DistMatch: Adaptive Binning via Distribution Matching for Robust Sequential Conformal Prediction

Enver Menadjiev[1*], Jihyeon Seong[1*], Jisu Yeo[1], and Jaesik Choi[1,2†]

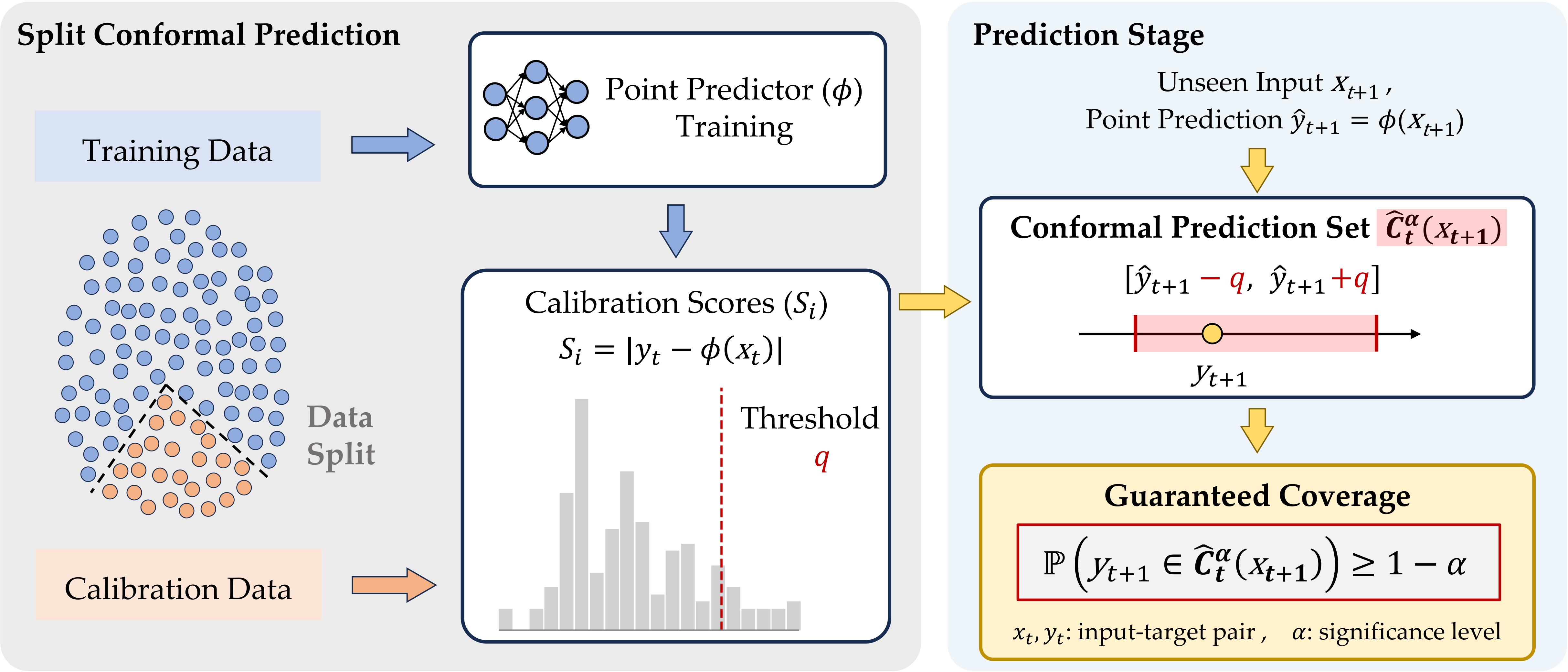
[1]Korea Advanced Institute of Science and Technology

[2]INEEJI

** Equal Contribution*
† Corresponding Author

Conformal Prediction

Conformal Prediction[1] produces a range guaranteed to contain the true outcome with a certain probability.

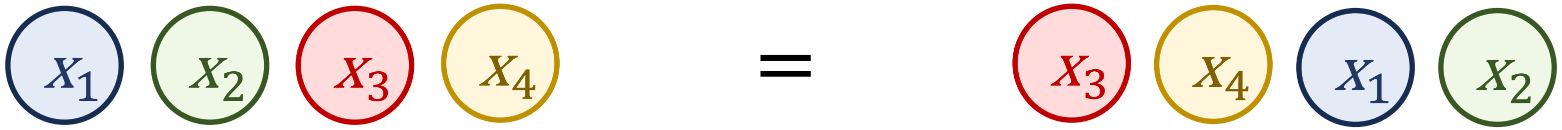


[1] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic Learning in a Random World. Springer US.

Exchangeability

Q: How can we guarantee **coverage** on the unseen set?

A: **Exchangeability.**



$P(x_1, x_2, x_3, x_4)$ (equal dist.) $P(x_3, x_4, x_1, x_2)$

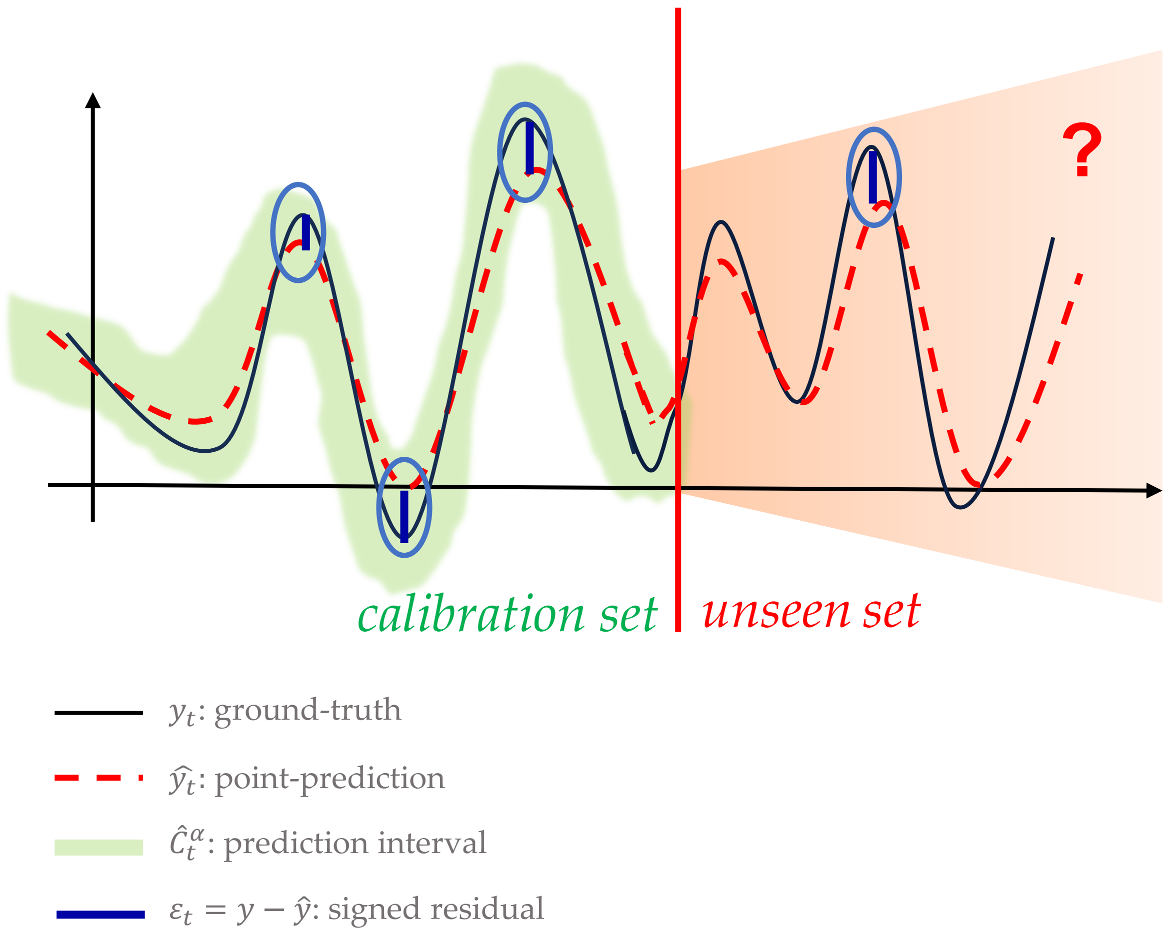
If the joint distribution of calibration samples (1, …, T−1) and unseen samples (T, …, T+n) is **invariant under finite permutations**, valid coverage on the unseen set is guaranteed.

$$P(x_1, x_2, \dots, x_T) = P\left(x_{\pi(1)}, x_{\pi(2)}, \dots, x_{\pi(T)}\right)$$

- dist.: distribution
- π: permutation of indices

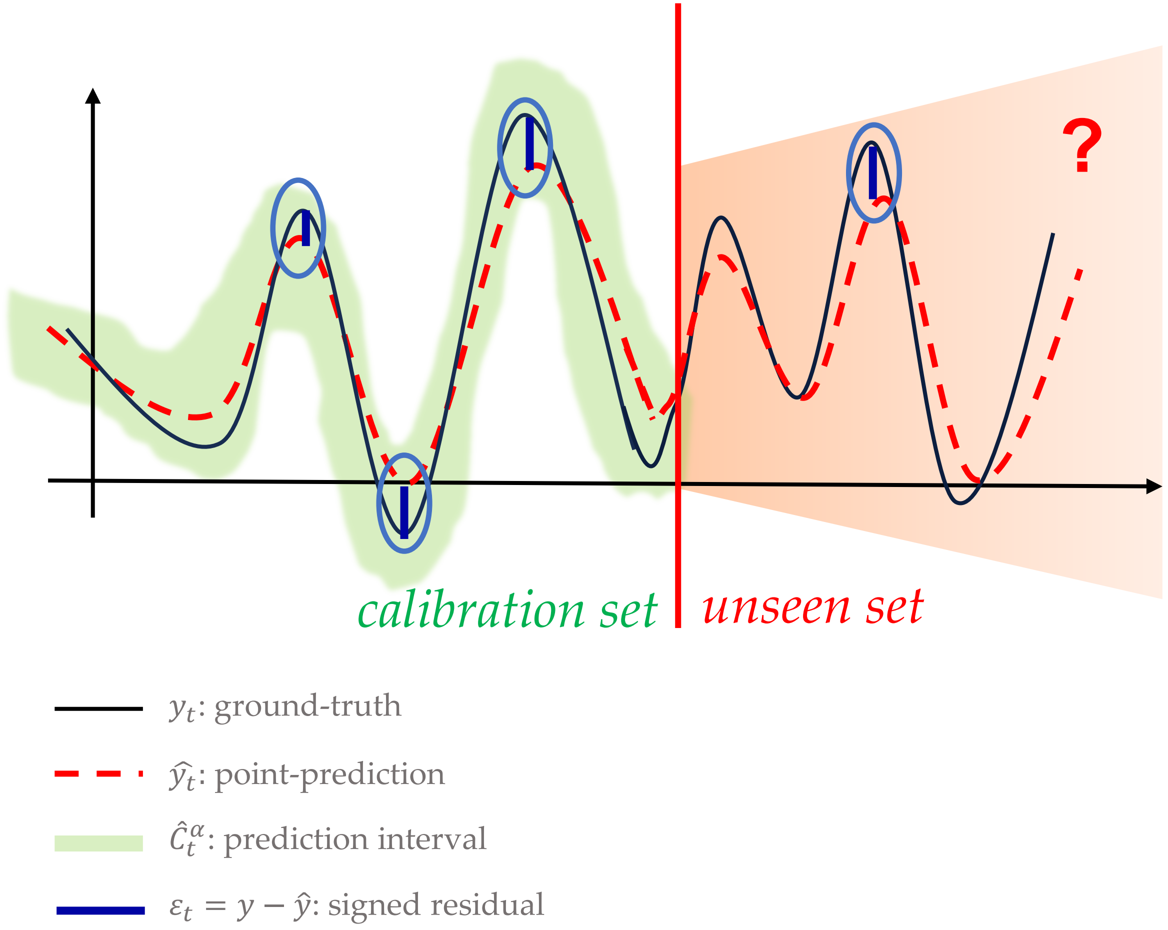
Exchangeability Failures in Time Series Data

In Sequential Conformal Prediction for Time Series, residuals exhibit temporal characteristics, so exchangeability is often violated.

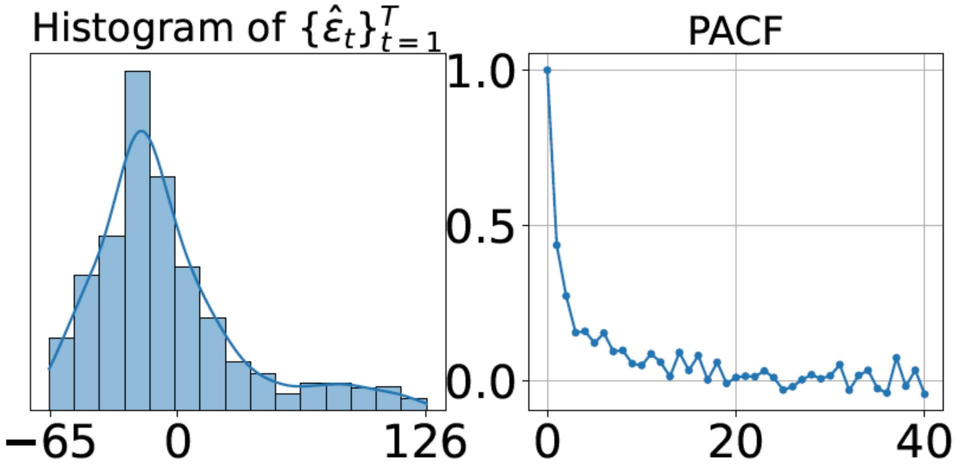


Exchangeability Failures in Time Series Data

In Sequential Conformal Prediction for Time Series, residuals exhibit temporal characteristics, so exchangeability is often violated.



Why does Exchangeability fail?



The histogram of prediction residuals (left) shows that the residual distribution is highly skewed, and the partial autocorrelation between residuals (right) shows a significant serial correlation among residuals [2].

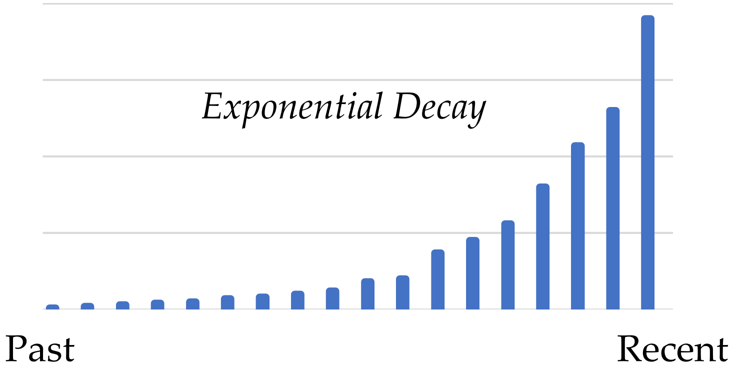
- Distributional shifts over time
- Temporal autocorrelation
- Non-stationary residuals

[2] Xu, Chen, and Yao Xie. "Sequential predictive conformal inference for time series." International Conference on Machine Learning. PMLR, 2023.

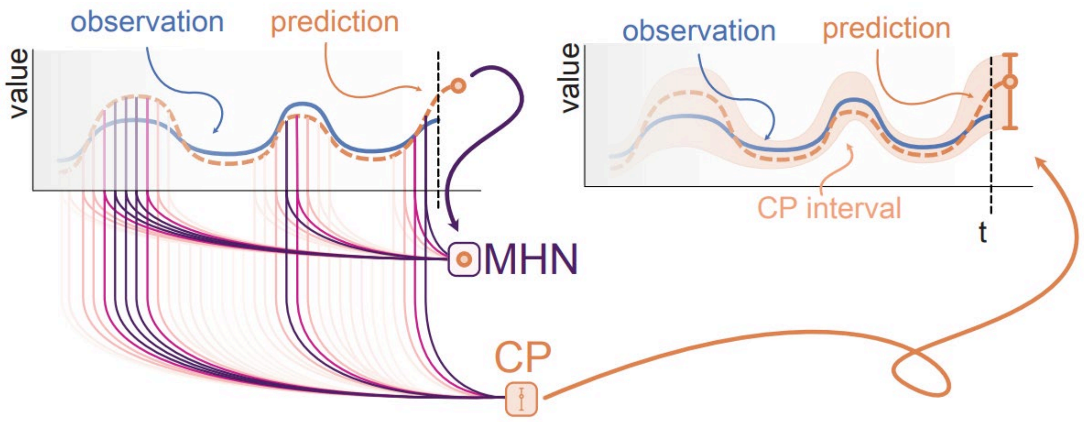
Related Work and Limitation

Q: How do existing methods address this?

A: **Reweighting.**



NexCP [3] computes weighted quantiles of past residuals using fixed, time-decaying weights to handle distribution drift.



HopCPT [4] retrieves past residuals with similar context to estimate the conditional residual distribution.

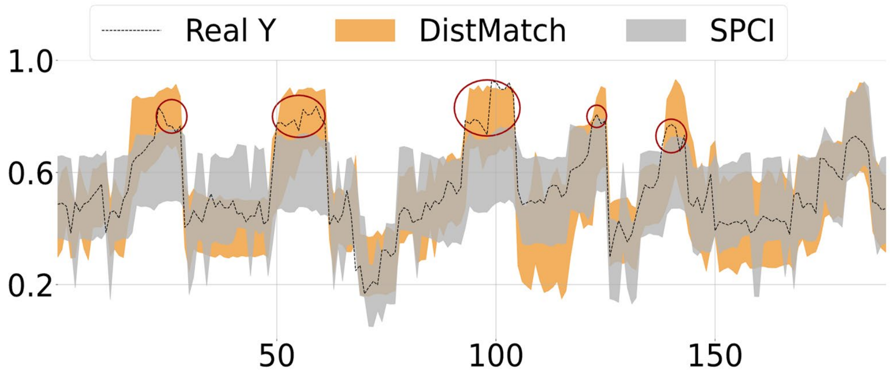
Reweighting approximates exchangeability by assigning data-dependent weights to past residuals.

⚠️ However, reweighting assigns ***continuous weights*** that can **distort the residual distribution** [3, 4].

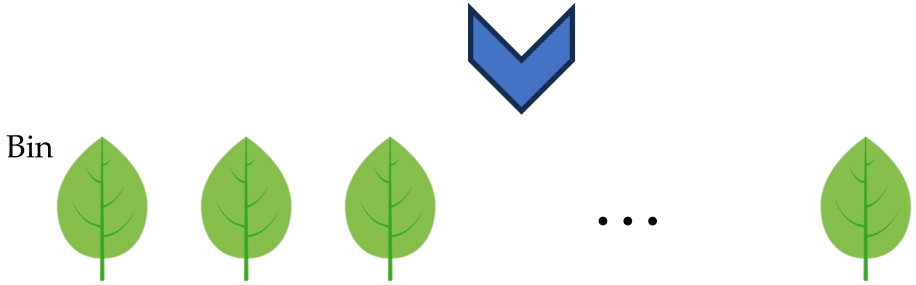
[3] Barber, Rina Foygel, et al. "Conformal prediction beyond exchangeability." The Annals of Statistics 51.2 (2023): 816-845.
[4] Auer, Andreas, et al. "Conformal prediction for time series with modern hopfield networks." Advances in neural information processing systems 36 (2023): 56027-56074.

Binning and KS statistic for Sequential CP

Binning preserves empirical distributions by avoiding explicit weighting schemes, provided that samples within each bin are approximately exchangeable.



Binning-based method (DistMatch-ours) produces valid coverage even under extreme distribution shifts.



Each bin should be approximately exchangeable!

Binning and KS statistic for Sequential CP

Q: What can be the best bin criteria?

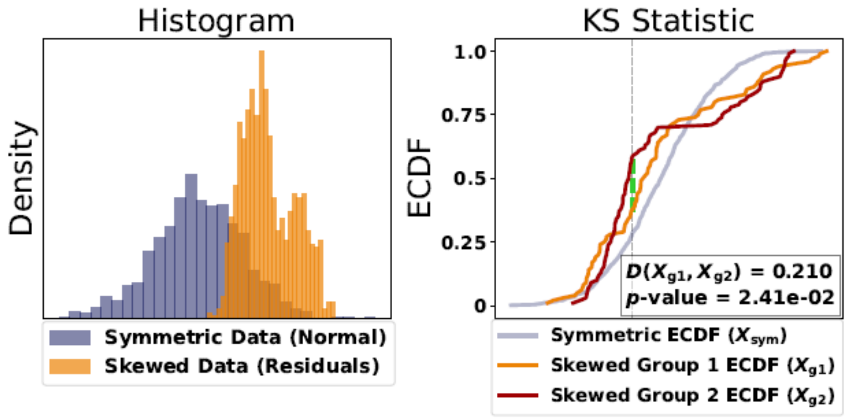
A: **Kolmogorov-Smirnov statistic.**[5]

Definition (Two-Sample KS Statistic). Let $S(\varepsilon) = \{\varepsilon_1, \dots, \varepsilon_t\}$ be a finite sample of residuals. Its empirical cumulative distribution function (ECDF) is defined as

$$F_{S(\varepsilon)}(x) = \frac{1}{|S(\varepsilon)|} \sum_{t \in \mathcal{I}(S(\varepsilon))} \mathbb{1}\{\varepsilon_t \le x\},$$

where $\mathcal{I}(\cdot)$ denotes the corresponding index set. Then, the *KS statistic (distance)* between $\mathrm{S}(\varepsilon_i)$ and $\mathrm{S}(\varepsilon_j)$ is defined as

$$D_{\mathrm{KS}}\left(S(\varepsilon_i), S(\varepsilon_j)\right) = \sup_{x \in \mathbb{R}} \left| F_{S(\varepsilon_i)}(x) - F_{S(\varepsilon_j)}(x) \right|.$$

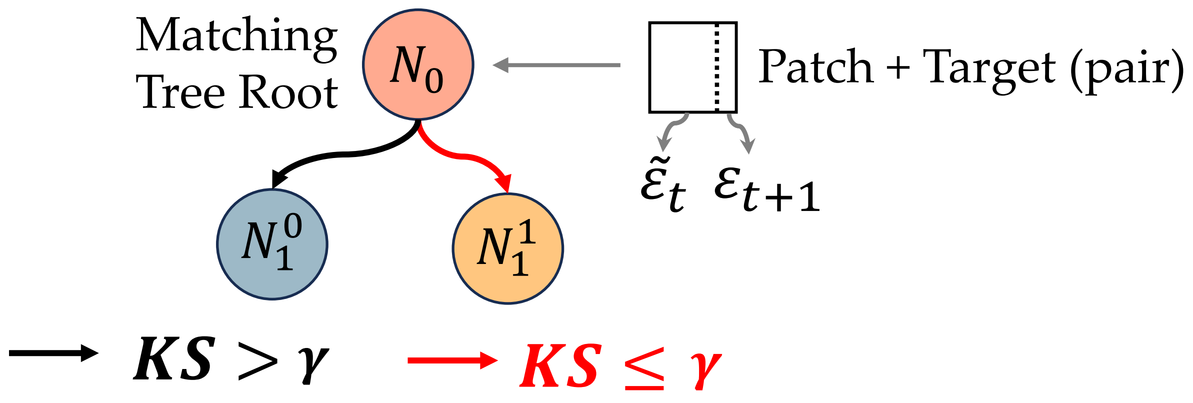


- Robust to skewed residuals
- Does not need temporal assumptions (i.e., global stationary)

[5] Massey Jr, Frank J. "The Kolmogorov-Smirnov test for goodness of fit." Journal of the American statistical Association 46.253 (1951): 68-78.

Distribution Matching - Train

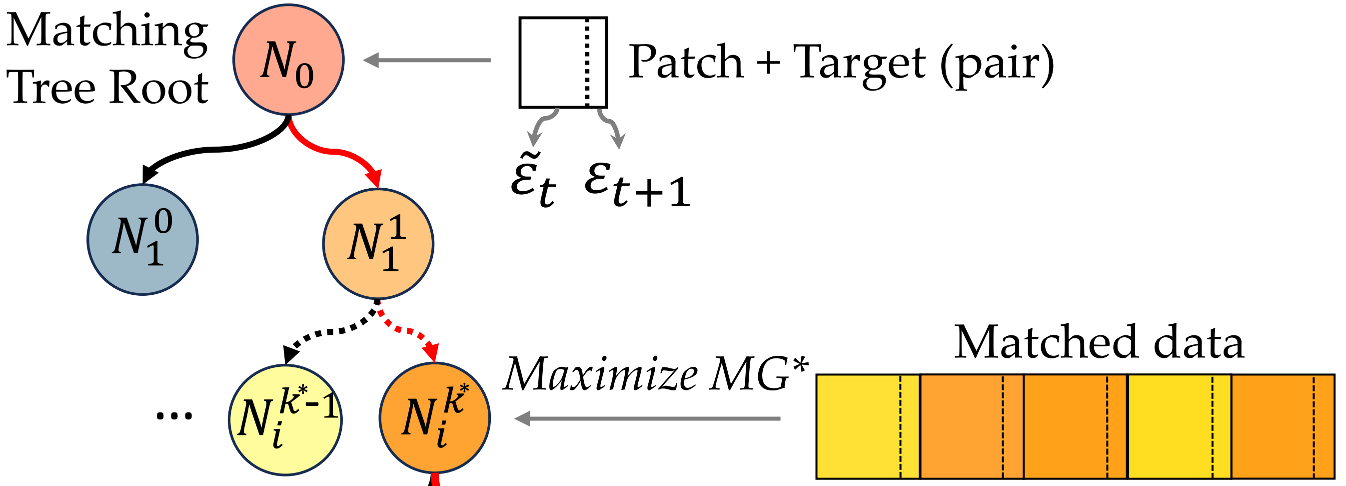
DistMatch builds a binary tree grouping similar residual patches via the KS statistic:

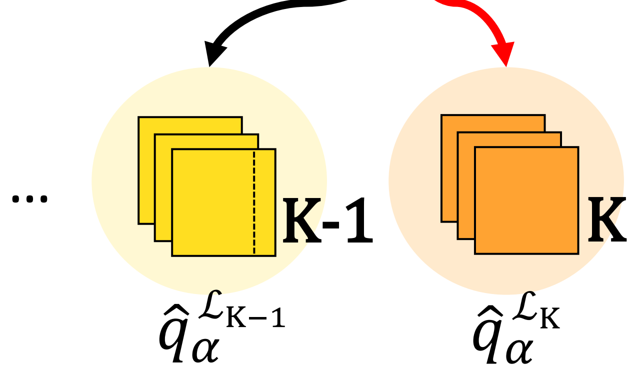


- $\tilde{\varepsilon}_t$: residual patch ($\{\varepsilon_{t-w+1}, \varepsilon_{t-w+2}, \dots, \varepsilon_t\}$)
- ε_{t+1}: target residual
- N_0: root node
- N_1^0, N_1^1: left/right child
- γ: KS threshold for tree splitting

Distribution Matching - Train

MG score quantifies how many samples are distributionally similar to a given candidate anchor:





* *MG* : Matching-based Gain score

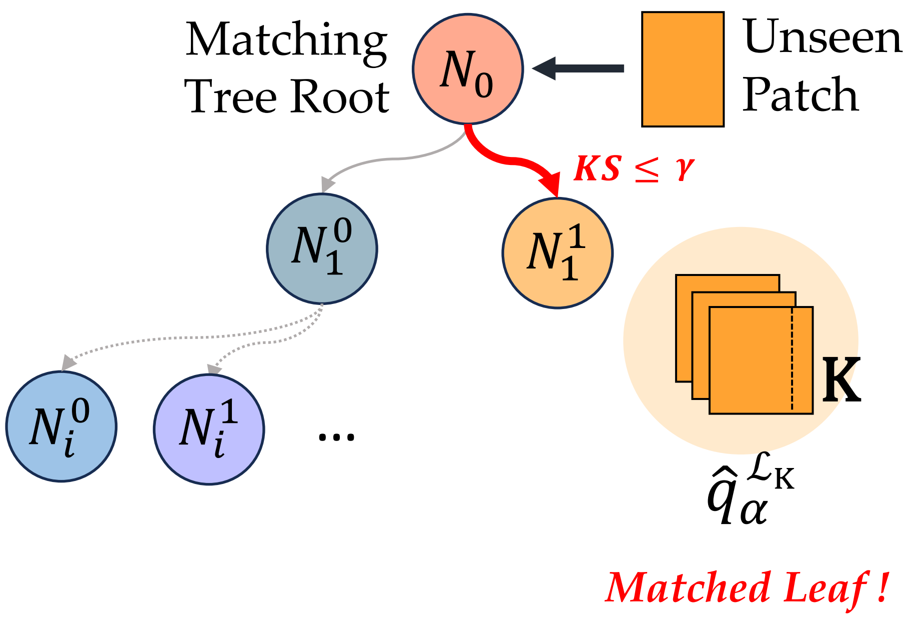
$$\mathrm{MG}(\tilde{\varepsilon}_i) = \sum_{j \in \mathcal{I}(\mathcal{D})} \mathbb{1}\{D_{\mathrm{KS}}(\tilde{\varepsilon}_i, \tilde{\varepsilon}_j) \le \gamma\}$$

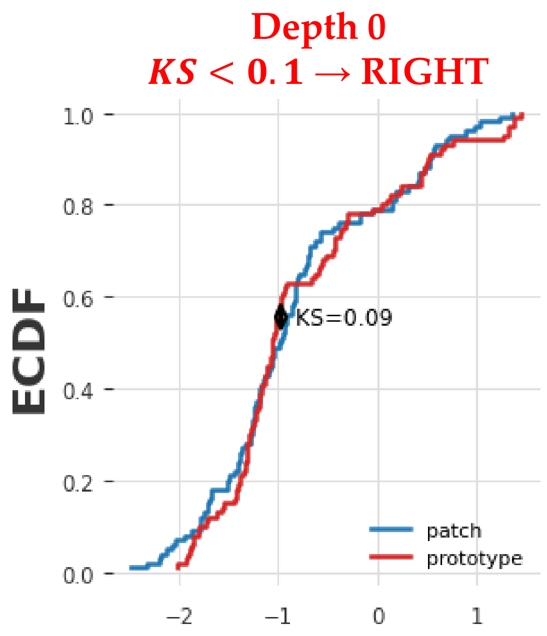
Get a split anchor (prototype)!

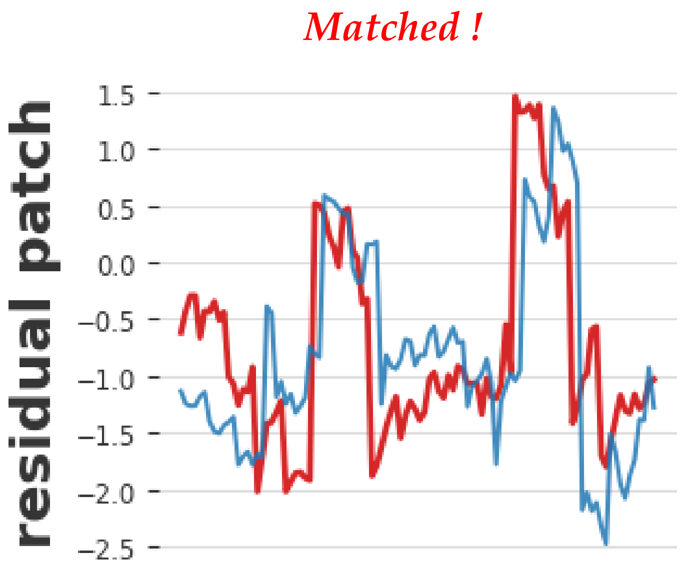
- $\mathcal{D}$: calibration data of patch-target pairs
- $\mathcal{I}(D)$: index set of $\mathcal{D}$
- $\mathbb{1}\{\cdot\}$: indicator function
- K: number of leaves
- $\mathcal{L}_k$: k-th leaf
- $\hat{q}_\alpha^{\mathcal{L}_k}$: quantile estimator within leaf

Distribution Matching – Inference (Real-world example 1)

When an unseen patch is matched directly to the right-most leaf:

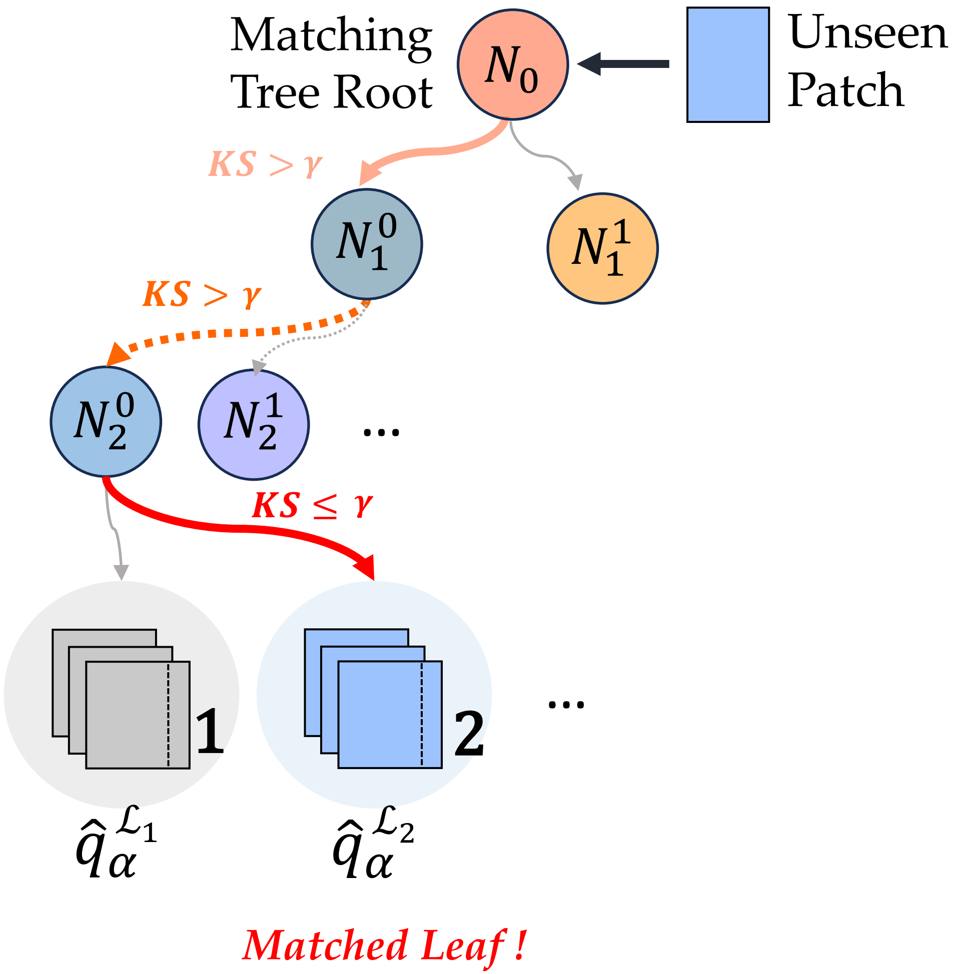


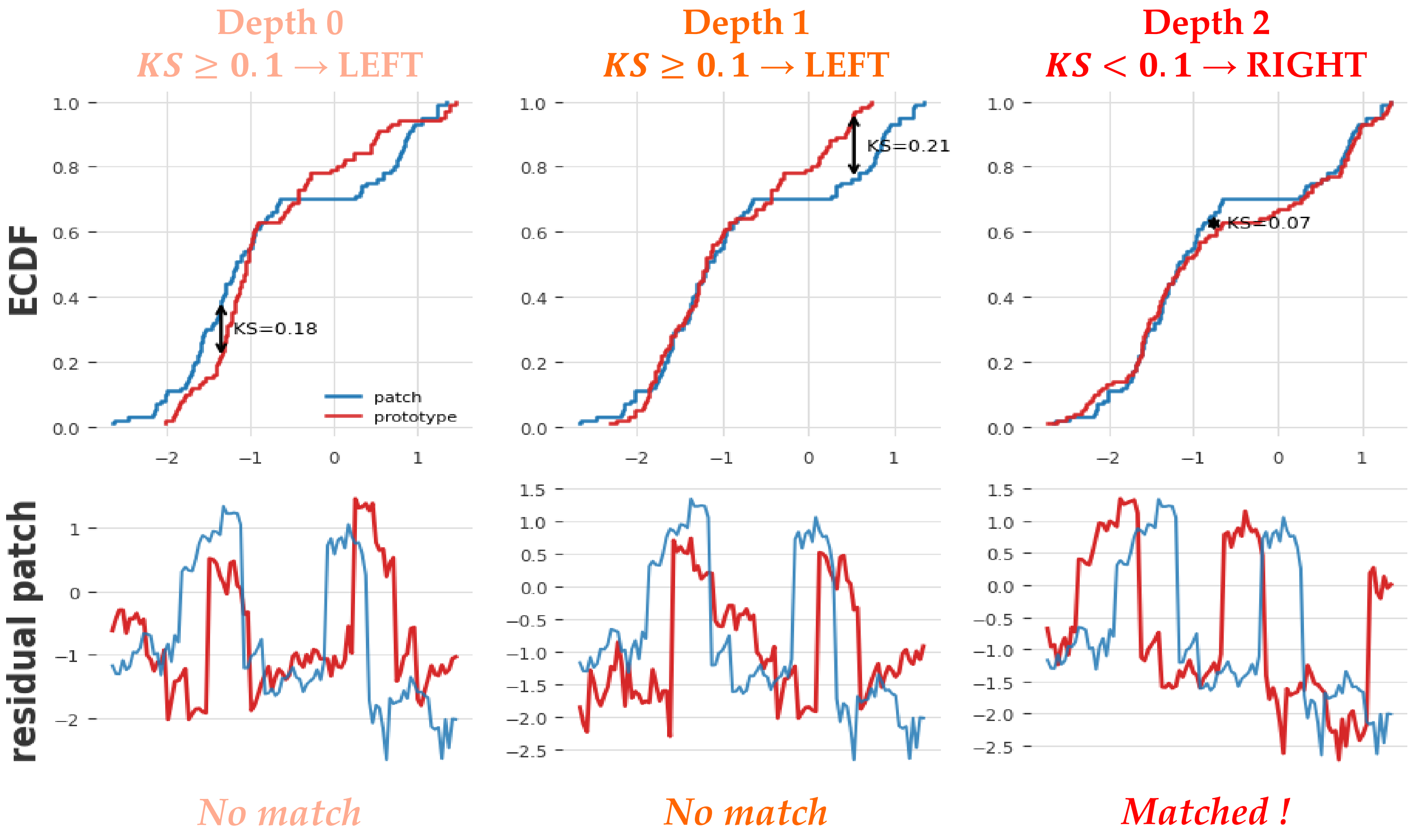




Distribution Matching – Inference (Real-world example 2)

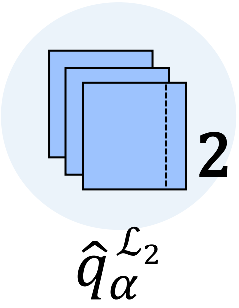
When an unseen patch is matched via tree traversal:

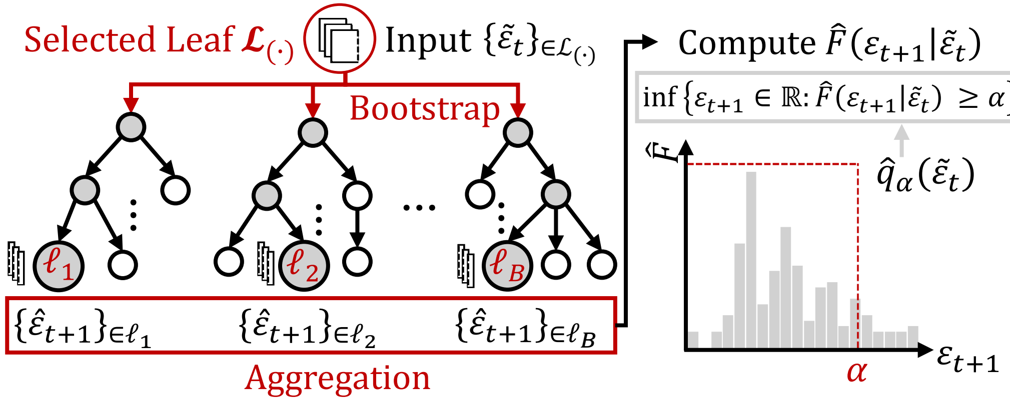




Distribution Matching - Inference

Given the selected leaf, the stored residual pairs are passed to QRF to estimate the unseen patch's target residual:





QRF [6] builds a forest by bootstrapping the input patches within a leaf and applies a quantile operator based on conditional ECDFs to estimate the target residual.

Additionally adopt an adaptive refinement strategy with $\delta^\star$.

$$\delta^\star = \mathrm{argmin}_{\delta\in[0,\alpha]} |\hat{q}_{1-\alpha+\delta}(\tilde{\varepsilon}_T) - \hat{q}_\delta(\tilde{\varepsilon}_T)|.$$

$$\hat{C}_T^\alpha(x_{T+1}) = [\hat{y}_{T+1} + \hat{q}_{\delta^\star}(\tilde{\varepsilon}_T),\ \hat{y}_{T+1} + \hat{q}_{1-\alpha+\delta^\star}(\tilde{\varepsilon}_T)].$$

[6] Meinshausen, N. Quantile regression forests. Journal of Machine Learning Research, 7:983–999, 2006.

Theory – Definition and Assumption

Definition ($\boldsymbol{\gamma^\star}$-Approximate Local Exchangeability). Let $\mathcal{L}_{(\ell^\star)}$ denote the leaf selected by $\tilde{\varepsilon}_T$. The target residuals $\{\varepsilon_{t+1}\}_{t\in\mathcal{I}(\mathcal{L}_{(\ell^\star)})}$ are $\gamma^\star$-approximately locally exchangeable with respect to the unseen target ε_{T+1} if

$$\max_{t\in\mathcal{I}(\mathcal{L}_{(\ell^\star)})} D_{\mathrm{KS}}(P_{t+1}, P_{T+1}) \le \gamma^\star,$$

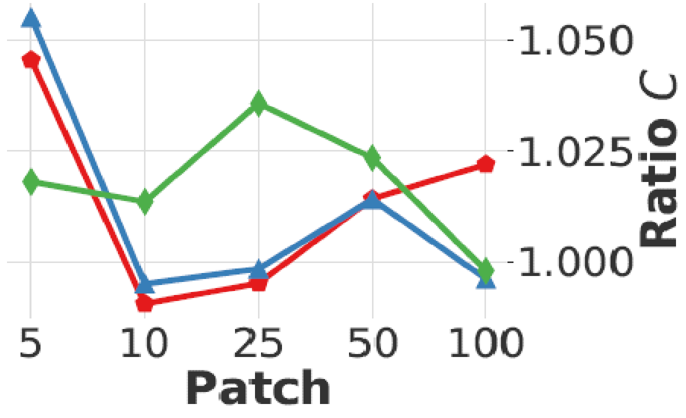
where P_{t+1} and P_{T+1} denote the distributions of ε_{t+1} and ε_{T+1}, respectively. Notably, $\gamma^\star = 0$ recovers the i.i.d. case, while $\gamma^\star > 0$ allows mild shifts between samples.

Theory – Definition and Assumption

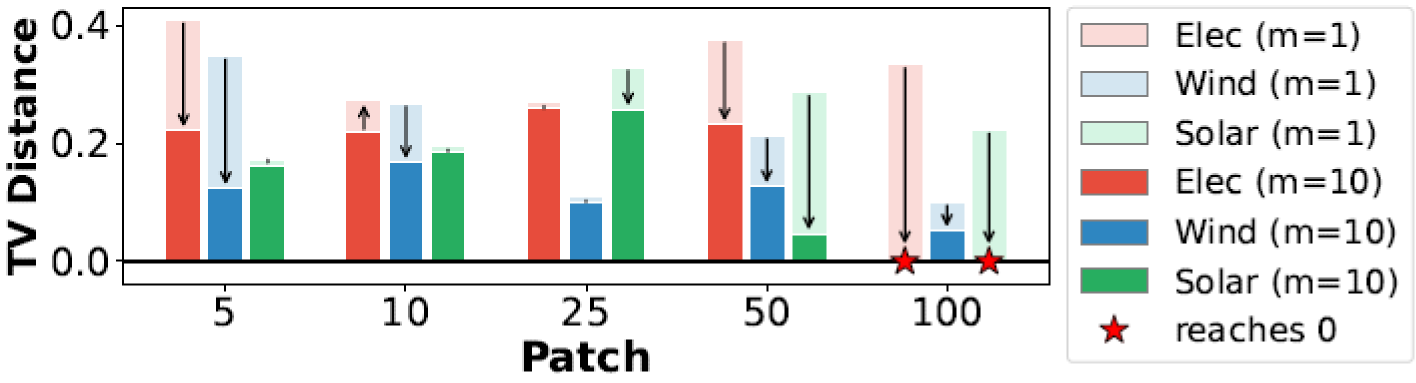
Assumption (Local Stationarity and Patch Mixing). The observed process is locally stationary in the sense of (Dahlhaus, 1997) [7], and the predictor ϕ is a smooth mapping, thus the residual process $\{\varepsilon_t = y_t - \phi(x_t)\}$ inherits local stationarity. Consequently, the patch sequence $\{\tilde{\varepsilon}_t\}$ is β-mixing with coefficients $\{\beta_m\}_{m\geq 1}$ satisfying $\sum \beta_m^{1/2} < \infty$. In addition, there exists a conditional distribution kernel $\mathcal{K}$ with $\varepsilon_{t+1} \mid \tilde{\varepsilon}_t \sim \mathcal{K}(\tilde{\varepsilon}_t)$ that is KS-stable: for any patch distributions $\tilde{P}, \tilde{Q}$,

$$D_{\mathrm{KS}}(\tilde{P}\mathcal{K}, \tilde{Q}\mathcal{K}) \leq C\, D_{\mathrm{KS}}(\tilde{P}, \tilde{Q}),$$

where $\tilde{P}\mathcal{K}$ denotes the marginal distribution of ε_{t+1} when $\tilde{\varepsilon}_t \sim \tilde{P}$.



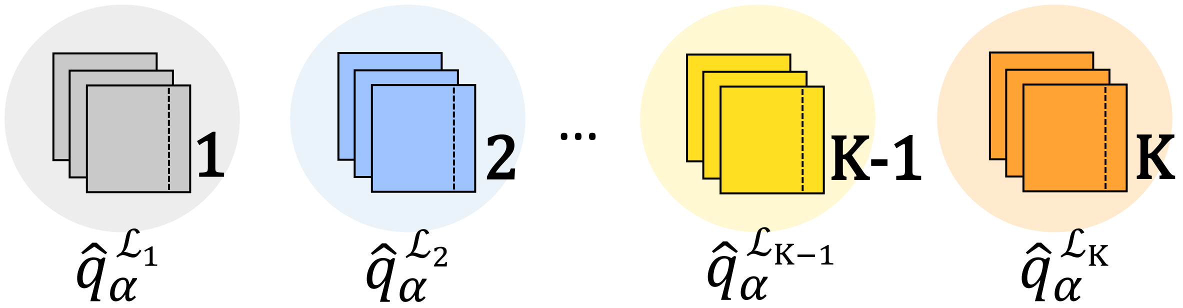
$C = \mathcal{O}(1)$ *in real-world datasets.*

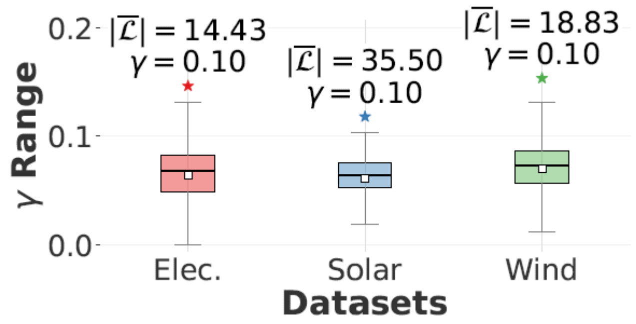


β-mixing holds in real-world datasets: TV distance generally decreases as the lag m increases across various datasets and patch sizes.

[7] Dahlhaus, R. Fitting time series models to nonstationary processes. The Annals of Statistics, 25(1):1–37, 1997.

Theory – Convergence of Algorithm





Theorem of convergence holds in practice: with fixed $\gamma = 0.1$, the range of γ within leaves satisfies < 0.2

Theorem (Convergence).

Every leaf satisfies:

$$\max_{i,j \in \mathcal{I}(\mathcal{L}_{(\cdot)})} D_{\text{KS}}(P_i, P_j) \leq 2\gamma$$

- $\mathcal{L}_{(k^*)}$: leaf selected by unseen patch
- P_i, P_j: patch distributions at indices i, j
- $\mathcal{L}_{(\cdot)}$: any leaf
- γ: KS threshold for tree splitting

Theory – Coverage Guarantee

UC Model		DistMatch (Ours)	KOWCPI (2025)	HopCPT (2023)	SPCI (2023)	EnbPI (2021)	NexCP (2023)	CP (2021)	Copula (2024)
Elec.	Win. ↓	**$1.97^{\pm 0.01}$**	$2.44^{\pm 0.00}$	$3.35^{\pm 0.21}$	$2.54^{\pm 0.06}$	3.36	3.51	3.93	3.77
	Cov. ↑	$0.92^{\pm 0.00}$	$0.90^{\pm 0.00}$	$0.91^{\pm 0.02}$	$0.90^{\pm 0.00}$	0.85	0.92	0.97	0.96
	Width ↓	**$0.27^{\pm 0.00}$**	$0.38^{\pm 0.00}$	$0.50^{\pm 0.05}$	$0.28^{\pm 0.00}$	0.45	0.57	0.71	0.67
Solar	Win. ↓	**$1.54^{\pm 0.01}$**	$2.07^{\pm 0.00}$	$1.90^{\pm 0.13}$	$1.98^{\pm 0.15}$	2.55	3.74	3.56	3.54
	Cov. ↑	$0.91^{\pm 0.00}$	$0.93^{\pm 0.00}$	$0.92^{\pm 0.01}$	$0.85^{\pm 0.01}$	0.88	0.89	0.86	0.88
	Width ↓	**$60.00^{\pm 0.40}$**	$109.31^{\pm 0.00}$	$97.04^{\pm 8.21}$	$47.36^{\pm 5.14}$	112.57	215.67	167.91	187.69
Wind	Win. ↓	**$2.15^{\pm 0.01}$**	$3.54^{\pm 0.00}$	$3.72^{\pm 0.31}$	$2.19^{\pm 0.00}$	4.37	3.98	4.40	4.10
	Cov. ↑	$0.90^{\pm 0.00}$	$0.89^{\pm 0.00}$	$0.90^{\pm 0.01}$	$0.83^{\pm 0.00}$	0.85	0.89	0.74	0.82
	Width ↓	**$69.04^{\pm 0.03}$**	$139.25^{\pm 0.00}$	$151.66^{\pm 12.82}$	$63.144^{\pm 0.18}$	145.63	171.67	147.57	159.00

Theorem (Coverage Guarantee). *Let $\mathcal{L}_{(\Bbbk^\star)}$ be the leaf selected by the unseen patch $\tilde{\varepsilon}_T$ among $\{\mathcal{L}_\Bbbk\}_{\Bbbk=1}^{\mathcal{K}}$, and let $\hat{q}_{1-\alpha+\delta^\star}^{\mathcal{L}_{(\Bbbk^\star)}}$ be any consistent estimator of the $(1-\alpha+\delta^\star)$-conditional quantile constructed within $\mathcal{L}_{(\Bbbk^\star)}$, with $\delta^\star \in [0,\alpha]$.*
Define the within-leaf target divergence

$$m_{\mathcal{L}_{(\Bbbk^\star)}} := \max_{t\in\mathcal{I}(\mathcal{L}_{(\Bbbk^\star)})} D_{\mathrm{KS}}(P_{t+1}, P_{T+1}) \le 2C\gamma + \mathcal{O}(\sigma_{\mathrm{mix}}),$$

where $C = \mathcal{O}(1)$ depends only on the regularity of the conditional kernel $\mathcal{K}$ and $\sigma_{\mathrm{mix}} = \mathcal{O}\left(\sum_{m\ge 1}\sqrt{\beta_m}\right)$.
Then, with probability at least $1-\xi$,

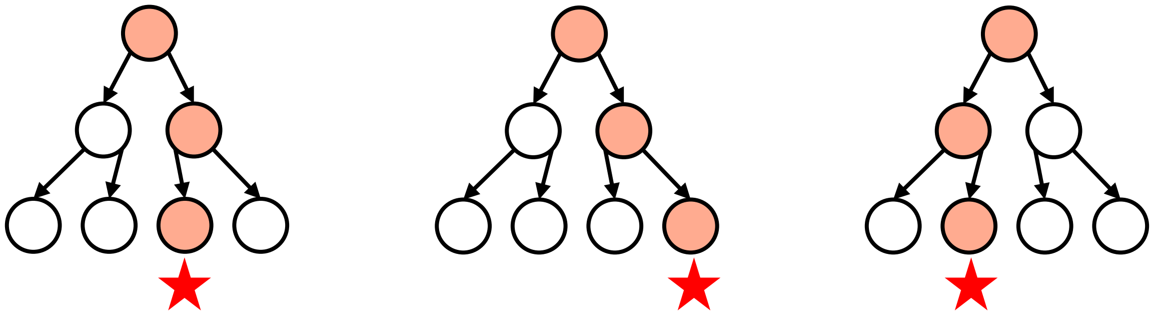
$$\left|\mathbb{P}\left(\varepsilon_{T+1} \le \hat{q}_{1-\alpha+\delta^\star}^{\mathcal{L}_{(\Bbbk^\star)}}\right) - (1-\alpha)\right| \le m_{\mathcal{L}_{(\Bbbk^\star)}} + \varrho_{\mathcal{L}_{(\Bbbk^\star)}}(\xi) + \delta^\star + 1/\left|\mathcal{L}_{(\Bbbk^\star)}\right|$$

$$\textit{where } \varrho_{\mathcal{L}_{(\Bbbk^\star)}}(\xi) = \mathcal{O}\left(\sigma_{\mathrm{mix}}\sqrt{\log(1/\xi)\,/\left|\mathcal{L}_{(\Bbbk^\star)}\right|}\right).$$

Theory – Bootstrapping and OOD

Case 1: mild shift

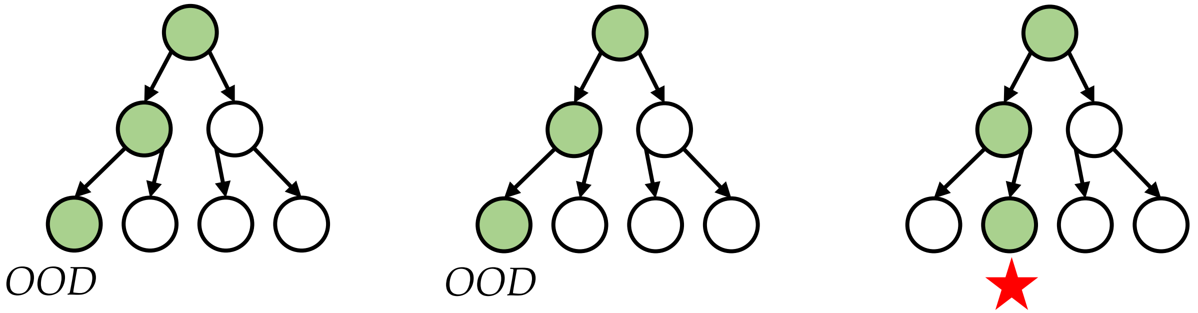
Unseen Patch **A**



Patch **A** is routed to [Matched] leaves in ALL trees.

Case 2: extreme shift

Unseen Patch **B**

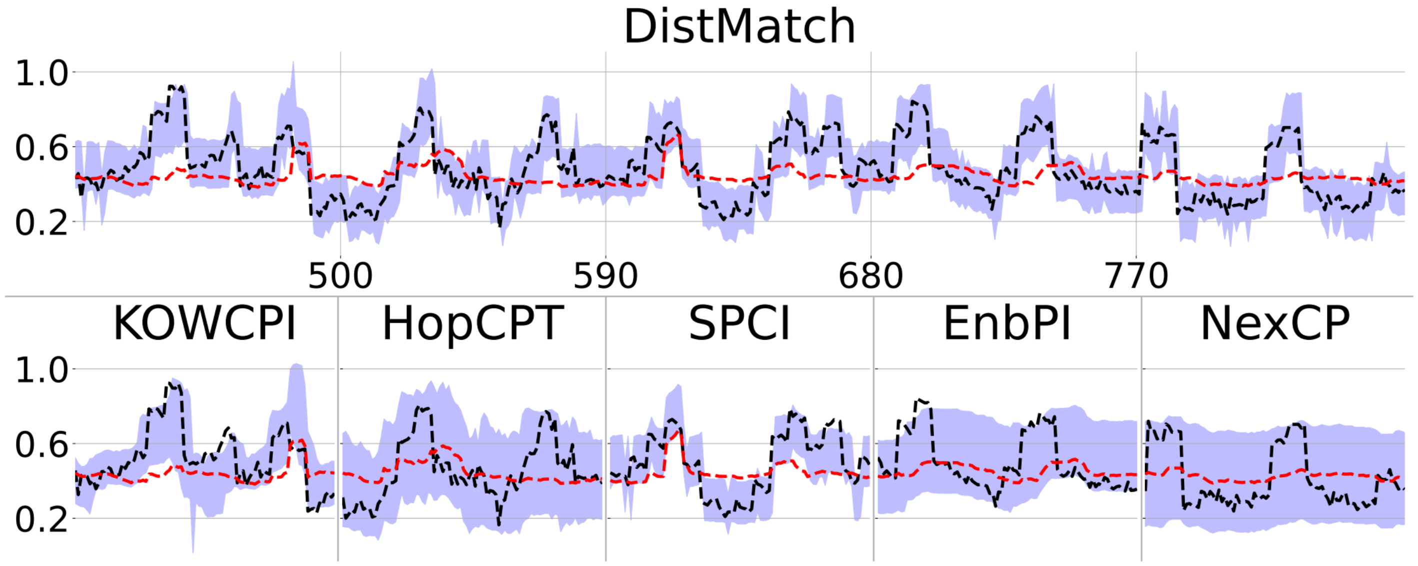


At least one tree routes the Patch **B** to a [Matched] leaf!

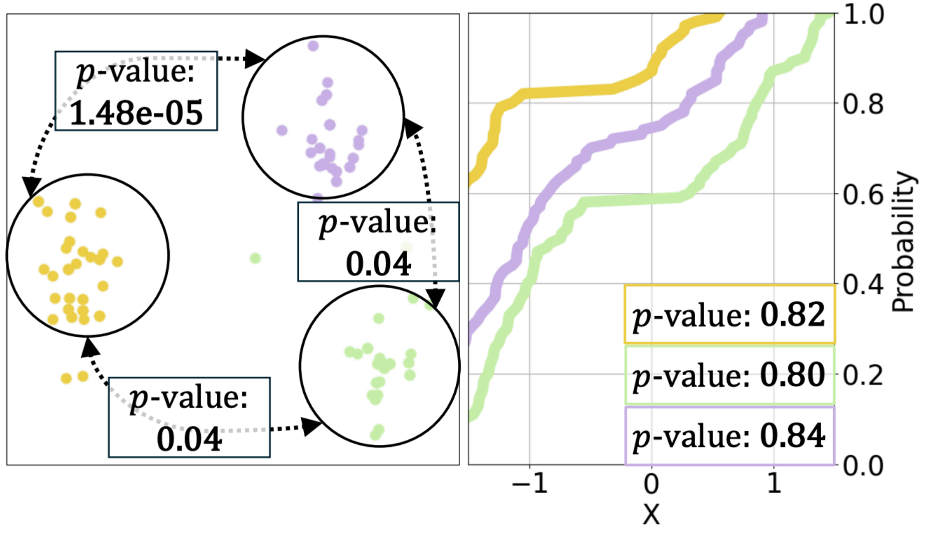
- *OOD*: routed to leftmost leaf

Qualitative Results

(a) DistMatch yields narrow, adaptive intervals under distribution shifts, while baselines struggle.
(b) Residuals within a leaf share similar distributions ($p > 0.05$), while those across leaves differ ($p \leq 0.05$).



(a) Visual Comparison of Prediction Intervals.



(b) Clustered Results and p-value from DistMatch.

Contributions

- First binning-based sequential CP framework using KS-based distribution matching, replacing the reweighting scheme.
- Proves approximate exchangeability in leaves, enabling a valid coverage guarantee even under distribution shifts.
- State-of-the-art performance across various real-world datasets.

Thank You!

Enver Menadjiev: enver@kaist.ac.kr

Jihyeon Seong: jihyeon.seong@kaist.ac.kr

Jisu Yeo: jisu.yeo@kaist.ac.kr

Jaesik Choi: jaesik.choi@kaist.ac.kr