Feature Conformal Prediction

Akash Choudhuri

11/3/2025

Predictive Inference with Feature Conformal Prediction

Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, Yang Yuan

ICLR 2023

Background

- Need for uncertainty Quantification- Prediction with confidence
- Conformal prediction:
  - Distribution Free
  - Model Agonistic
- Prior works used Conformal Prediction on the output space
- Claim: Using Conformal prediction on feature space, output band size decreases

Setting

- Data Pair: $(X, Y) \sim \mathcal{P}$
- Confidence Band: $\mathcal{C}_{1-\alpha}(X)$ $\mathbb{P}\left(Y \in \mathcal{C}_{1-\alpha}(X)\right) \geq 1 - \alpha.$
- Overall Dataset: $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$

$\mathcal{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{tr}}}$ $\mathcal{D}_{\text{ca}} = \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{ca}}}$

- Testing Point: (X', Y')

Algorithm 1 Conformal Prediction

Require: Desired confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, test point X', non-conformity score function $s(\cdot)$

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tr}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{tr}}}$ and a calibration fold $\mathcal{D}_{\text{ca}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{ca}}}$;
2: Train a base machine learning model $\hat{\mu}(\cdot)$ with $\mathcal{D}_{\text{tr}}$ to estimate the response Y_i;
3: For each $i \in \mathcal{I}_{\text{ca}}$, calculate its non-conformity score $V_i = s(X_i, Y_i, \hat{\mu})$;
4: Calculate the $(1 - \alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{ca}}|+1} \sum_{i \in \mathcal{I}_{ca}} \delta_{V_i} + \delta_{\infty}$.

Ensure: $\mathcal{C}_{1-\alpha}(X') = \{Y : s(X', Y, \hat{\mu}) \leq Q_{1-\alpha}\}$.

General Conditions of Conformal Prediction

Assumption 1 (exchangeability). *Assume that the calibration data* $(X_i, Y_i), i \in \mathcal{I}_{ca}$ *and the test point* (X', Y') *are exchangeable. Formally, define* $Z_i, i = 1, \ldots, |\mathcal{I}_{ca} + 1|$, *as the above data pair, then* Z_i *are exchangeable if arbitrary permutation leads to the same distribution, i.e.,*

$$(Z_1, \ldots, Z_{|\mathcal{I}_{ca}|+1}) \stackrel{d}{=} (Z_{\pi(1)}, \ldots, Z_{\pi(|\mathcal{I}_{ca}|+1)}), \tag{2}$$

with arbitrary permutation π *over* $\{1, \cdots, |\mathcal{I}_{ca} + 1|\}$.

Theorem 2 (theoretical guarantee for conformal prediction (Law, 2006; Lei et al., 2018; Tibshirani et al., 2019)). *Under Assumption 1, the confidence band* $\mathcal{C}_{1-\alpha}(X')$ *returned by Algorithm 1 satisfies*

$$\mathbb{P}(Y' \in \mathcal{C}_{1-\alpha}(X')) \geq 1 - \alpha.$$

Method: Setup

- Assume the mean estimator to be a composition of 2 functions

Definition 3 (Surrogate feature)**.** *Consider a trained neural network $\hat{\mu} = \hat{g} \circ \hat{f}$ where $\circ$ denotes the composition operator. For a sample (X, Y), we define $\hat{v} = \hat{f}(X)$ to be the trained feature. Besides, we define the surrogate feature to be any feature v such that $\hat{g}(v) = Y$.*

- Proposed NC Score: $s(X, Y, \hat{g} \circ \hat{f}) = \inf_{v \in \{v : \hat{g}(v) = Y\}} \|v - \hat{f}(X)\|.$ Intractable. Why?
- Take Upper Bound of NC Score:

Algorithm 2 Non-conformity Score

Require: Data point (X, Y), trained predictor $\hat{g} \circ \hat{f}(\cdot)$, step size η, number of steps M;

1: $u \leftarrow \hat{f}(X)$;
2: $m \leftarrow 0$;
3: **while** $m < M$ **do**
4: $\quad u \leftarrow u - \eta \frac{\partial \|\hat{g}(u) - Y\|^2}{\partial u}$;
5: $\quad m \leftarrow m + 1$;
6: **end while**

Ensure: $s(X, Y, \hat{g} \circ \hat{f}) = \|u - \hat{f}(X)\|$.

Band Estimation and Overall Algorithm

- NC score-> Feature Space. How to transfer the score to the output space? $\{\hat{g}(v) : \|v - \hat{v}\| \leq Q_{1-\alpha}\}$,
- Band Estimation (LIRPA):
  - Treat as perturbation analysis problem
  - Treat v as perturbation of trained feature $\hat{v}$
  - Analyze the output bounds of $\hat{g}(.)$

Algorithm 3 Feature Conformal Prediction

Require: Level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, test point X';

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tr}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{tr}}}$ together with a calibration fold $\mathcal{D}_{\text{ca}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{ca}}}$;
2: Train a base machine learning model $\hat{g} \circ \hat{f}(\cdot)$ using $\mathcal{D}_{\text{tr}}$ to estimate the response Y_i;
3: For each $i \in \mathcal{I}_{\text{ca}}$, calculate the non-conformity score V_i based on Algorithm 2;
4: Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{ca}}|+1} \sum_{i \in \mathcal{I}_{ca}} \delta_{V_i} + \delta_{\infty}$;
5: Apply Band Estimation on test data feature $\hat{f}(X')$ with perturbation $Q_{1-\alpha}$ and prediction head $\hat{g}$, which returns $\mathcal{C}^{\text{fcp}}_{1-\alpha}(X)$;

Ensure: $\mathcal{C}^{\text{fcp}}_{1-\alpha}(X)$.

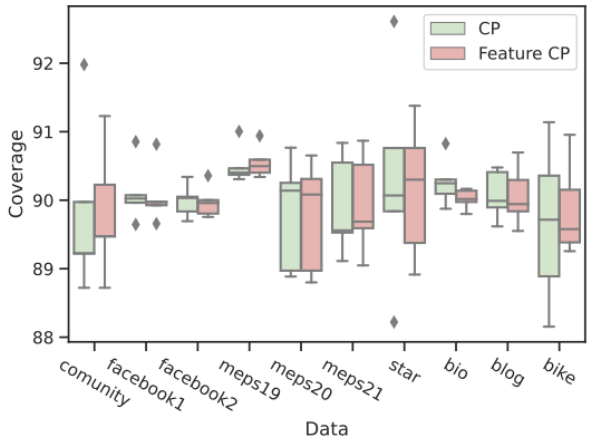
Theoretical Guarantee

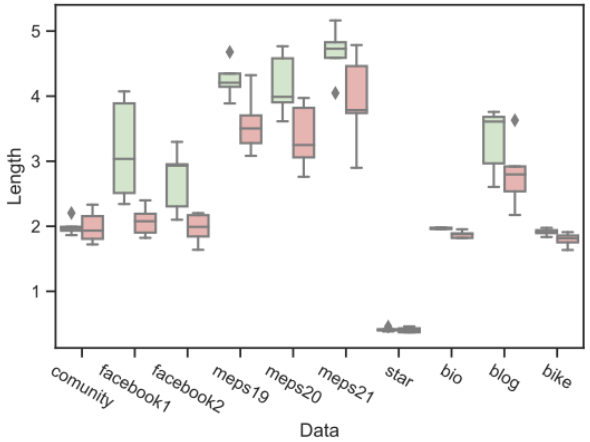
Theorem 4 (Informal Theorem on the Efficiency of Feature CP)**.** *Under mild assumptions, if the following cubic conditions hold:*

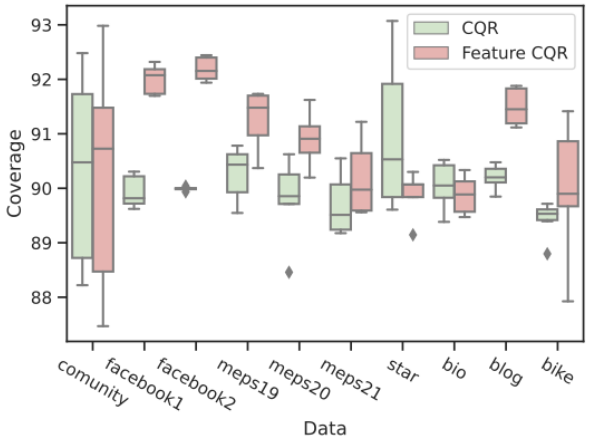
1. ***Length Preservation.*** *Feature CP does not cost much loss in feature space.*
2. ***Expansion.*** *The Band Estimation operator expands the differences between individual length and their quantiles.*
3. ***Quantile Stability.*** *The band length is stable in both the feature space and the output space for a given calibration set.*

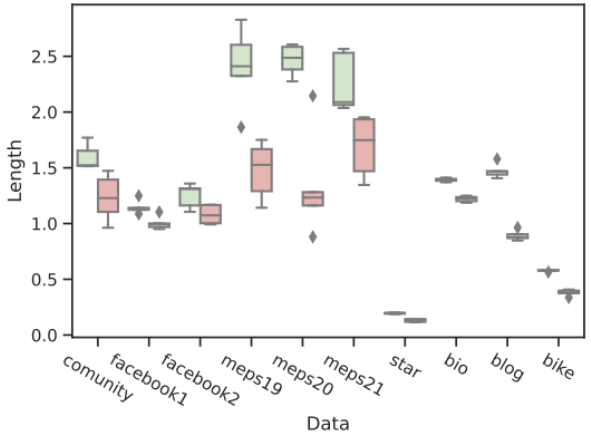
then Feature CP outperforms vanilla CP in terms of average band length.

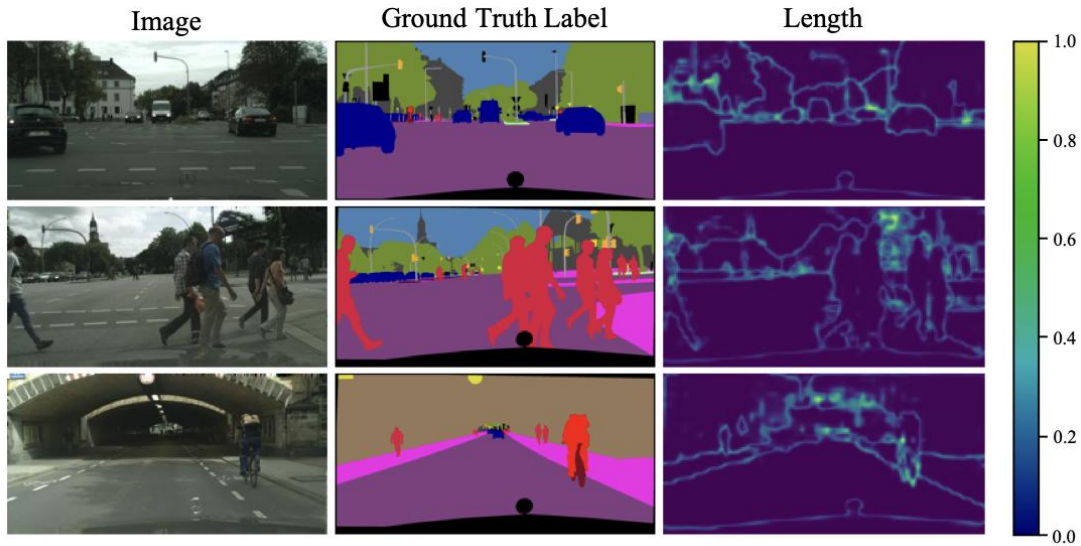
Experiments











Accelerating Feature Conformal Prediction via Taylor Approximation

Zihao Tang Boyuan Wang Chuan Wen Jiaye Teng

NeurIPS 2025

Background

- Band Estimation:
  - Time-consuming
  - Only returns estimated bands on the output space

$$s_{\mathrm{f}}(X, Y, g \circ h) = \inf_{v \in \{v : g(v) = Y\}} \|v - \hat{v}\|. \quad = \quad s_{\mathrm{ff}}(X, Y, g \circ h) = |Y - f(X)| / \|\nabla g(\hat{v})\|, \quad \nabla g(\hat{v}) = \frac{\mathrm{d}g \circ h(X)}{\mathrm{d}h(X)}$$

$$\{g(v) : \|v - \hat{v}\| \leq Q_{1-\alpha}\}. \quad = \quad C^{\mathrm{ffcp}}_{1-\alpha}(X) = \left[g(\hat{v}) - \|\nabla g(\hat{v})\| Q_{1-\alpha}, g(\hat{v}) + \|\nabla g(\hat{v})\| Q_{1-\alpha}\right].$$

Overall Algorithm

Algorithm 2 Fast Feature Conformal Prediction

Input: Confidence level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, tesing point X'

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tra}} \triangleq \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{tra}}}$ and a calibration fold $\mathcal{D}_{\text{cal}} \triangleq \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\text{cal}}}$;

2: Train a base neural network with training fold $f(\cdot) = g \circ h(\cdot)$ with training fold $\mathcal{D}_{\text{tra}}$;

3: For each $i \in \mathcal{I}_{\text{cal}}$, calculate the non-conformity score $\tilde{R}_i = |Y_i - f(X_i)|/\|\nabla g(\hat{v}_i)\|$, where $\nabla g(\hat{v}_i)$ denotes the gradient of $g(\cdot)$ on the feature $\hat{v}_i \triangleq h(X_i)$, namely $\nabla g(\hat{v}_i) = \frac{\mathrm{d}g \circ h(X_i)}{\mathrm{d}h(X_i)}$;

4: Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{cal}}|+1} \sum_{i \in \mathcal{I}_{\text{cal}}} \delta_{\tilde{R}_i} + \delta_{\infty}$;

Output: $C^{\text{ffcp}}_{1-\alpha}(X') = [f(X') - \|\nabla g(\hat{v}')\| Q_{1-\alpha}, f(X') + \|\nabla g(\hat{v}')\| Q_{1-\alpha}]$, where $\hat{v}' = h(X')$.

Theorem 5 (Band Length). *Under mild assumptions, if the following square conditions hold:*

1. ***Expansion.*** *The feature space expands the differences between individual length and their quantiles.*

2. ***Quantile Stability.*** *Given a calibration set $\mathcal{D}_{cal}$, the quantile of the band length is stable in both feature space and output space.*

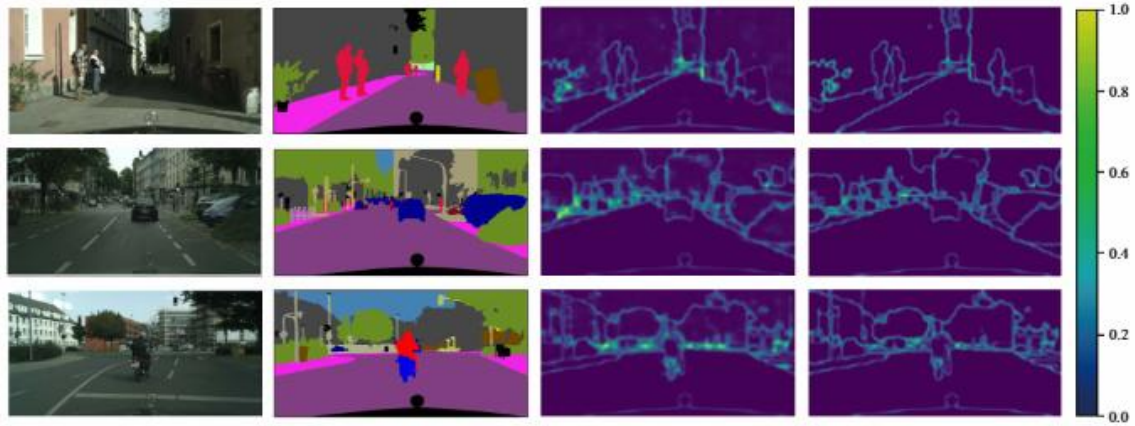
Then FFCP provably outperforms vanilla CP in terms of average band length.

Experiments

Dataset	Vanilla CP	FCP	FFCP	Faster
synthetic	0.0088±0.0003	3.8939±0.3725	0.0902±0.0056	43x
com	0.0047±0.0010	4.9804±0.8588	0.0844±0.0187	59x
fb1	0.0245±0.0059	5.9822±0.9871	0.1940±0.0564	31x
fb2	0.0414±0.0070	9.3534±0.0927	0.2510±0.0058	37x
meps19	0.0106±0.0010	3.3237±0.0431	0.0755±0.0037	44x
meps20	0.0152±0.0016	5.4003±0.3945	0.0948±0.0077	57x
meps21	0.0137±0.0008	4.1657±0.0670	0.0854±0.0146	49x
star	0.0030±0.0006	3.5842±0.3722	0.0332±0.0066	108x
bio	0.0291±0.0053	7.5417±1.1028	0.2042±0.0344	37x
blog	0.0340±0.0024	8.0913±1.2072	0.2239±0.0261	36x
bike	0.0072±0.0007	3.5806±0.0285	0.0534±0.0021	67x

Method	Vanilla CP		FCP		FFCP	
Dataset	Coverage	Length	Coverage	Length	Coverage	Length
synthetic	90.080±0.951	0.176±0.015	89.930±0.956	**0.081**±0.041	90.080±0.951	0.176±0.015
com	89.875±0.985	1.974±0.071	89.724±1.087	1.939±1.408	90.226±2.179	**1.838**±0.180
fb1	90.254±0.170	2.004±0.191	90.198±0.207	2.010±0.182	90.168±0.220	**1.472**±0.232
fb2	89.933±0.206	2.016±0.218	89.966±0.130	**1.371**±0.370	89.868±0.062	1.425±0.109
meps19	90.567±0.311	3.982±0.614	90.605±0.340	3.493±2.734	90.352±0.469	**3.134**±0.309
meps20	89.923±0.715	4.184±0.316	89.929±0.770	**2.730**±0.962	89.615±0.661	3.268±0.283
meps21	90.019±0.341	3.732±0.555	90.038±0.303	3.393±1.313	89.745±0.344	**3.146**±0.506
star	90.393±1.494	0.208±0.004	90.300±1.362	**0.174**±0.038	90.393±1.494	0.208±0.004
bio	89.875±0.488	1.661±0.019	89.930±0.501	**1.412**±0.265	89.875±0.488	1.661±0.019
blog	90.176±0.241	3.524±0.850	90.151±0.405	2.795±1.385	90.059±0.101	**2.741**±0.517
bike	89.871±0.568	0.703±0.016	89.394±0.633	2.147±0.249	89.624±0.688	**0.635**±0.030

Layer	layer1		layer2		layer3		layer4	
Dataset	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
synthetic	89.810±0.784	0.184±0.018	90.050±0.534	0.184±0.017	89.960±0.910	**0.182**±0.023	90.220±0.983	0.189±0.033
com	90.476±1.889	1.878±0.224	90.226±2.179	**1.838**±0.180	89.674±1.465	1.853±0.136	89.825±0.646	2.037±0.188
fb1	90.112±0.199	3.540±0.327	90.212±0.357	2.860±0.327	90.083±0.216	1.597±0.052	90.168±0.220	**1.472**±0.232
fb2	89.953 ±0.250	3.530±0.384	89.897±0.235	3.048±0.510	89.956±0.159	2.077±0.517	89.868±0.062	**1.425**±0.109
meps19	90.155±0.643	3.251±0.396	90.352±0.469	**3.134**±0.309	90.440±0.183	3.184±0.482	90.586±0.246	3.795±0.640
meps20	89.934±0.520	4.302±1.377	89.889±0.621	3.573±0.488	89.615±0.661	**3.268**±0.283	89.82±0.689	3.817±0.308
meps21	89.496±0.262	3.443±0.487	89.623±0.275	3.218±0.239	89.745±0.344	**3.146**±0.506	90.026±0.301	3.452±0.711
star	90.901±1.732	0.221±0.002	90.993±1.807	0.217±0.003	91.039±1.442	0.210±0.004	90.300±1.248	**0.209**±0.004
bio	89.937±0.391	2.292±0.077	90.022±0.375	2.042±0.067	89.991±0.594	2.080±0.063	90.127±0.476	**1.822**±0.025
blog	89.968±0.420	4.772±0.614	89.918±0.319	3.404±0.598	90.059±0.101	**2.741**±0.517	90.017±0.197	3.058±0.873
bike	89.917±0.791	1.701±0.254	89.568±0.476	1.138±0.114	89.495±0.579	0.794±0.068	89.624±0.688	**0.635**±0.030



(a) Image (b) Ground Truth Label (c) FCP Length (d) FFCP Length

Feature CQR

Algorithm 4 Feature Conformalized Quantile Regression (Feature CQR)

Require: Level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, test point X';

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{\text{tr}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{tr}}}$ together with a calibration fold $\mathcal{D}_{\text{ca}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\text{ca}}}$;

2: Train a base machine learning model $\hat{g}^{\text{lo}} \circ \hat{f}^{\text{lo}}(\cdot)$ and $\hat{g}^{\text{hi}} \circ \hat{f}^{\text{hi}}(\cdot)$ using $\mathcal{D}_{\text{tr}}$ to estimate the quantile of response Y_i, which returns $[\hat{Y}_i^{\text{lo}}, \hat{Y}_i^{\text{hi}}]$;

3: For each $i \in \mathcal{I}_{\text{ca}}$, calculate the index $c_i^{\text{lo}} = \mathbb{I}(\hat{Y}_i^{\text{lo}} \leq Y)$ and $c_i^{\text{hi}} = \mathbb{I}(\hat{Y}_i^{\text{hi}} \geq Y)$;

4: For each $i \in \mathcal{I}_{\text{ca}}$, calculate the non-conformity score $V_i^{\text{lo}} = \tilde{V}_i^{\text{lo}} c_i^{\text{lo}}$ where $\tilde{V}_i^{\text{lo}}$ is derived on the lower bound function with Algorithm 2;

5: Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}^{\text{lo}}$ of the distribution $\frac{1}{|\mathcal{I}_{\text{ca}}|+1} \sum_{i \in \mathcal{I}_{ca}} \delta_{V_i^{\text{lo}}} + \delta_{\infty}$;

6: Apply Band Estimation on test data feature $\hat{f}^{\text{lo}}(X')$ with perturbation $Q_{1-\alpha}^{\text{lo}}$ and prediction head $\hat{g}^{\text{lo}}$, which returns $[C_0^{\text{lo}}, C_1^{\text{lo}}]$;

7: Apply STEP 4-6 similarly with higher quantile, which returns $[C_0^{\text{hi}}, C_1^{\text{hi}}]$;

8: Derive $C_{1-\alpha}^{\text{fcqr}}(X)$ based on Equation (12);

Ensure: $C_{1-\alpha}^{\text{fcqr}}(X)$.