# Wasserstein-regularized Conformal Prediction under General Distribution Shift

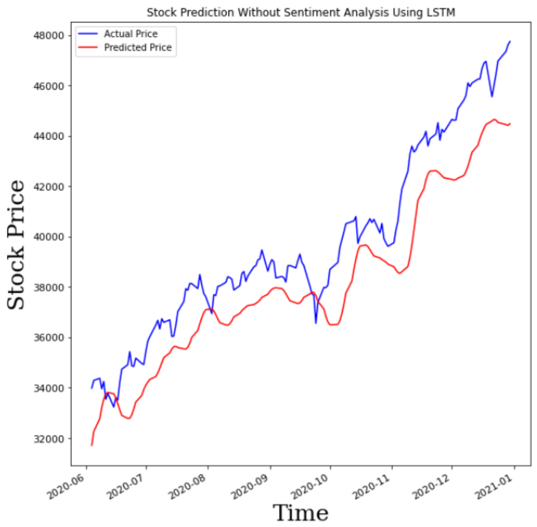
Rui Xu[1], Chao Chen[2], Yue Sun[3],
Parvathinathan Venkitasubramaniam[3], Sihong Xie[1]

1. The Hong Kong University of Science and Technology (Guangzhou), HKUST(GZ)
2. Harbin Institute of Technology, HIT
3. Lehigh University

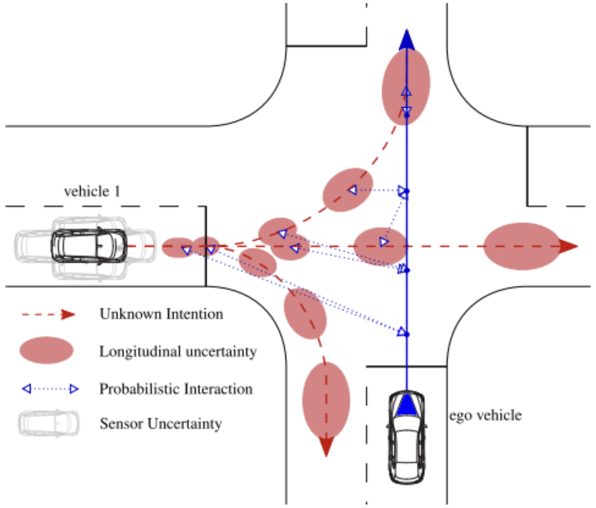
HKUST(GZ) HIT Lehigh

# Background: Uncertainty Quantification

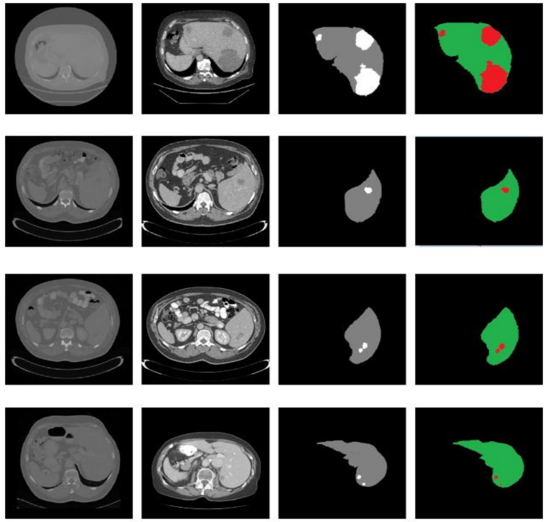
It is necessary to consider prediction uncertainty to be aware of potential risks in high-stake fields .




Fintech: Stock Forecasting [1]




Auto-driving Path Planning [2]



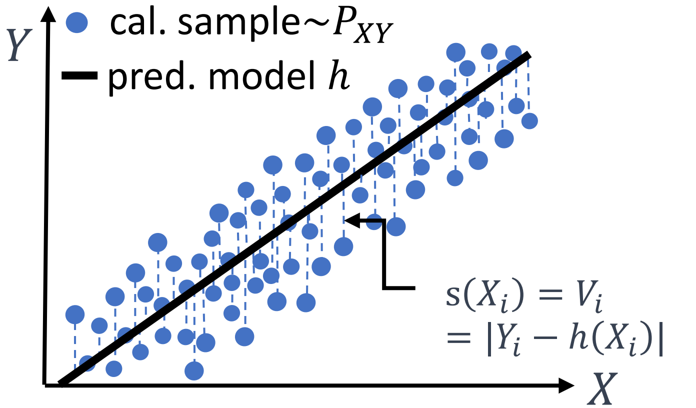
Medicine Image Segment [3]


[1] Kasture, P., and K. Shirsath. "Enhancing Stock Market Prediction: A Hybrid RNN-LSTM Framework with Sentiment Analysis." *Indian Journal of Science and Technology* 17.18 (2024): 1880-1888.
[2] Hubmann, Constantin, et al."Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction." IEEE transactions on intelligent vehicles 3.1(2018).
[3] Alalwan, Nasser, et al. "Efficient 3D deep learning model for medical image semantic segmentation." *Alexandria Engineering Journal* 60.1 (2021): 1231-1239.

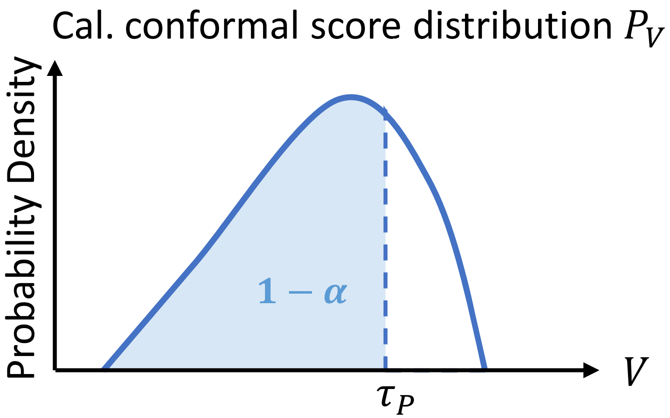
# Background: Conformal Prediction

**Conformal Prediction (CP)** uses calibration samples $\{(X_i, Y_i)\}_{i=1}^{n}$ to output a prediction set of **independent and identically distributed (i.i.d.)** test sample $(X_{n+1}, Y_{n+1})$ with a $1-\alpha$ **coverage guarantee:**

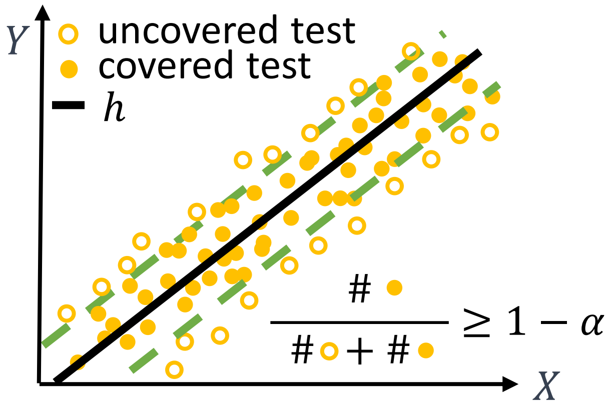
$$\Pr\left(Y_{n+1} \in C(X_{n+1})\right) \in \left[1-\alpha, 1-\alpha+\frac{1}{n+1}\right).$$




Conformal scores $\{V_i\}_{i=1}^{n}$ can be residuals



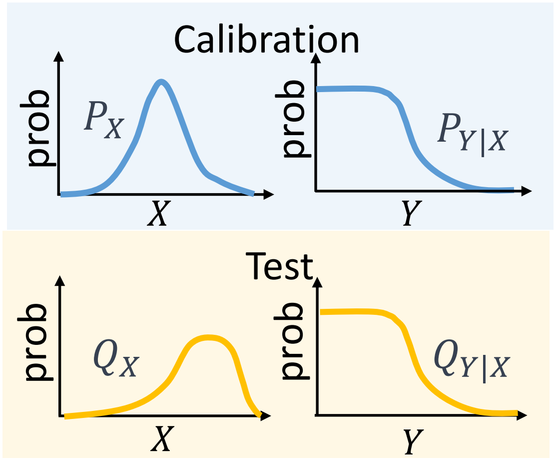

$\tau_P$ is empirical $1-\alpha$ quantile

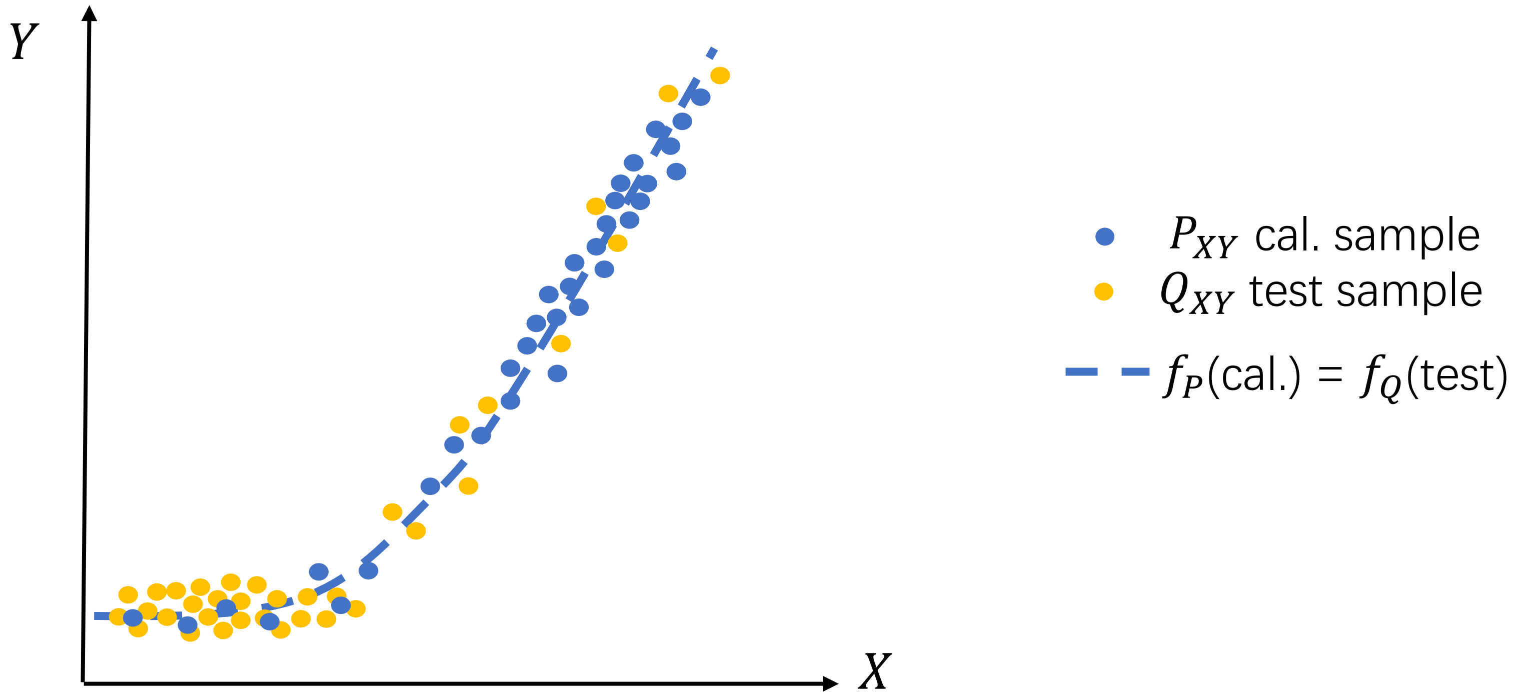



Prediction Set $C(X_{n+1}) = \{y: |y - h(X_{n+1})| \leq \tau_P\}$

# Challenge: CP under Distribution Shift

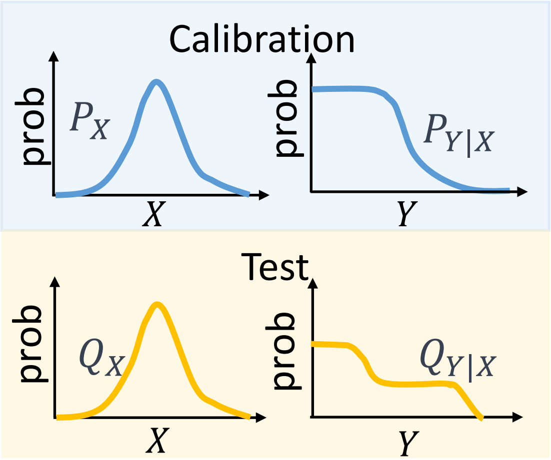
Covariate shift ($P_X \neq Q_X, P_{Y|X} = Q_{Y|X}$)

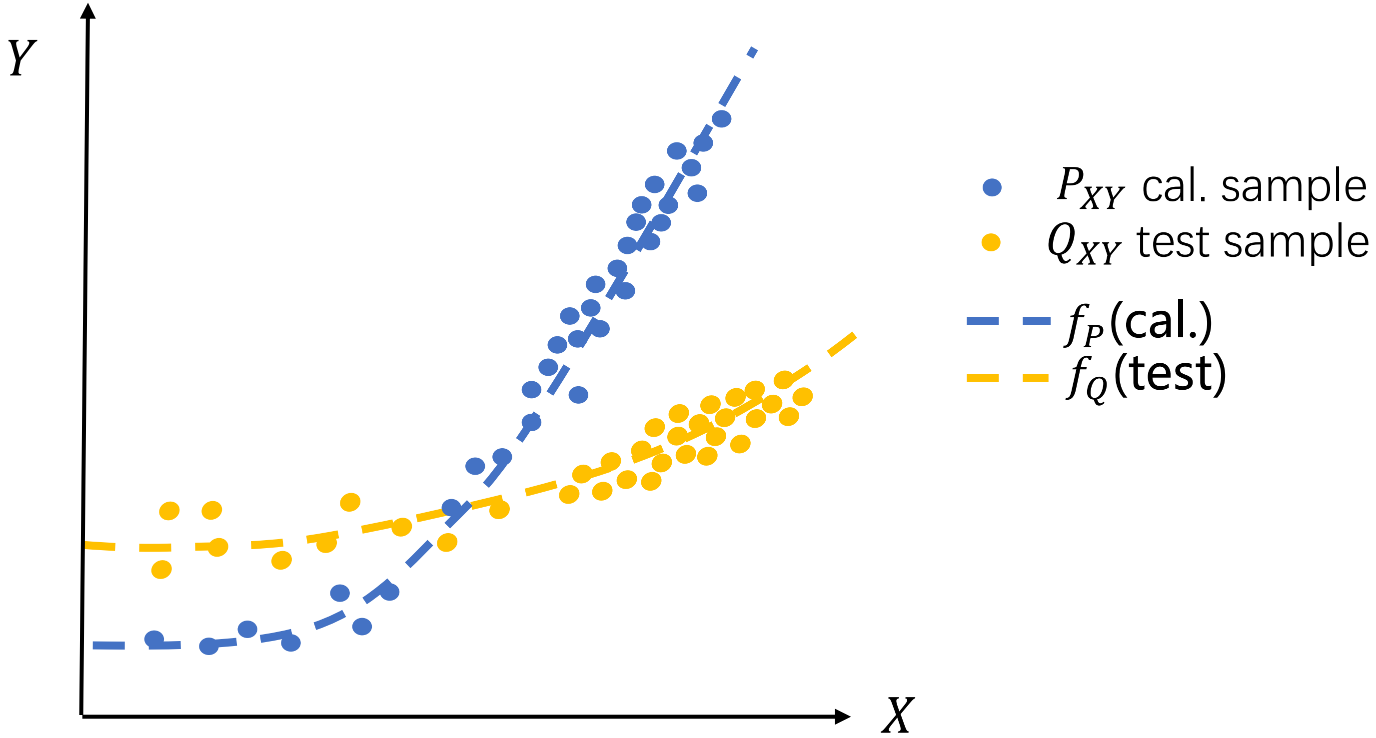


# Challenge: CP under Distribution Shift

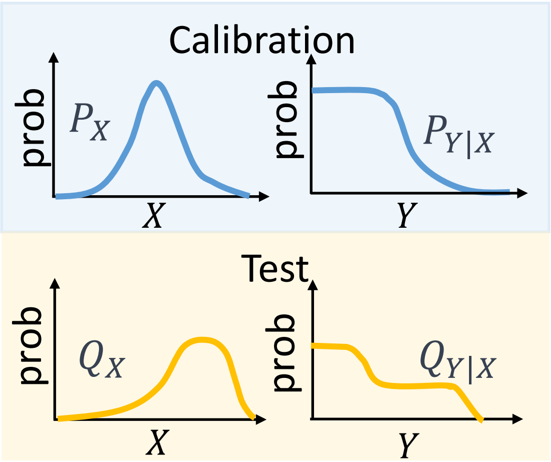
Concept shift $(P_X = Q_X, P_{Y|X} \neq Q_{Y|X})$

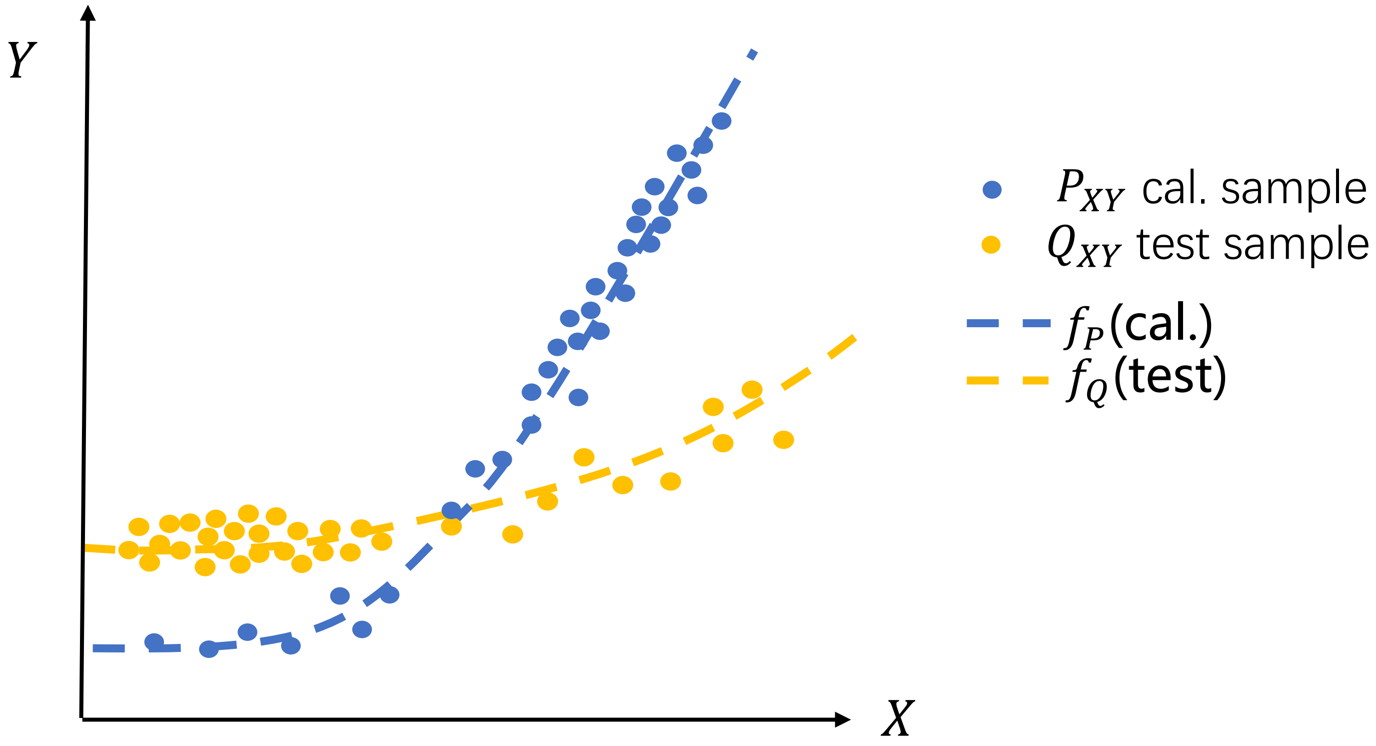


# Challenge: CP under Distribution Shift

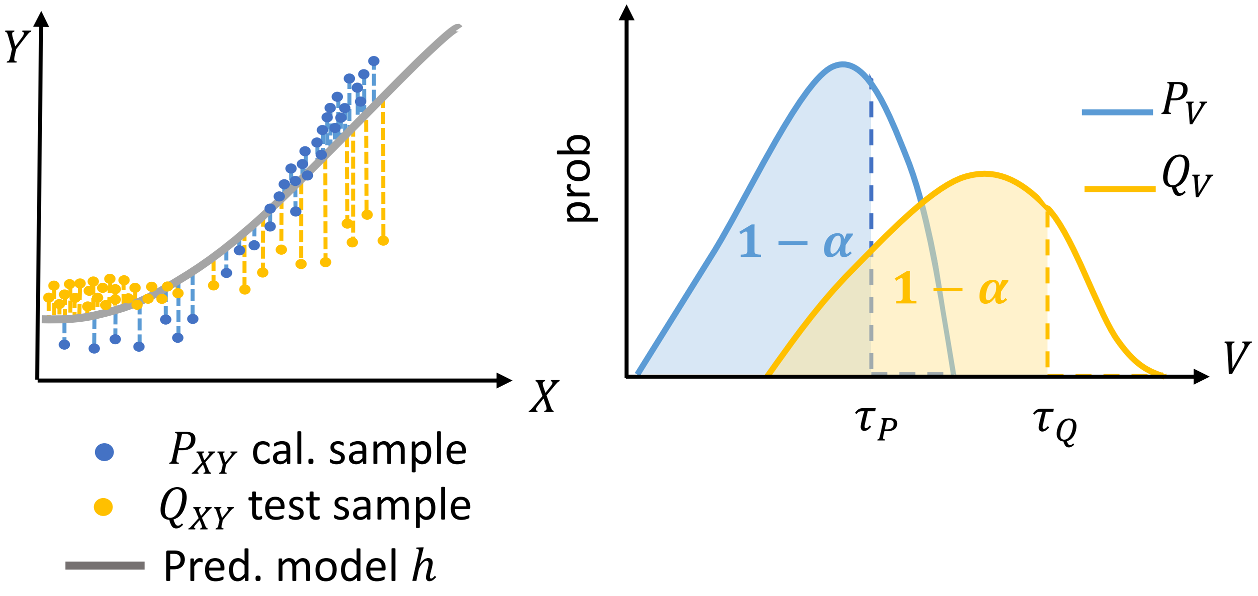
Joint distribution shift $(P_X = Q_X, P_{Y|X} \neq Q_{Y|X})$



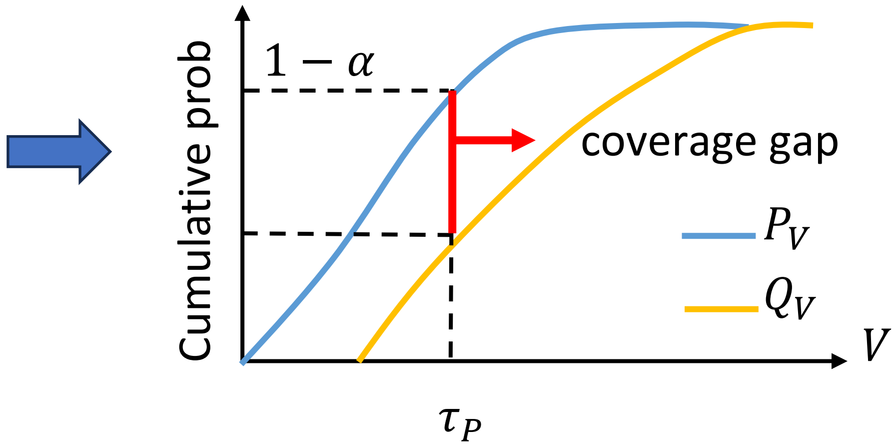



# Challenge: CP under Distribution Shift

Calibration and test conformal score distributions, $P_V$ and $Q_V$, are different.



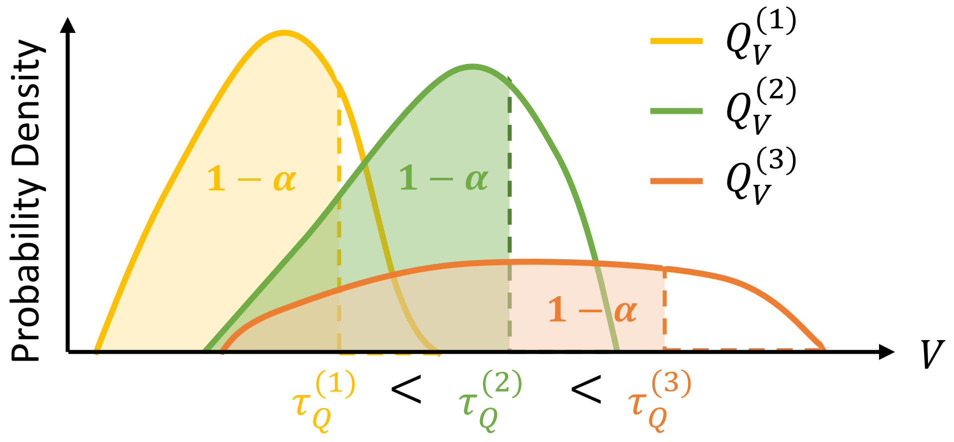

**Coverage guarantee does not hold.**



# Challenge: Prediction Inefficiency

Existing worst-case (WC) solutions for robust coverage make prediction set $C(X_{n+1})$ excessively large [4,5,6], which are less informative to locate the true label.

Potential test conformal score distributions

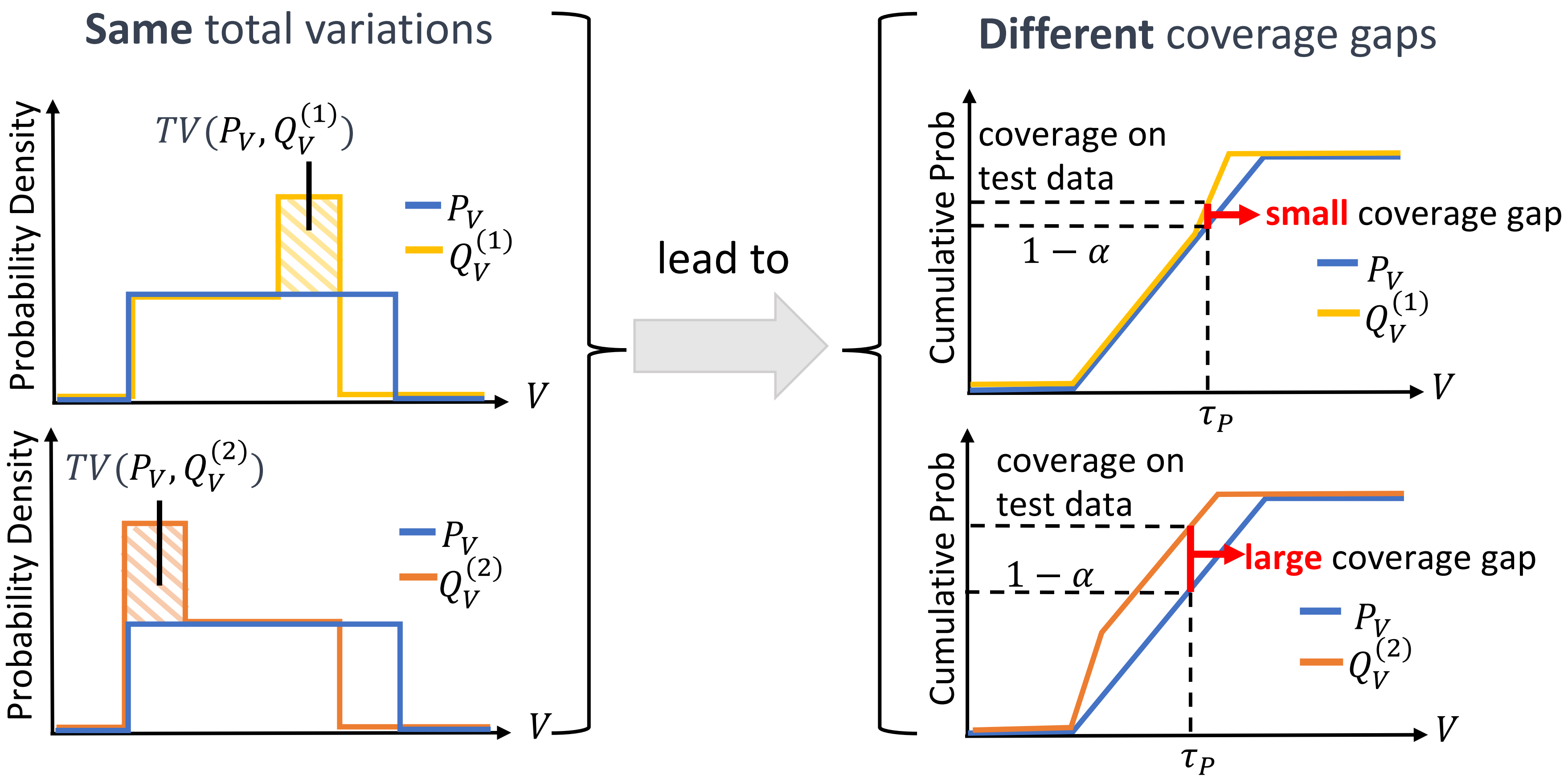



$C(X_{n+1}) = \{y: |y - h(X_{n+1})| \leq \tau_Q^{(3)}\}$ is unnecessarily large for test samples from $Q_{XY}^{(1)}$ and $Q_{XY}^{(2)}$.


. [4] Gendler, Asaf, et al. "Adversarially robust conformal prediction." International Conference on Learning Representations. 2021.
[5] Cauchois, Maxime, et al. "Robust validation: Confident predictions even when distributions shift." Journal of the American Statistical Association 119.548 (2024): 3033-3044.
[6] Zou, Xin, and Weiwei Liu. "Coverage-guaranteed prediction sets for out-of-distribution data." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 15. 2024

# Challenge: Agnostic Distribution Divergence Location

Total variation ($TV$) between calibration and test conformal score distributions, $P_V$ and $Q_V$, fails to indicate coverage gap changes without knowing where two distributions diverge[7].

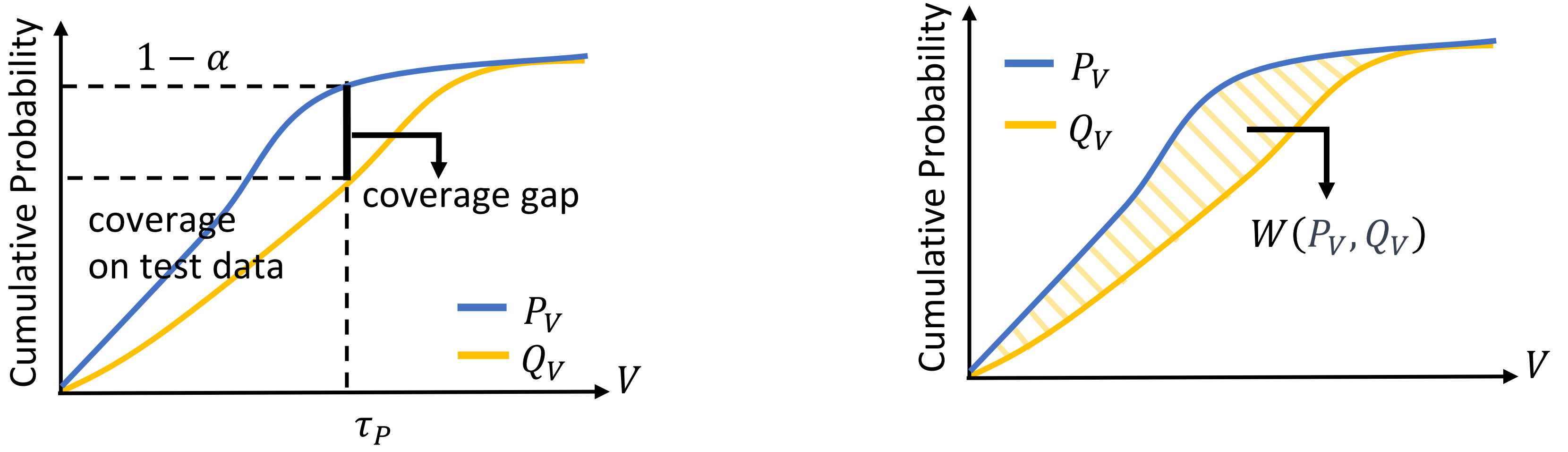




[7] Barber, Rina Foygel, et al. "Conformal prediction beyond exchangeability." *The Annals of Statistics* 51.2 (2023): 816-845.

# Method: Upper-bounding Coverage Gap by Wasserstein Distance

Wasserstein distance integrate coverage gap over ***all*** quantiles, consistently indicating coverage robustness.




Definition of Coverage Gap

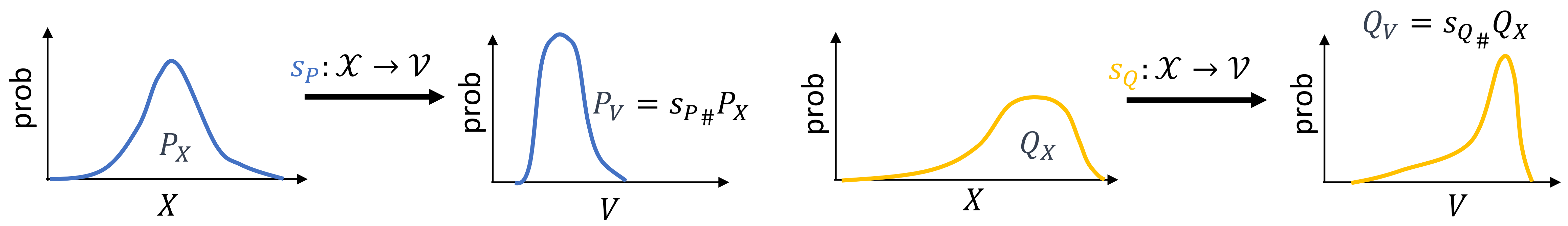
Wasserstein-1 Distance $W(P_V, Q_V)$

Given $L$ the Lebesgue density bound of $P_V$, we derive that coverage gap $\leq \sqrt{2L \cdot W(P_V, Q_V)}$.

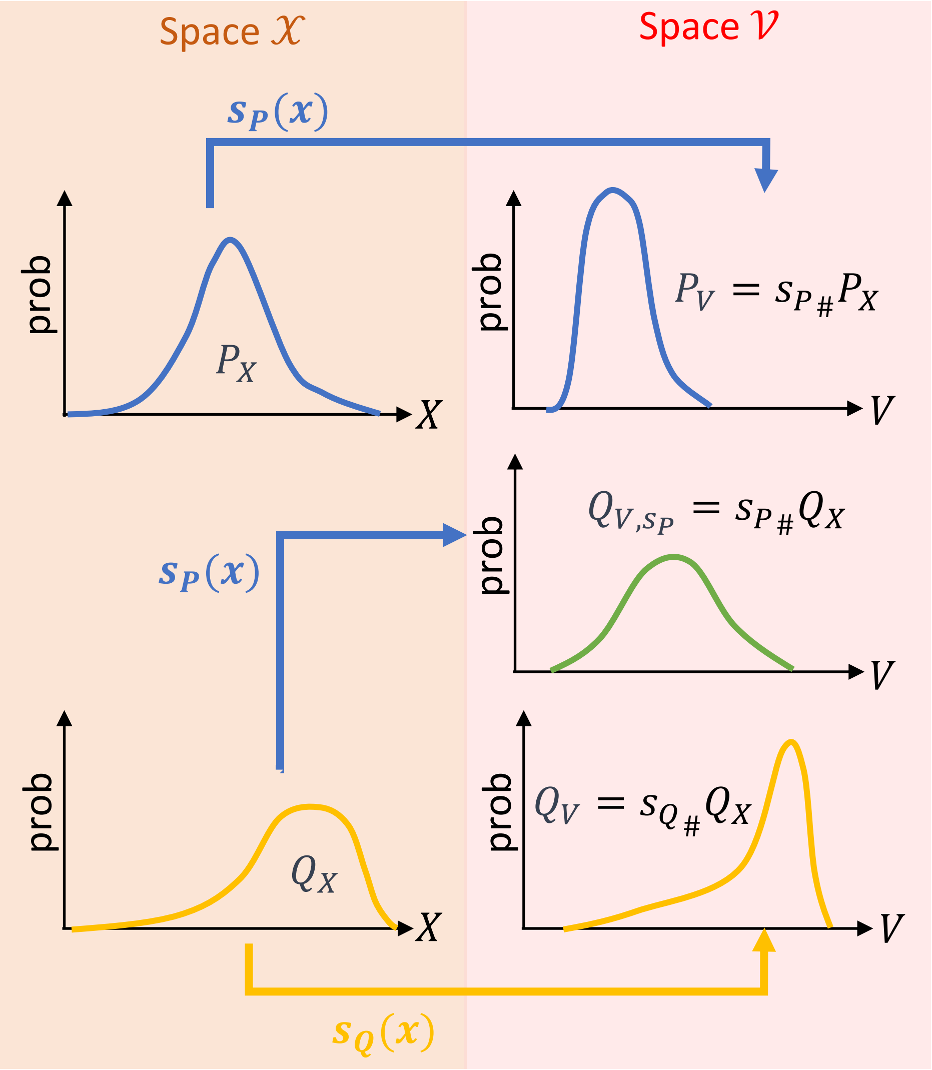
# Method: Pushforward Measure

Pushforward measure helps explain how covariate and concept shifts result in $W(P_V, Q_V)$.

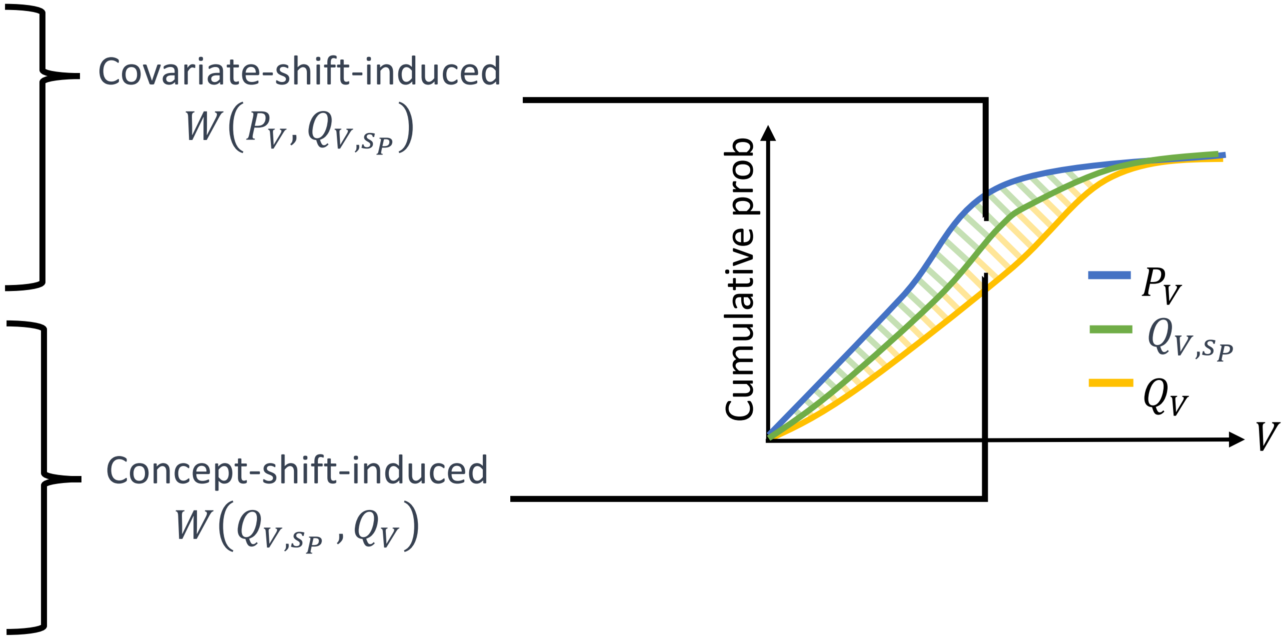
We denote $s_P(x) = |f_P(x) - h(x)|$ and $s_Q(x) = |f_Q(x) - h(x)|$.



# Method: Wasserstein Distance Decomposition

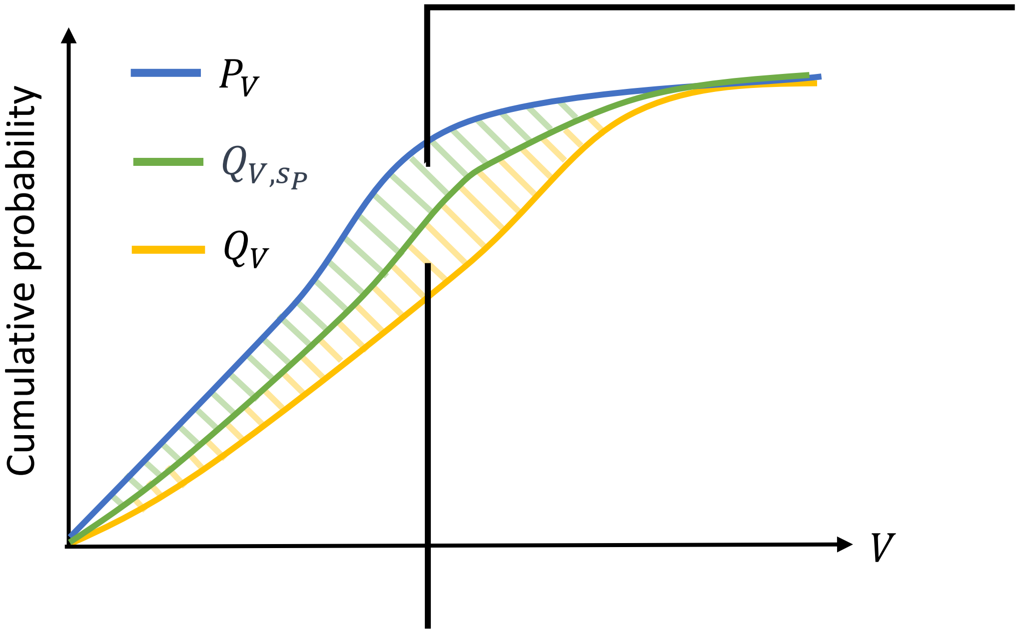



We quantify how the two components of joint distribution shift, namely covariate and concept shifts, impact $W(P_V, Q_V)$.

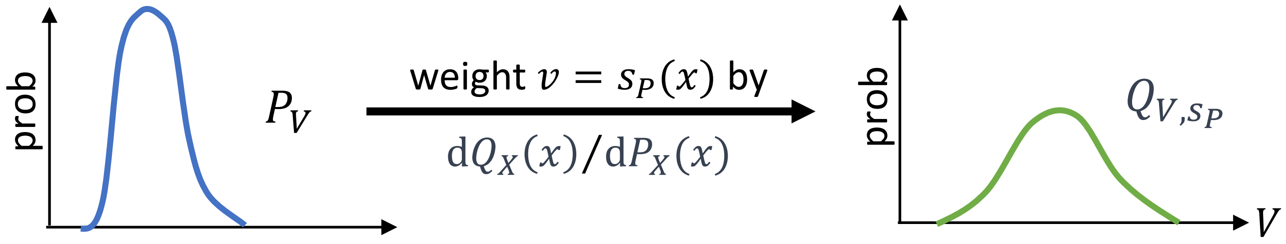



By triangular inequality,
$W(P_V, Q_V) \leq W(P_V, Q_{V,s_P}) + W(Q_{V,s_P}, Q_V)$.

# Method: Wasserstein Distance Minimization




**Covariate-shift-induced** $W\left(P_V, Q_{V,s_P}\right)$ can be minimized by weighting $P_V$ with $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$.




**Concept-shift-induced** $W\left(Q_{V,s_P}, Q_V\right)$ is minimized by Wasserstein-regularization of $\theta$-parameterized prediction model $h$.

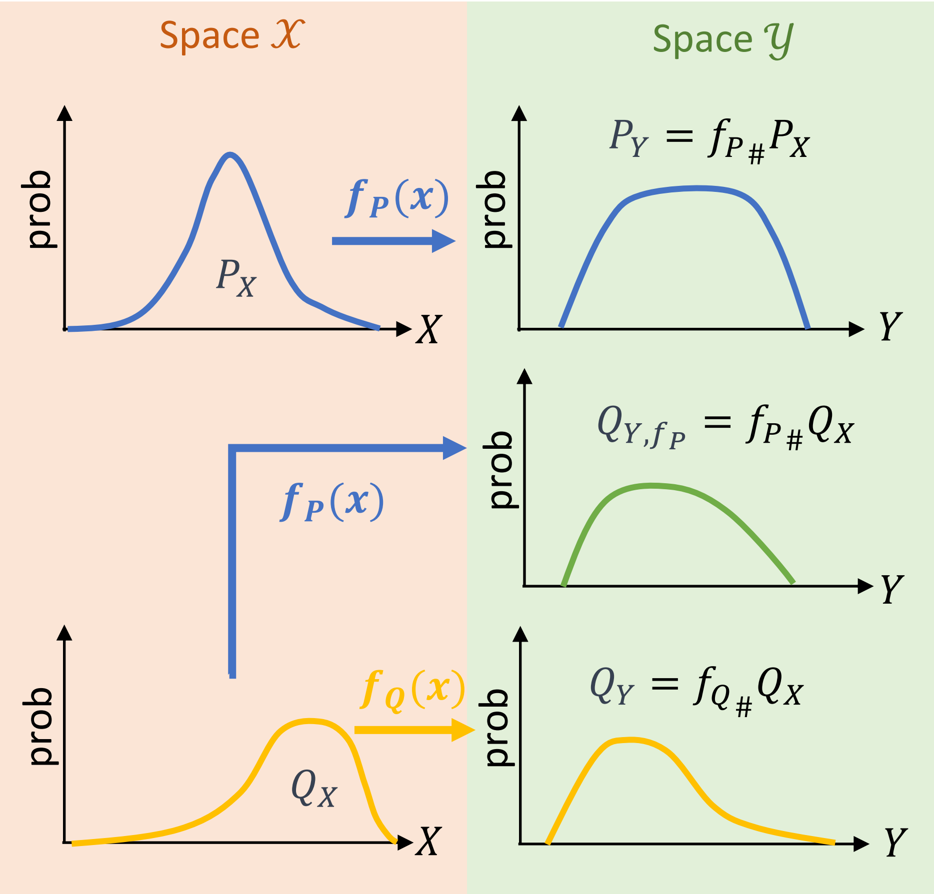
$$\min_{\theta} W\left(Q_{V,s_P}, Q_V\right)$$

# Theory: Upper-Bounding Wasserstein Distance by Covariate and Concept Shifts

$W(P_X, Q_X)$ more accurately quantifies covariate shift in $\mathcal{X}$.

**Upper bound**

Covariate-shift-induced $W(P_V, Q_{V,s_P})$ in $\mathcal{V}$




$W(Q_{Y,s_P}, Q_Y)$ is a more direct measure of concept shift in $\mathcal{Y}$.

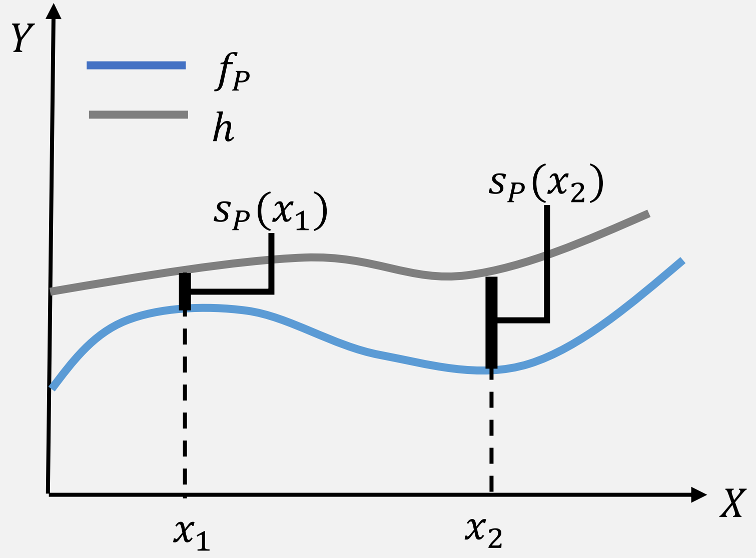
**Upper bound**

Concept-shift-induced $W(Q_{V,s_P}, Q_V)$ in $\mathcal{V}$

# Theory: Upper-Bounding Wasserstein Distance by Covariate and Concept Shifts

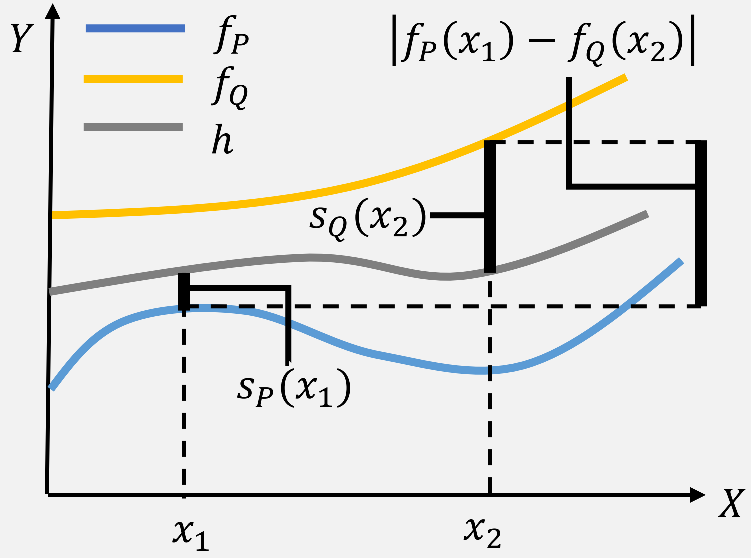
$W\left(P_V, Q_{V,s_P}\right) \leq \kappa \cdot W(P_X, Q_X)$, where $\kappa$ is the Lipschitz constant of $s_P(x) = |f_P(x) - h(x)|$.

$$\frac{|s_P(x_1) - s_P(x_2)|}{\|x_1 - x_2\|_2} \leq \kappa, \forall x_1, x_2 \in \mathcal{X}.$$



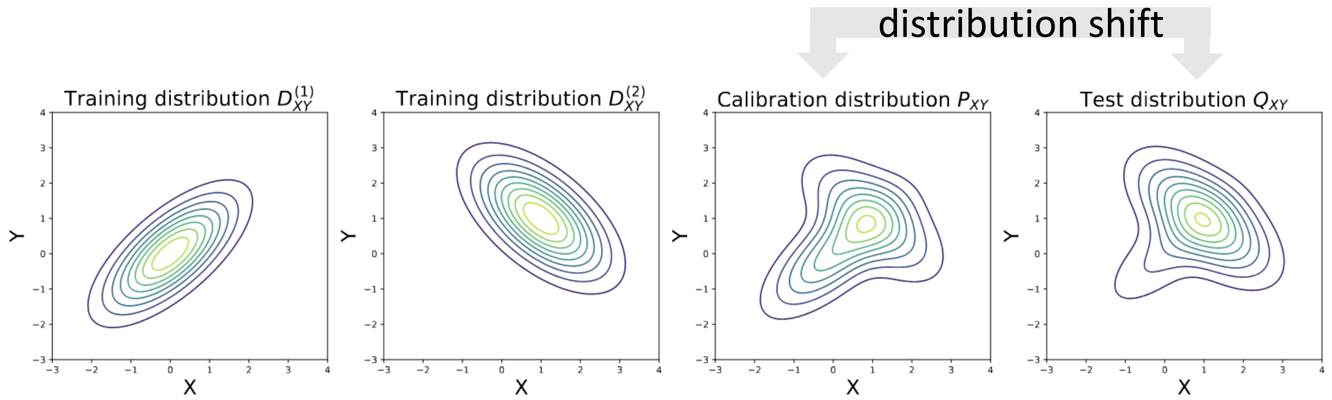

$W\left(Q_{V,s_P}, Q_V\right) \leq \eta \cdot W\left(Q_{Y,s_P}, Q_Y\right)$, where $\eta$ satisfies

$$\eta = \max \frac{|s_P(x_1) - s_Q(x_2)|}{|f_P(x_1) - f_Q(x_2)|}, \forall x_1, x_2 \in \mathcal{X}.$$




By triangular inequality, $W(P_V, Q_V) \leq \kappa \cdot W(P_X, Q_X) + \eta \cdot W\left(Q_{Y,s_P}, Q_Y\right)$.

# Algorithm: Wasserstein-regularized Conformal Prediction (WR-CP)




**Setup**: CP under multi-source generalization:

- Training distributions $D_{XY}^{(i)}$ for $i = 1, \dots, k$.
- Calibration distribution $P_{XY}$ is known.
- Test distribution $Q_{XY} \in \{\sum_{i=1}^{k} w_i \, D_{XY}^{(i)} : w_1, \dots, w_k \geq 0, \sum_{i=1}^{k} w_i = 1\}$.

**Training**:

(1) Importance weighting: obtain $D_{V,s_P}^{(i)} = s_{P\#} D_X^{(i)}$ by weighting on $P_V$ with $\frac{\mathrm{d}D_X^{(i)}}{\mathrm{d}P_X}$.

(2) Regularized optimization: $\min_{\theta} \sum_{i=1}^{k} \mathbb{E}_{(x,y)\sim D_{XY}^{(i)}} [l(h_\theta(x), y)] + \beta \sum_{i=1}^{k} W(D_{V,s_P}^{(i)}, D_V^{(i)})$.

**Inference**: Weight $P_V$ with $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ to generate prediction sets.

# Experiment: Setup

**Datasets**: (a) the airfoil self-noise dataset [8] (b) Seattle-loop[9], PeMSD4, PeMSD8 [10] for traffic speed prediction; (c) Japan-Prefectures, and U.S.-States [11] for epidemic spread forecasting.

**Baselines and proposed method:**

| Category | Methods | Functionality |
|---|---|---|
| Baseline | Vanilla CP [12] | Coverage guarantee under **i.i.d. assumption** |
| | Importance weighted CP (IW-CP) [13] | Coverage guarantee under **covariate shift** |
| | Conformalized Quantile Regression (CQR) [14] | Adaptive coverage under **i.i.d. assumption** |
| | Worst-Case CP (WC-CP) [6,7,8] | Coverage guarantee under distribution shift, but leads to large prediction sets (**inefficiency**) |
| Proposed | Wasserstein-regularized CP (WR-CP) | Robust coverage under distribution shift with small prediction sets (**efficiency**) |


[8] Brooks, Thomas, D. Pope, and Michael Marcolini. "Airfoil Self-Noise." UCI Machine Learning Repository, 1989, https://doi.org/10.24432/C5VW2C.
[9] Cui, Zhiyong, et al. "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting." *IEEE TITS* 21.11 (2019): 4883-4894.
[10] Guo, Shengnan, et al. "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.
[11] Deng, Songgaojun, et al. "Cola-GNN: Cross-location attention-based graph neural networks for long-term ILI prediction." *Proceedings of the 29th ACM CIKM*. 2020.
[12] Papadopoulos, Harris, et al. "Inductive confidence machines for regression." Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13. Springer Berlin Heidelberg, 2002.
[13] Tibshirani, Ryan J., et al. "Conformal prediction under covariate shift." *Advances in neural information processing systems* 32 (2019).
[14] Romano, Yaniv, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression." *Advances in neural information processing systems* 32 (2019).

# Experiment: Correlation between Wasserstein Distance and Coverage Gap

Compared with total variation (TV) distance, Kullback-Leibler (KL)-divergence and expectation differenec ($\Delta\mathbb{E}$), Wasserstein distance is an **effective** indicator of coverage gap with a **consistently high** Spearman's coefficient.

Spearman's coefficients between distance measures and the average coverage gap
(The highest coefficient is bold, and the second-highest coefficient is underlined.)

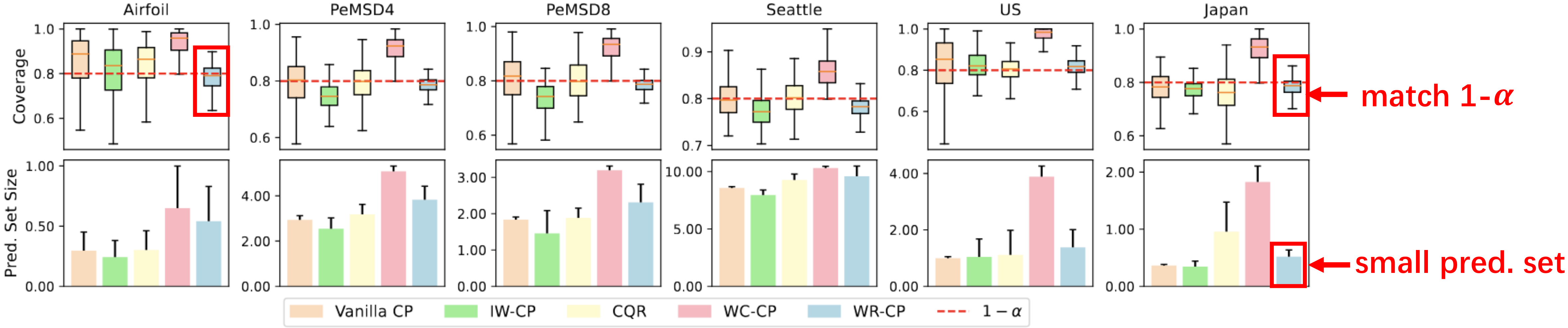
| Dataset | Airfoil | PeMSD4 | PeMSD8 | Seattle | U.S. | Japan |
|---|---|---|---|---|---|---|
| $W$ | **0.59** (0.24) | <u>0.84</u> (0.03) | **0.90** (0.03) | **0.84** (0.05) | **0.77** (0.06) | **0.57** (0.05) |
| TV | 0.45 (0.16) | **0.88** (0.03) | <u>0.86</u> (0.06) | <u>0.75</u> (0.09) | 0.67 (0.10) | 0.37 (0.06) |
| KL | 0.40 (0.21) | 0.49 (0.17) | 0.51 (0.09) | 0.45 (0.17) | 0.60 (0.11) | <u>0.53</u> (0.05) |
| $\Delta\mathbb{E}$ | <u>0.55</u> (0.19) | 0.78 (0.05) | 0.85 (0.04) | 0.71 (0.06) | <u>0.68</u> (0.08) | 0.37 (0.09) |

positive correlation

Wasserstein distance integrates the vertical gap between two cumulative probability distributions overall *all* quantiles, and is sensitive to coverage gap changes at *any* quantile.

# Experiment: Robust and Efficient Prediction Sets by WR-CP

As the Wasserstein distance between calibration and test conformal scores is minimized, WR-CP achieves robust coverage under distribution shift. As a result, WR-CP makes coverages on test data more concentrated around $1 - \alpha$ level compared to vanilla CP, IW-CP, and CQR.

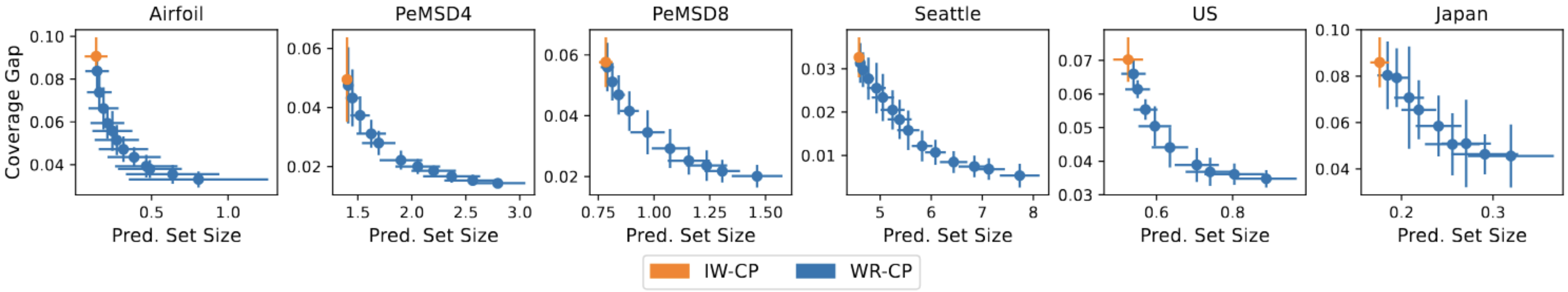



Coverages and prediction set sizes of WR-CP and baselines with $1 - \alpha = 0.8$.

While WC-CP also ensures coverage guarantees, it leads to inefficient predictions due to large set sizes, whereas WR-CP mitigates this inefficiency. Yet, due to regularization, WR-CP may lead to larger prediction sets than CP methods based on empirical risk minimization, such as vanilla CP, IW-CP, and CQR.

# Experiment: Ablation Study with Various $\beta$ Values

WR-CP effectively balances conformal prediction accuracy and efficiency, providing a flexible and customizable solution. When $\beta = 0$, WR-CP returns to IW-CP.

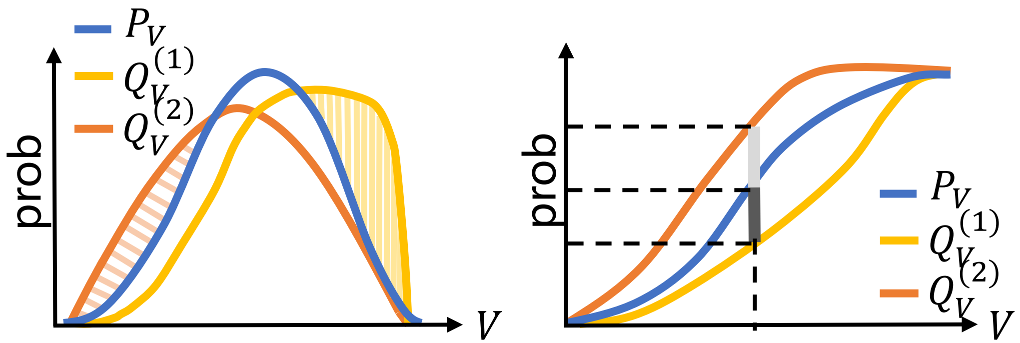



Pareto fronts of coverage gap and prediction set size obtained from WR-CP with varying $\beta$.

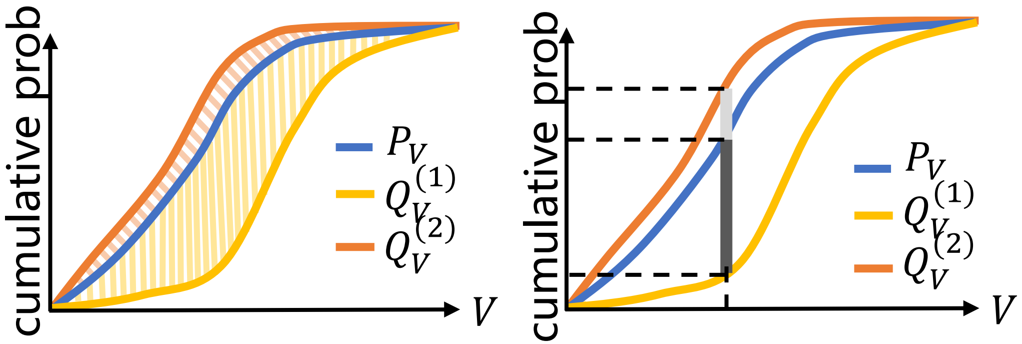
As $\beta$ increases, model $h_\theta$ prioritizes minimizing the Wasserstein distance, resulting in a smaller coverage gap. However, the regularization term inevitably impacts prediction accuracy, leading to larger prediction set sizes.

# Conclusion

- Wasserstein distance captures coverage gap and enhances interpretability of gap variations across various confidence level $1-\alpha$.



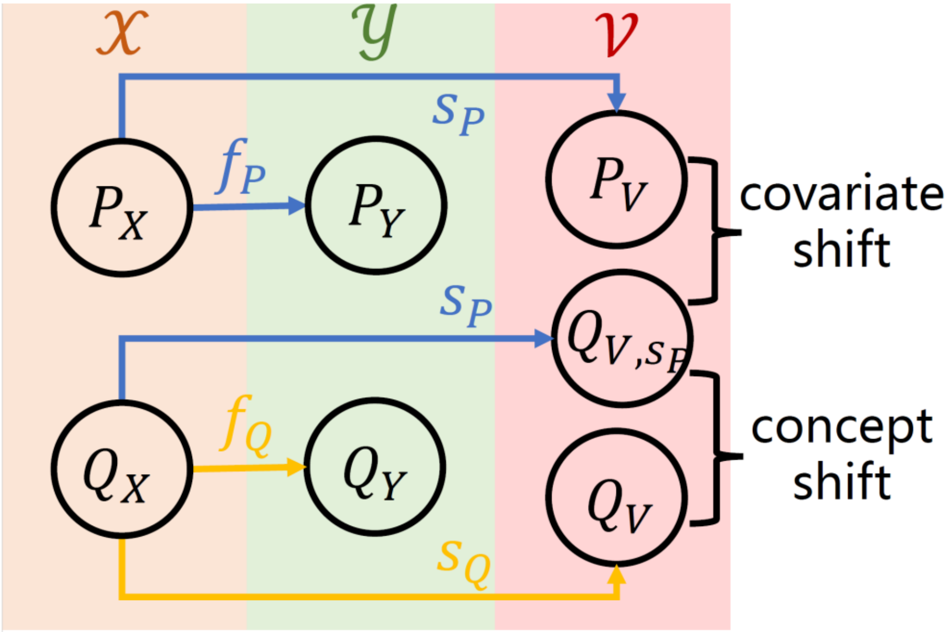

Prior: Total variation distance




Proposed: Wasserstein distance

# Conclusion

- We disentangle how covariate shift (input distribution changes) and concept shift (labeling function changes) independently impact coverage gaps.



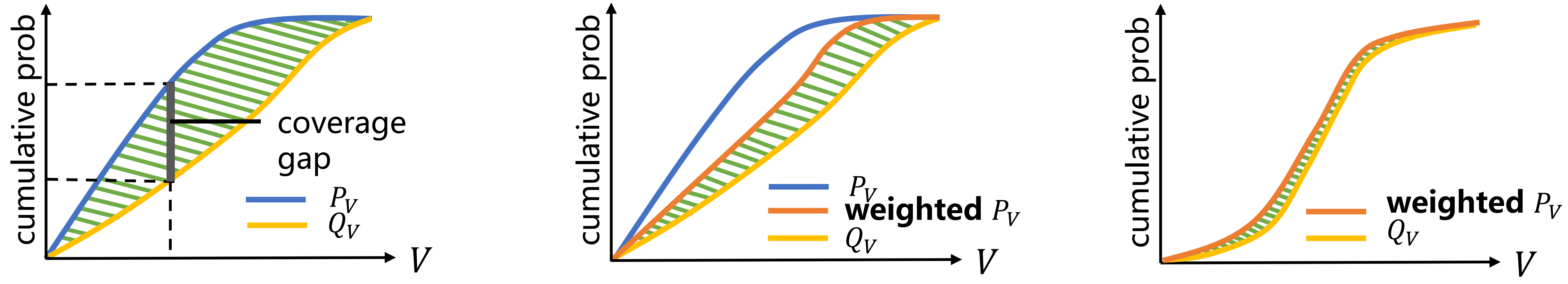
# Conclusion

- A novel method, WR-CP, combining importance weighting and representation learning regularization is proposed to optimize and balance prediction set accuracy and efficiency.

1. Initial Data Distribution and Model

2. Importance Weighting for Covariate Shift

3. Wasserstein Minimization for Concept Shift



Sihong XIE

Associate Professor

Rui XU

PhD Student

Code Repository

github.com/rxu0112/WR-CP

# Thank you

HKUST(GZ)

HIT

Lehigh