}{=} \mathbb{P}(E_{n+1} \leq Q_{n+1}(\delta))$$

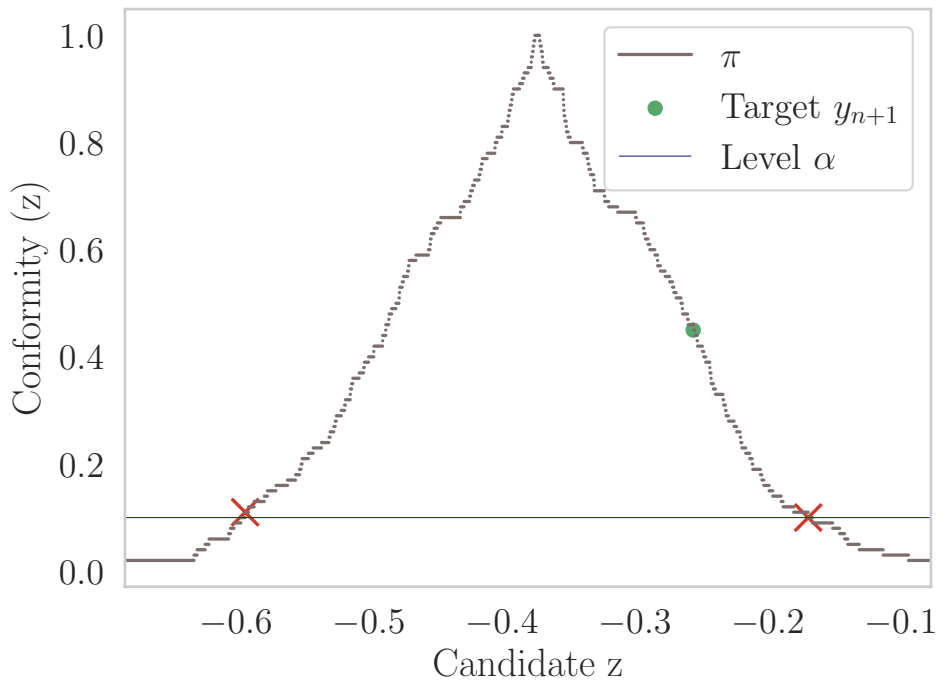
[6] (V. Vovk, A. Gammerman, and G. Shafer, 2005)
[7] (G. Shafer and V. Vovk, 2008)
[8] (J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, and L. Wasserman, 2018)
[9] (Romano, Patterson, and Candes, 2019)

Visualization/interpretation

$$\Gamma(x_{n+1}) = \{z : |z - \mu_z(x_{n+1})| \leq Q_{n+1}(1-\alpha, z)\} = \{z : \ \pi(z) \geq \alpha\}$$

$$\pi(z) = 1 - F_{n+1}(|z - \mu_z(x_{n+1})|) = \quad \texttt{conformity function}$$

- $\pi(\cdot)$ acts as `pi-value` to test $H_0 : y_{n+1} = z$ vs $H_1 : y_{n+1} \neq z$



- Statistical guarantee: $\mathbb{P}(y_{n+1} \in \Gamma(x_{n+1})) \geq 1 - \alpha$

Computational Limitations

$$\Gamma(x_{n+1}) = \{z : |z - \mu_z(x_{n+1})| \leq Q_{n+1}(1-\alpha, z)\}$$

Issue: we need to refit the prediction model $\mu_z(\cdot)$ **for all** z

- Ok if y_{n+1} has a *finite small* number of possibilities
- Ok for model with simple explicit formula [10] [11]

Otherwise, it is an *open problem.*

[10] (I. Nouretdinov, T. Melluish, and V. Vovk, 2001)
[11] (J. Lei, 2019)

Full CP Set

$$\left\{ z : \ |z - \mu_z(x_{n+1})| \leq Q_{n+1}(1-\alpha, z) \right\}$$

[12](Papadopoulos, Proedrou, Vovk, and Gammerman, 2002)

Full CP Set

$$\left\{ z : \ |z - \mu_z(x_{n+1})| \leq Q_{n+1}(1-\alpha, z) \right\}$$

split[12] CP Set

$$\left\{ z : |z - \mu_{\mathcal{D}_1}(x_{n+1})| \leq Q_{\mathcal{D}_2}(1-\alpha) \right\}$$

- $\mu_{\mathcal{D}_1}$ is fitted on $\mathcal{D}_1$
- $Q_{\mathcal{D}_2}(1-\alpha)$ is the quantile of errors $|y - \mu_{\mathcal{D}_1}(x)|$ for $(x, y) \in \mathcal{D}_2$

[12](Papadopoulos, Proedrou, Vovk, and Gammerman, 2002)

Full CP Set

$$\left\{ z : \ |z - \mu_z(x_{n+1})| \leq Q_{n+1}(1-\alpha, z) \right\}$$

split[12] CP Set

$$\left\{ z : |z - \mu_{\mathcal{D}_1}(x_{n+1})| \leq Q_{\mathcal{D}_2}(1-\alpha) \right\}$$

- $\mu_{\mathcal{D}_1}$ is fitted on $\mathcal{D}_1$
- $Q_{\mathcal{D}_2}(1-\alpha)$ is the quantile of errors $|y - \mu_{\mathcal{D}_1}(x)|$ for $(x, y) \in \mathcal{D}_2$

$$(\sim \_ \sim)$$

- (often) results in **wider** interval
- Introduces an additional randomness
- Suitable for `pre-trained` model?

[12] (Papadopoulos, Proedrou, Vovk, and Gammerman, 2002)

Cross-Conformal Prediction

Use a K-fold fit and calibration [13] [14] [15] [16]

$$(\sim \_ \sim)$$

- Instead of $(1-\alpha)$, the coverage is

$$\mathbb{P}(y_{n+1} \text{ in CP set}) \geq 1 - 2\alpha$$

  obtained with $K = n$ (`Jackknife`) and so needs n model fit!
- It gets worse when $K < n$
- Not trivial to maintain validity with aggregated CP

[13] (Carlsson, Eklund, and Norinder, 2014)
[14] (Vovk, 2015)
[15] (Linusson, Norinder, Boström, Johansson and Löfström, 2017)
[16] (Barber, Candes, Ramdas and, Tibshirani, 2021)

Main contribution

Compute a conformal prediction set

- **Data efficiency**: **no** data splitting
- **Statistical coverage**: $1 - \alpha$
- **Computational efficiency**: **one** model fit

$$\texttt{Dataset} = \{(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1})\}$$

We might want to bet that the estimator $\mu_z(\cdot)$ does not change much when a single input z changes[17].

[17](O. Bousquet and A. Elisseeff, 2002)

$$\texttt{Dataset} = \{(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, \not{y}_{n+1})\}$$

We might want to bet that the estimator $\mu_z(\cdot)$ does not change much when a single input z changes[17].

Definition (Algorithmic Stability)

A prediction function $\mu_\cdot$ is stable if for any observation (x_i, y_i)

$$|S(q, \mu_z(x_i)) - S(q, \mu_{\hat{z}}(x_i))| \leq \tau_i \qquad \forall z, \hat{z}, q$$

[17] (O. Bousquet and A. Elisseeff, 2002)

$$\texttt{Dataset} = \{(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1})\}$$

We might want to bet that the estimator $\mu_z(\cdot)$ does not change much when a single input z changes[17].

Definition (Algorithmic Stability)

A prediction function $\mu_\cdot$ is stable if for any observation (x_i, y_i)

$$|S(q, \mu_z(x_i)) - S(q, \mu_{\hat{z}}(x_i))| \leq \tau_i \qquad \forall z, \hat{z}, q$$

- **Upper** and **Lower** bounds on the prediction error

[17] (O. Bousquet and A. Elisseeff, 2002)

$$\texttt{Dataset} = \{(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, \not{y}_{n+1})\}$$

We might want to bet that the estimator $\mu_z(\cdot)$ does not change much when a single input z changes[17].

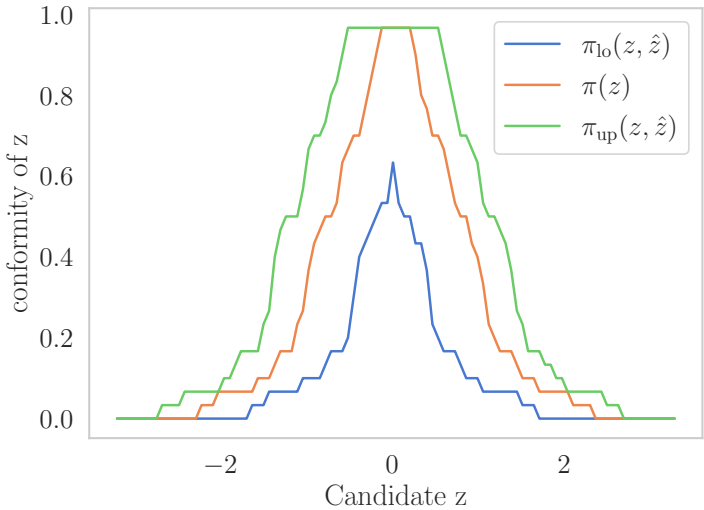
Definition (Algorithmic Stability)

A prediction function $\mu_\cdot$ is stable if for any observation (x_i, y_i)

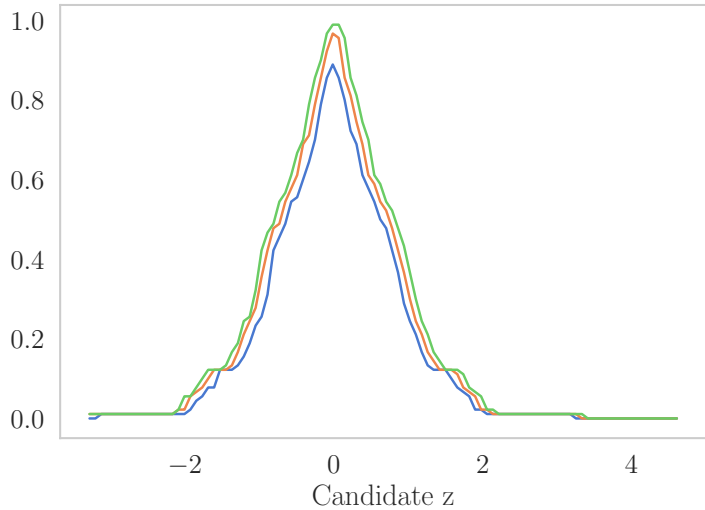
$$|S(q, \mu_z(x_i)) - S(q, \mu_{\hat{z}}(x_i))| \leq \tau_i \qquad \forall z, \hat{z}, q$$

- **Upper** and **Lower** bounds on the prediction error
- A **single** model fit $\mu_{\hat{z}}$

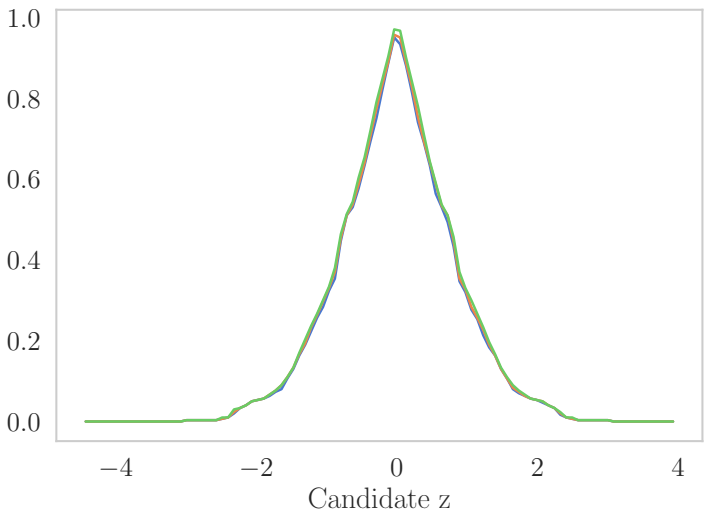
[17] (O. Bousquet and A. Elisseeff, 2002)



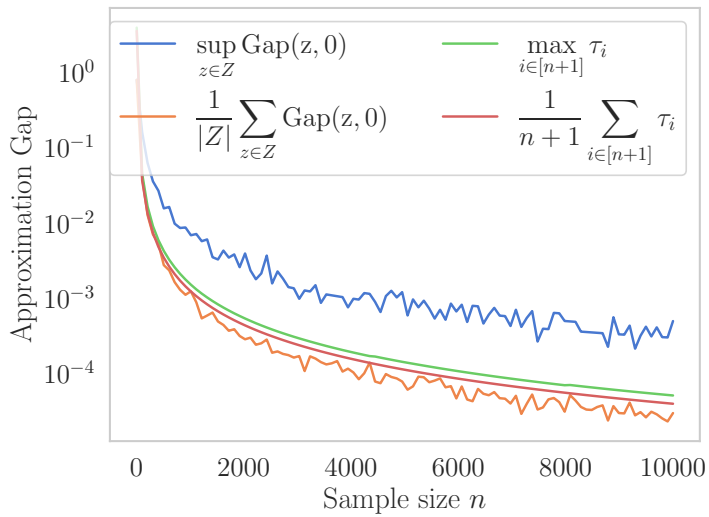
(e) $n = 30$



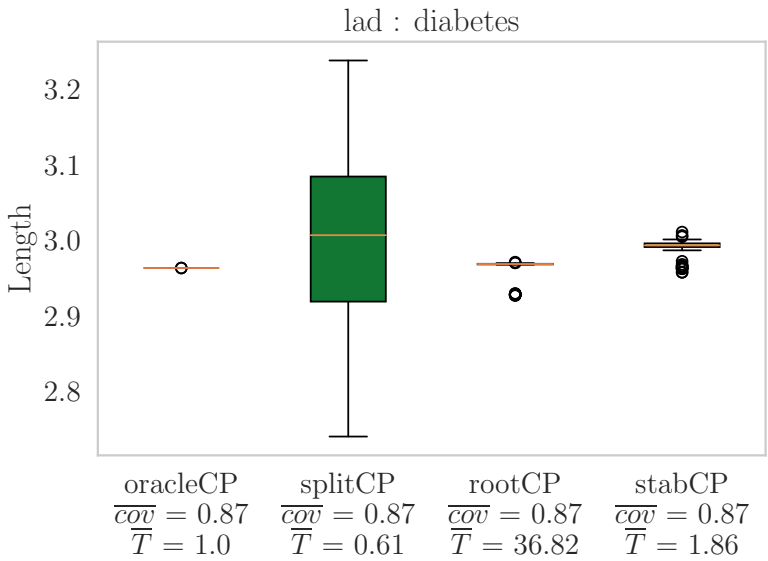
(f) $n = 90$



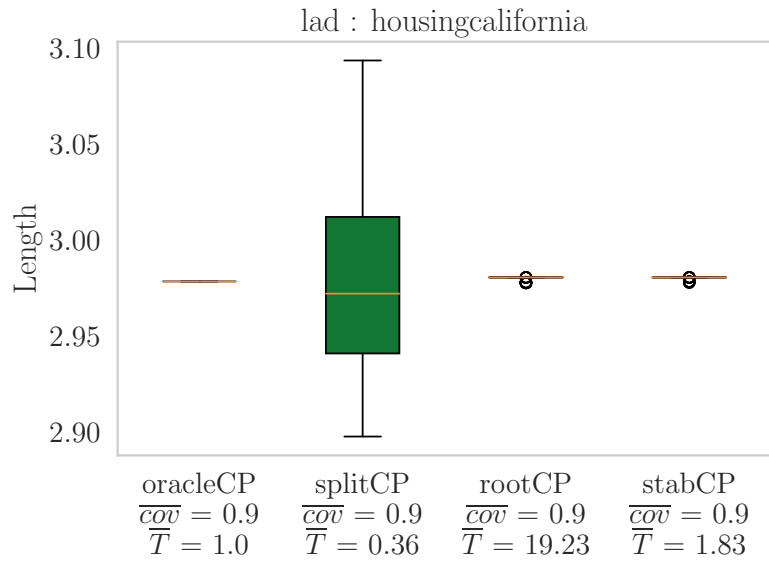
(g) $n = 300$



(h) Approximation error



(i) `Diabetes` $(442, 10)$

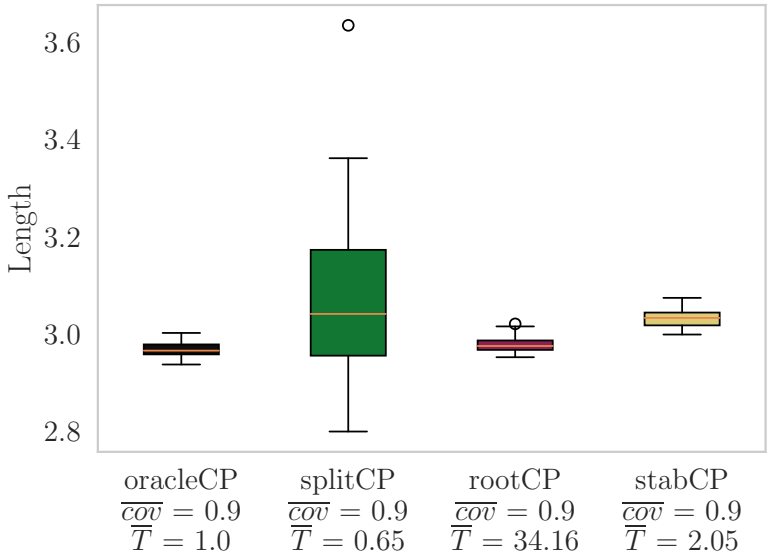


(j) `Housingcalifornia` $(20640, 8)$

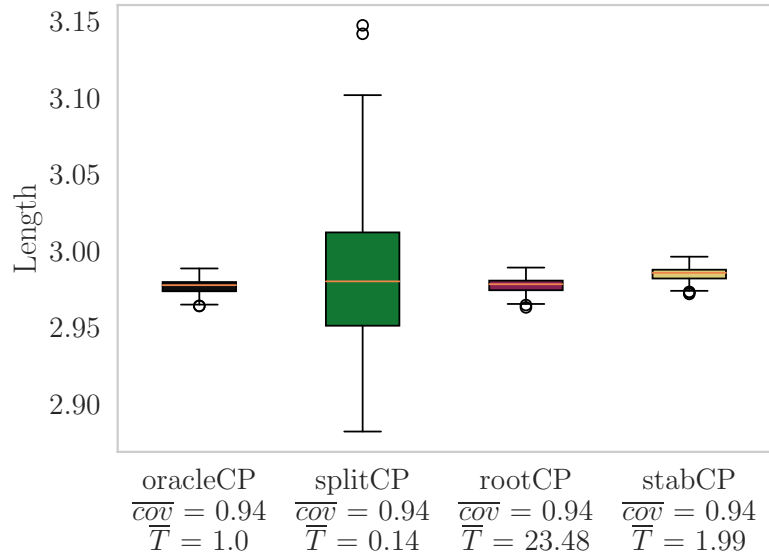
$$\min_{\beta} \frac{1}{n+1} \|Y - X\beta\|_1 + \lambda \|\beta\|_2^2$$

Stability bounds

$$\tau_i = \frac{2 \|x_i\|}{\lambda(n+1)}$$



(k) `Diabetes` $(442, 10)$



(l) `Housingcalifornia` $(20640, 8)$

MLP regression models with a ridge regularization. The parameter of the model is obtained after $T = n/10$ iterations of stochastic gradient descent. We use a rough stability bound estimate of

$$\tau_i = \frac{T \|x_i\|}{n+1}$$

Current trend in ML:

- we focus a lot on optimizing the predictive **performance**
- On a dual side, how should we optimize **confidence**?

- **Codes:** `https://github.com/EugeneNdiaye/stable_conformal_prediction`
- **Papers:** `https://eugenendiaye.github.io`