Acknowledgment

Sihong Xie was supported by the Department of Science and Technology of Guangdong Province (Grant No. 2023CX10X079), the National Key R&D Program of China (Grant No. 2023YFF0725001), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), and Education Bureau Guangzhou Municipality.

References

Salim I Amoukou and Nicolas JB Brunel. Adaptive conformal prediction by reweighting nonconformity score. *arXiv preprint arXiv:2303.12695*, 2023.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Liviu Aolaritei, Nicolas Lanzetti, Hongruyu Chen, and Florian Dörfler. Distributional uncertainty propagation via optimal transport. *arXiv preprint arXiv:2205.00343*, 2022.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Alberto Bernacchia and Simone Pigolotti. Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3):407–422, 2011.

Pope D. Brooks, Thomas and Michael Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5VW2C.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pp. 1–66, 2024.

Nicolo Colombo. Normalizing flows for conformal regression. *arXiv preprint arXiv:2406.03346*, 2024.

Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 245–254, 2020.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Bat-Sheva Einbinder, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to label noise. *arXiv preprint arXiv:2209.14295*, 2, 2022a.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 35:22380–22395, 2022b.

Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems*, 34:2060–2071, 2021.

Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Robert E Gaunt and Siqi Li. Bounding kolmogorov distances through wasserstein and related integral probability metrics. *Journal of Mathematical Analysis and Applications*, 522(1):126985, 2023.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 922–929, 2019.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction, 2023. URL `https://arxiv.org/abs/2206.13092`.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pp. 14153–14177. PMLR, 2023.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Yi Liu, Alexander W Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. *arXiv preprint arXiv:2405.09331*, 2024.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pp. 22942–22964. PMLR, 2023.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

A Tuan Nguyen, Toan Tran, Yarin Gal, Philip HS Torr, and Atılım Güneş Baydin. Kl guided domain adaptation. *arXiv preprint arXiv:2106.07780*, 2021.

Travis A O'Brien, Karthik Kashinath, Nicholas R Cavanaugh, William D Collins, and John P O'Brien. A fast and objective multidimensional kernel density estimation method: fastkde. *Computational Statistics & Data Analysis*, 101:148–160, 2016.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356. Springer, 2002.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pp. 27907–27947. PMLR, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Nathan Ross. Fundamentals of stein's method. 2011.

Hyun-Sun Ryu and Kwang Sun Ko. Sustainable development of fintech: Focused on uncertainty and perceived quality issues. *Sustainability*, 12(18):7669, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine*, pp. 107441, 2023.

Matteo Sesia, YX Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *arXiv preprint arXiv:2309.05092*, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction, 2007. URL `https://arxiv.org/abs/0706.3188`.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. *arXiv preprint arXiv:2404.19651*, 2024.

Xin Zou and Weiwei Liu. Coverage-guaranteed prediction sets for out-of-distribution data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17263–17270, 2024.

A Proofs of theorems

A.1 Proof of theorem 1

Proof. We define $f \times g$ by $f \times g(x_1, x_2) = (f(x_1), g(x_2)) = (y_1, y_2)$. Let $\mathrm{Id}_{\mathcal{X}}$ be the identity mapping function on $\mathcal{X}$, and let π_i be the mapping function to the i-th marginal. The proof follows Proposition 3 in the work by Aolaritei et al. (2022).

First, we prove the inclusion that $(f \times g)\#\Gamma(\mu, \nu) \subset \Gamma(f_{\#}\mu, g_{\#}\nu)$. Consider $\gamma \in \Gamma(\mu, \nu)$, so it is equivalent to prove that $(f \times g)_{\#}\gamma \in \Gamma(f_{\#}\mu, g_{\#}\nu)$, which means the marginals of $(f \times g)_{\#}\gamma$ are $f_{\#}\mu$ and $g_{\#}\nu$. For any continuous and bounded function $\phi : \mathcal{Y} \to \mathbb{R}$, we have

$$\begin{aligned}\int_{\mathcal{Y}\times\mathcal{Y}} \phi(y_1)\,\mathrm{d}((f \times g)_{\#}\gamma)(y_1, y_2) &= \int_{\mathcal{X}\times\mathcal{X}} \phi(f(x_1))\,\mathrm{d}\gamma(x_1, x_2) \\ &= \int_{\mathcal{X}} \phi(f(x_1))\,\mathrm{d}\mu(x_1) = \int_{\mathcal{Y}} \phi(y_1)\,\mathrm{d}(f_{\#}\mu)(y_1),\end{aligned} \tag{20}$$

so we obtain $\pi_{1\#}((f \times g)_{\#}\gamma) = f_{\#}\mu$ and similarly derive $\pi_{2\#}((f \times g)_{\#}\gamma) = g_{\#}\nu$.

Secondly, we need to prove $\Gamma(f_{\#}\mu, g_{\#}\nu) \subset (f \times g)\#\Gamma(\mu, \nu)$. With $\gamma' \in \Gamma(f_{\#}\mu, g_{\#}\nu)$, we seek $\gamma \in \Gamma(\mu.\nu)$ such that $(f \times g)_{\#}\gamma = \gamma'$. To do so, let $\gamma_{12} := (\mathrm{Id}_{\mathcal{X}} \times f)_{\#}\mu \in \Gamma(\mu, f_{\#}\mu)$, $\gamma_{23} := \gamma' \in \Gamma(f_{\#}\mu, g_{\#}\nu)$, and $\gamma_{34} := (g \times \mathrm{Id}_{\mathcal{X}})_{\#}\nu \in \Gamma(g_{\#}\nu, \nu)$. As $\pi_{2\#}\gamma_{12} = \pi_{1\#}\gamma_{23} = f_{\#}\mu$, and $\pi_{1\#}\gamma_{34} = \pi_{2\#}\gamma_{23} = g_{\#}\nu$, Santambrogio (2015) ensures a joint probability measure $\bar{\gamma}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \times \mathcal{X}$ satisfying $(\pi_1 \times \pi_2)_{\#}\bar{\gamma} = \gamma_{12}$, $(\pi_2 \times \pi_3)_{\#}\bar{\gamma} = \gamma_{23}$, and $(\pi_3 \times \pi_4)_{\#}\bar{\gamma} = \gamma_{34}$. We demonstrate that $\gamma := (\pi_1 \times \pi_4)_{\#}\bar{\gamma}$ is the probability measure we are seeking. For this, we prove $\gamma \in \Gamma(\mu, \nu)$ with any continuous and bounded function $\phi : \mathcal{X} \to \mathbb{R}$ by

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{X}} \phi(x_i)\,\mathrm{d}\gamma(x_1, x_2) &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\times\mathcal{X}} \phi(x_1)\,\mathrm{d}\bar{\gamma}(x_1, y_1, y_2, x_2) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \phi(x_1)\,\mathrm{d}\gamma_{12}(x_1, y_1) = \int_{\mathcal{X}} \phi(x_1)\,\mathrm{d}\mu(x_1)\,.\end{aligned} \tag{21}$$

Eq. (21) indicates $\pi_{1\#}\gamma = \mu$. Similarly, we can derive $\pi_{2\#}\gamma = \nu$. As a result, we can prove $(f \times g)_{\#}\gamma = \gamma'$ with any continuous and bounded function $\phi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ by

$$\begin{aligned}&\int_{\mathcal{Y}\times\mathcal{Y}} \phi(y_1, y_2)\,\mathrm{d}((f \times g)_{\#}\gamma)(x_1, x_2) \\ &= \int_{\mathcal{X}\times\mathcal{X}} \phi(f(x_1), g(x_2))\,\mathrm{d}\gamma(x_1, x_2) \\ &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\times\mathcal{X}} \phi(f(x_1), g(x_2))\,\mathrm{d}\bar{\gamma}(x_1, y_1, y_2, x_2) \\ &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\times\mathcal{X}} \phi(y_1, y_2)\,\mathrm{d}\bar{\gamma}(x_1, y_1, y_2, x_2) \\ &= \int_{\mathcal{Y}\times\mathcal{Y}} \phi(y_1, y_2)\,\mathrm{d}\gamma_{23}(y_1, y_2) = \int_{\mathcal{Y}\times\mathcal{Y}} \phi(y_1, y_2)\,\mathrm{d}\gamma'(y_1, y_2).\end{aligned} \tag{22}$$

As $(f \times g)\#\Gamma(\mu, \nu) \subset \Gamma(f_{\#}\mu, g_{\#}\nu)$ and $\Gamma(f_{\#}\mu, g_{\#}\nu) \subset (f \times g)\#\Gamma(\mu, \nu)$, we obtain $(f \times g)\#\Gamma(\mu, \nu) = \Gamma(f_{\#}\mu, g_{\#}\nu)$. Finally, we prove Theorem 1 by

$$\begin{aligned}W(\mu_f, \nu_g) &= W(f_{\#}\mu, g_{\#}\nu) \\ &= \inf_{\gamma' \in \Gamma(f_{\#}\mu, g_{\#}\nu)} c_{\mathcal{Y}}(y_1, y_2)\,\mathrm{d}\gamma'(y_1, y_2) \\ &= \inf_{\gamma' \in (f \times g)_{\#}\Gamma(\mu, \nu)} \int_{\mathcal{Y}\times\mathcal{Y}} c_{\mathcal{Y}}(y_1, y_2)\,\mathrm{d}\gamma'(y_1, y_2) \\ &= \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{Y}\times\mathcal{Y}} c_{\mathcal{Y}}(y_1, y_2)\,\mathrm{d}((f \times g)_{\#}\gamma)(y_1, y_2) \\ &= \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{Y}\times\mathcal{Y}} c_{\mathcal{Y}}(f(x_1), g(x_2))\,\mathrm{d}\gamma(y_1, y_2)\end{aligned} \tag{23}$$

□

A.2 Proof of theorem 2

Proof. Let $\gamma' \in \Gamma(\mu_f, \nu_f)$ be the pushforward of $\gamma \in \Gamma(\mu, \nu)$ via function $f \times f$. We can apply Theorem 1 to $W(\mu_f, \nu_f)$ and obtain

$$W(\mu_f, \nu_f) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{Y}}(f(x_1), f(x_2))\, \mathrm{d}\gamma(x_1, x_2)\,. \tag{24}$$

If the optimal transport plan for $W(\mu, \nu)$ is γ^*, and κ bounds the Lipschitz continuity of f, we have

$$\begin{aligned} W(\mu_f, \nu_f) &\leq \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{Y}}(f(x_1), f(x_2))\, \mathrm{d}\gamma^*(x_1, x_2) \\ &\leq \int_{\mathcal{X}\times\mathcal{X}} \kappa c_{\mathcal{X}}(x_1, x_2)\, \mathrm{d}\gamma^*(x_1, x_2) = \kappa W(\mu, \nu). \end{aligned} \tag{25}$$

In Eq. (25), the first inequality holds because γ^* may not be the optimal transport plan for $W(\mu_f, \nu_f)$, and the second inequality holds due to the definition of κ. □

A.3 Proof of theorem 3

Proof. As Wasserstein distance satisfies triangle inequality, $W(\mu, \nu)$ and $W(\hat{\mu}_n, \hat{\nu}_m)$ follow

$$W(\mu, \nu) \leq W(\hat{\mu}_n, \mu) + W(\hat{\mu}_n, \nu) \leq W(\hat{\mu}_n, \mu) + W(\hat{\mu}_n, \hat{\nu}_m) + W(\hat{\nu}_m, \nu). \tag{26}$$

Given $\mathbb{E}[W(\mu, \hat{\mu}_n)] \leq \lambda_\mu n^{-1/\sigma_\mu}$ and $\mathbb{E}[W(\nu, \hat{\nu}_m)] \leq \lambda_\nu m^{-1/\sigma_\nu}$ from Proposition 2, with probabilities at least $1 - e^{-2nt_\mu{}^2}$ and $1 - e^{-2mt_\nu{}^2}$, respectively, we have

$$W(\mu, \hat{\mu}_n) \leq \lambda_\mu n^{-1/\sigma_\mu} + t_\mu,\ W(\nu, \hat{\nu}_m) \leq \lambda_\nu m^{-1/\sigma_\nu} + t_\nu. \tag{27}$$

It is reasonable to assume the two events in Eq. (27) are independent, so we can apply them to Eq. (26), and thus obtain Eq. (15) with probability at least $(1 - e^{-2nt_\mu{}^2})(1 - e^{-2mt_\nu{}^2})$. □

A.4 Proof of theorem 4

Proof. We denote F_μ, F_ν, and $F_{\nu^{(i)}}$ the corresponding CDFs of μ, ν, and $\nu^{(i)}$ for $i = 1, ..., k$.

When two distributions are on the real number set $\mathbb{R}$ with Euclidean distance, W of the two distributions equals the area between their CDFs. Therefore, the 1-Wasserstein distance between μ and ν is given by

$$W(\mu, \nu) = \int_{\mathcal{X}} |F_\mu(x) - F_\nu(x)|\, \mathrm{d}x\,. \tag{28}$$

Since $\nu = \sum_{i=1}^k w_i \nu^{(i)}$, we have $F_\nu(x) = \sum_{i=1}^k w_i F_{\nu^{(i)}}(x)$. As ν, $\nu^{(i)}$, and μ are definded on $\mathcal{X} \subseteq \mathbb{R}$, we can derive

$$\begin{aligned} W(\mu, \nu) &= \int_{\mathcal{X}} |F_\mu(x) - F_\nu(x)|\, \mathrm{d}x = \int_{\mathcal{X}} \left| F_\mu(x) - \sum_{i=1}^k w_i F_{\nu^{(i)}}(x) \right| \mathrm{d}x \\ &= \int_{\mathcal{X}} \left| \sum_{i=1}^k w_i F_\mu(x) - \sum_{i=1}^k w_i F_{\nu^{(i)}}(x) \right| \mathrm{d}x = \int_{\mathcal{X}} \left| \sum_{i=1}^k w_i \left(F_\mu(x) - F_{\nu^{(i)}}(x)\right) \right| \mathrm{d}x \\ &\leq \int_{\mathcal{X}} \sum_{i=1}^k w_i \left|F_\mu(x) - F_{\nu^{(i)}}(x)\right| \mathrm{d}x = \sum_{i=1}^k w_i \int_{\mathcal{X}} \left|F_\mu(x) - F_{\nu^{(i)}}(x)\right| \mathrm{d}x \\ &= \sum_{i=1}^k w_i W(\mu, \nu^{(i)}). \end{aligned} \tag{29}$$

□

B Comparison between total variation and Wasserstein distance

The total variation (TV) distance between two univariate distributions is defined as half of the absolute area between their probability density functions (PDFs). For instance, given two distributions μ and ν with PDFs p_μ and p_ν, respectively, on space $\mathbb{R}_{\geq 0}$, the TV distance is given by

$$TV(\mu,\nu) = \frac{1}{2}\int_{\mathbb{R}_{\geq 0}} |p_\mu(x) - p_\nu(x)|\,\mathrm{d}x\,. \tag{30}$$

In contrast, we expand $W(\mu,\nu)$ according to Eq. (28) by

$$\begin{aligned} W(\mu,\nu) &= \int_{\mathbb{R}_{\geq 0}} |F_\mu(x) - F_\nu(x)|\,\mathrm{d}x = \int_{\mathbb{R}_{\geq 0}} \left|\int_0^x p_\mu(t)\,\mathrm{d}t - \int_0^x p_\nu(t)\,\mathrm{d}t\right|\mathrm{d}x \\ &= \int_{\mathbb{R}_{\geq 0}} \left|\int_0^x p_\mu(t) - p_\nu(t)\,\mathrm{d}t\right|\mathrm{d}x\,. \end{aligned} \tag{31}$$

The inner integration between 0 and x indicates Wasserstein distance cares where two distributions μ and ν differ, whereas the total variation distance in Eq. (30) does not take this into consideration.

We would like to introduce a toy example to illustrate further why total variation distance can not consistently capture the closeness between two cumulative distribution functions (CDFs). Consider three conformal score distributions $P_V, Q_V^{(1)}, Q_V^{(2)}$ on space $\mathbb{R}_{\geq 0}$ with their PDFs:

$$p_{P_V}(v) = 1, v \in [0,1];$$

$$p_{Q_V^{(1)}}(v) = \begin{cases} 1 & \text{if } v \in [0, 0.9], \\ 2 & \text{if } v \in (0.9, 0.95]; \end{cases}$$

$$p_{Q_V^{(2)}}(v) = \begin{cases} 2 & \text{if } v \in [0, 0.04], \\ 1 & \text{if } v \in (0.04, 0.96]. \end{cases}$$

Therefore, we calculate $TV(P_V, Q_V^{(1)}) = 0.05$ and $TV(P_V, Q_V^{(2)}) = 0.04$, while $W(P_V, Q_V^{(1)}) = 0.0025$ and $W(P_V, Q_V^{(2)}) = 0.0384$. In this example, a reduction in total variation distance results in a larger Wasserstein distance between two CDFs. Intuitively, TVD only measures the overall difference between two distributions without accounting for the specific locations where they diverge. In contrast, the Wasserstein distance will be high when divergence occurs early (i.e., at a small quantile), especially if the discrepancy persists until the "lagging" CDF catches up. We visualize the example in Figure 6.

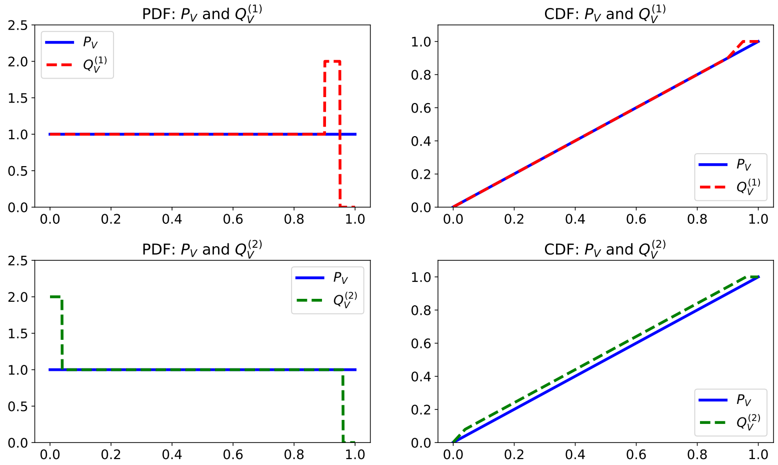

Figure 6: **Comparison between total variation distance and Wasserstein distance**: a reduction in the total variation distance does not necessarily result in CDFs becoming closer.

C Rationale for differentiating covariate and concept shifts

There are two key reasons to differentiate between covariate and concept shifts. First, making this distinction enables the application of importance weighting. Minimizing the Wasserstein regularization term inevitably increases prediction residuals. By applying importance weighting, we expect to reduce the distance, mitigating the adverse effects of regularization on optimizing the regression loss function in Eq. (19). Figure 3 shows this expectation is met on five out of the six datasets. This occurs because, in most cases, covariate shifts exacerbate the distance caused by concept shifts ($f_P \neq f_Q$). Consequently, importance weighting effectively reduces this distance, as illustrated in Figure 7(a) and evidenced by the results for the airfoil self-noise, PeMSD4, PeMSD8, U.S.-States, and Japan-Prefectures datasets in Figure 3. However, there are instances where covariate shifts can alleviate the Wasserstein distance induced by concept shifts. In such cases, applying importance weighting may increase the distance, as demonstrated in the results for the Seattle-loop dataset in Figure 3. This phenomenon is further illustrated in Figure 7(b).

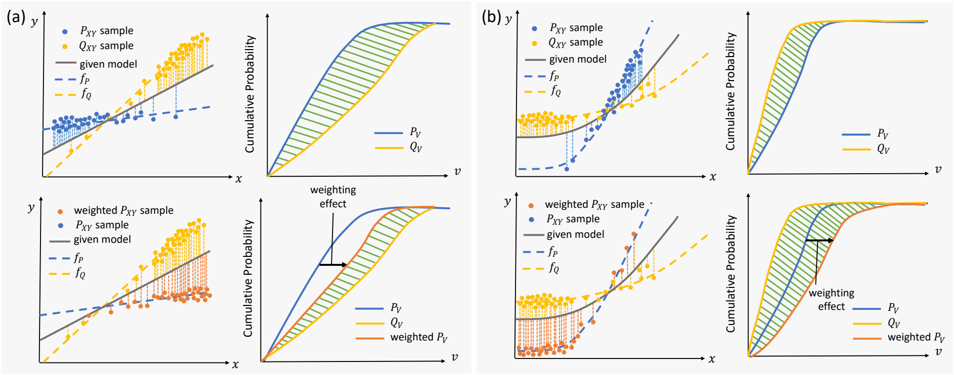

Figure 7: **Effect of importance weighting on Wasserstein distance:** (a) Scenario where importance weighting reduces Wasserstein distance; (b) Scenario where importance weighting enlarges Wasserstein distance.

Secondly, in multi-source CP, different training distributions $D_{XY}^{(i)}$ can suffer from different degrees of covariate and concept shifts. Importance weighting allows the regularized loss in Eq. (19) to minimize the distance between training conformal score distribution $D_V^{(i)}$ and its correspondingly weighted calibration conformal score distribution $D_{V,s_P}^{(i)}$, so the model can be more targeted on those whose remaining Wasserstein distances are large. Also, since various non-exchangeable test distributions will weight calibration conformal score distribution differently in the inference phase, prediction set sizes can be adaptive to different test distributions. In contrast, without importance weighting, the model can only regularize $\sum_{i=1}^{k} W(P_V, D_V^{(i)})$, and use the same quantile of P_V to generate prediction sets for samples from all test distributions, resulting in the same prediction set size and lack of adaptiveness.

To further demonstrate the two reasons we mentioned above, we modify Wasserstein-regularization based on unweighted calibration conformal scores (i.e. $\sum_{i=1}^{k} W(P_V, D_V^{(i)})$) during training. Also, the weighting operation in the prediction phase in Algorithm 1 is removed accordingly. This method is denoted as WR-CP(uw). We performed WR-CP(uw) on the sampled data from the 10 trials of each dataset at $\alpha = 0.2$ and compared its results with those of WR-CP.

The comparison is depicted in Figure 8. Although the average coverage gaps between WR-CP and WR-CP(uw) are quite similar, at 3.1% and 2.3% respectively, the average prediction set size for WR-CP is 28.0% smaller than that of WR-CP(uw). This observation proves our first reason that importance weighting effectively reduces the Wasserstein distance between calibration and test conformal scores. By doing so, it mitigates the side effect of optimizing the regularized objective function in Eq. (19), which increases prediction residuals. Since larger residuals result in larger prediction sets, reducing residuals directly helps minimize prediction set size. Additionally, the standard deviations of the prediction set sizes observed in WR-CP(uw) are typically smaller than those found in WR-CP. This proves the second reason that removing importance weighting will make prediction sets less adaptive to different test distributions.

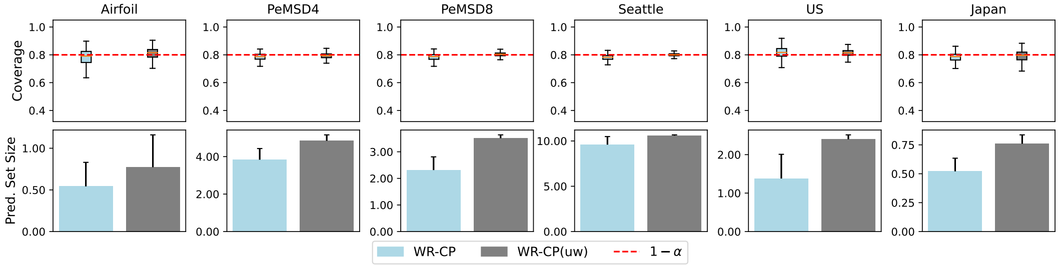

Figure 8: **Comparison between WR-CP and WR-CP(uw) at** $\alpha = 0.2$**.** Both methods were implemented using the same β values of 4.5, 9, 9, 6, 8, and 20 across the datasets.

D Geometric intuition of η

To provide a geometric intuition of η, we expand the definition of η as

$$\begin{aligned}\eta &= \max_{x_1,x_2\in\mathcal{X}} \frac{|s_P(x_1) - s_Q(x_2)|}{|f_P(x_1) - f_Q(x_2)|} \\ &= \max_{x_1,x_2\in\mathcal{X}} \frac{|s\left(x_1, f_P(x_1)\right) - s\left(x_2, f_Q(x_2)\right)|}{|f_P(x_1) - f_Q(x_2)|} \\ &= \max_{x_1,x_2\in\mathcal{X}} \frac{||h(x_1) - f_P(x_1)| - |h(x_2) - f_Q(x_2)||}{|f_P(x_1) - f_Q(x_2)|}.\end{aligned} \tag{32}$$

We first simplify the definition by assuming $x_1 = x_2$, so the denominator is the absolute difference between two ground-truth mapping functions f_P and f_Q at x_1, and the numerator is the absolute difference of the residuals of f_P and f_Q with a given model h at x_1. η is the largest ratio between the two absolute differences. A small η means even if f_P and f_Q differ significantly, h results in similar prediction residuals on f_P and f_Q. When $x_1 \neq x_2$, η is the largest ratio of the two absolute differences at two positions, x_1 and x_2, so a small η means that h can lead to similar residuals when $f_P(x_1)$ and $f_Q(x_2)$ differ. The expanded definition above includes both $x_1 = x_2$ and $x_1 \neq x_2$ conditions and Figure 9 (a) and (b) present the two conditions, respectively. Intuitively, the residual difference caused by concept shift will be constrained by η.

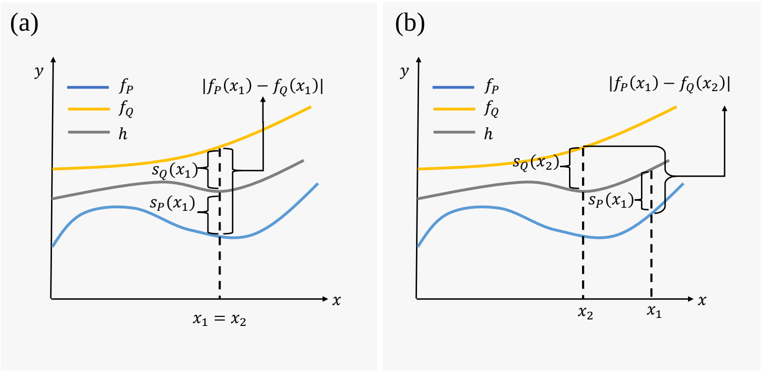

Figure 9: **Geometric intuition of** η **when (a)** $x_1 = x_2$ **and (b)** $x_1 \neq x_2$: Intuitively, the residual difference caused by concept shift will be constrained by η.

E Distribution estimation

E.1 Kernel density estimation

$\hat{P}_X$ and $\hat{D}_X^{(i)}$ for $i = 1, ..., k$ are obtained by kernel density estimation (KDE), and based on the estimated distributions we calculate the likelihood ratio.

In our experiments, we applied the Gaussian kernel, which is a positive function of $x \in \mathcal{X} \subseteq \mathbb{R}^d$ given by

$$\mathrm{K}(x, b) = \frac{1}{(\sqrt{2\pi}b)^d} e^{-\frac{\|x\|^2}{2b^2}}, \tag{33}$$

where $\|\cdot\|$ is Euclidean distance and b is bandwidth. Given this kernel form, the estimated probability density, denoted by $\hat{p}$, at a position x_a within a group of points $x_1, ..., x_n$ is

$$\hat{p}(x_a, \mathrm{K}) = \sum\nolimits_{i=1}^{n} \mathrm{K}(x_a - x_i, b). \tag{34}$$

To find the optimized bandwidth value of $\hat{P}_X$ and $\hat{D}_X^{(i)}$ for $i = 1, ..., k$ on each dataset, we applied the grid search method with a bandwidth pool using scikit-learn package (Pedregosa et al., 2011). With the approximated marginal distribution densities, we can calculate the likelihood ratio to implement the weighting technique proposed by Tibshirani et al. (2019).

E.2 Point-wise distribution estimation

$\hat{D}_V^{(i)}$ and $\hat{D}_{V,s_P}^{(i)}$ for $i = 1, ..., k$ are estimated as discontinuous, point-wise distributions to ensure differentiability during training. Specifically, as $\hat{D}_V^{(i)}$ and $\hat{D}_{V,s_P}^{(i)}$ are conformal score distributions on real number set $\mathbb{R}$, $W(\hat{D}_V^{(i)}, \hat{D}_{V,s_P}^{(i)})$ is equal to area between their CDFs, as Eq. (28) shows. Hence, our focus is on estimating the CDFs of $\hat{D}_V^{(i)}$ and $\hat{D}_{V,s_P}^{(i)}$ for $i = 1, ..., k$.

For the details of point-wise distribution estimation, consider we have a $x_1, ..., x_n$ drawn from a probability measure μ in space $\mathcal{X} \subseteq \mathbb{R}$, so the approximated CDF of μ is given by

$$F_{\hat{\mu}}(x) = \frac{1}{n} \sum\nolimits_{j=1}^{n} \delta_{x_i} \mathbb{1}_{x_i < x}, \tag{35}$$

where $\mathbb{1}$ is the indicator function and δ_{x_i} represents the point mass at x_i (i.e., the distribution placing all mass at the value x_i). In other words, Eq. (35) counts the partition of samples that are smaller than x. This point-wise estimation ensures that the Wasserstein-1 distance between the estimated distributions is differentiable.

F Supplementary experimental insights

F.1 Datasets

The airfoil self-noise dataset from the UCI Machine Learning Repository (Brooks & Marcolini, 2014) was intentionally modified to introduce covariate shift and concept shift among them. It includes 1503 instances. The target variable is the scaled sound pressure level of NASA airfoils, and there are 5 features: log frequency, angle of attack, chord length, free-stream velocity, and log displacement thickness of the suction side. To introduce covariate shift, we divided the original dataset into three subsets based on the 33% and 66% quantiles of the first dimension feature, log frequency, and partially shuffled them. Therefore, $k = 3$ for this dataset. We further introduced concept shifts among the three subsets by modifying target values. With ξ following a normal distribution $N(0, 10)$, for y in the first set, $y+ = y/1000 * \xi$; for y in the second set, $y+ = y/\xi$; for y in the third set, $y+ = \xi$. With the modified data, we conducted sampling trials to generate 10 randomly sampled datasets.

The Seattle-loop dataset Cui et al. (2019), as well as the PeMSD4 and PeMSED8 datasets Guo et al. (2019), consist of sensor-observed traffic volume and speed data gathered in Seattle, San Francisco, and San Bernardino, respectively. The data was collected at 5-minute intervals. Our goal for each dataset is to forecast the traffic speed of a specific interested local road segment in the next time step by utilizing the current traffic speed and volume data from both the local segment and its neighboring segments. Before sampling, we selected 10 segments of interest for each dataset randomly, setting $k = 10$ for them. There are natural joint distribution shifts present among these segments because of the varying local traffic patterns.

The U.S.-States and Japan-Prefectures datasets Deng et al. (2020) contain data on the number of patients infected with influenza-like illness (ILI) reported by the U.S. Department of Health and Human Services, Center for Disease Control and Prevention (CDC), and the Japan Infectious Diseases Weekly Report, respectively. The data in each dataset is structured based on the collection region. Our objective is to utilize the regional predictive features, including population, the increase in the number of infected patients observed in the current week, and the annual cumulative total of infections, to forecast the rise in infections for the following week in the corresponding region. We also randomly selected 10 regions for both datasets, so $k = 10$. Due to the diverse regional epidemiological conditions, there are inherent joint distribution shifts among these regions.

For each dataset, we began by sampling $\mathcal{S}_{XY}^{(i)}$ from each subset i, for $i = 1, ..., k$, without replacement. After this step, we allocated the remaining elements within each subset for calibration and testing purposes. The parts intended for calibration across all subsets were then unified to form $\mathcal{S}_{XY}^{P}$. Lastly, to create diverse testing scenarios, we generated multiple test sets by randomly mixing the parts designated for testing from each subset with replacement. For each dataset, we conducted the sampling trial for 10 times, and calculated the mean and standard deviation of the results from these trials, as shown in Figure 3, Figure 4, and Figure 5. For efficiency, all CP methods were conducted as split conformal prediction.

We introduce a toy example to further illustrate that exchangeability does not hold. Consider we have two training distributions:

$$D_{XY}^{(1)} = N\left([0,0], \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right); D_{XY}^{(2)} = N\left([1,1], \begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}\right).$$

A calibration distribution is a mixture of these two training distributions with known weights, such as a uniformly weighted mixture ($w_1 = w_2 = 0.5$). A test distribution is a mixture of $D_{XY}^{(1)}$ and $D_{XY}^{(2)}$ with unknown random weights. To visualize the non-exchangeability in Figure 10, we assume the unknown test distribution has weights of 0.2 for $D_{XY}^{(1)}$ and 0.8 for $D_{XY}^{(2)}$.

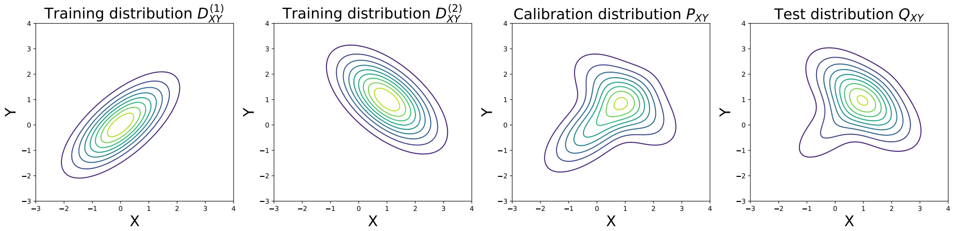

Figure 10: **Calibration and test samples are not exchangeable as they are from different distributions.**

F.2 Spearman's coefficient

We first provide the definition of Pearson coefficient.

Definition 5 (Pearson coefficient)**.** *With n pairs of samples, (x_i, y_i) for $i = 1, ..., n$, of two random variables X and Y, Pearson coefficient, r_p, is calculated as the covariance of the samples divided by the product of their standard deviations. Formally, it is given by*

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{36}$$

where $\overline{x}$ and $\overline{y}$ are the means of the samples of X and Y, respectively.

Based on Pearson coefficient, the definition of Spearman's coefficient is given as follows.

Definition 6 (Spearman's coefficient)**.** *With n pairs of samples, (x_i, y_i) for $i = 1, ..., n$, of two random variables X and Y, letting $r(\cdot)$ be the rank function (i.e., $r(x_1) = 3$ indicates that x_1 is the third largest sample among $x_1, ..., x_n$), Spearman's coefficient, r_s, is defined as the Pearson coefficient between the ranked samples:*

$$r_s = \frac{\sum_{i=1}^{n} \left(r(x_i) - r(\overline{x})\right)\left(r(y_i) - r(\overline{y})\right)}{\sqrt{\sum_{i=1}^{n} \left(r(x_i) - r(\overline{x})\right)^2}\sqrt{\sum_{i=1}^{n} \left(r(y_i) - r(\overline{y})\right)^2}}, \tag{37}$$

where $\overline{x}$ and $\overline{y}$ are the means of the samples of X and Y, respectively.

We calculated Spearman's coefficient between each distance measure and the largest coverage gap in Section 6 to confirm that Wasserstein distance holds the strongest positive correlation compared with other distance measures.

F.3 Additional experiment results of subsection 6.4

In addition to the results shown in Figure 4, we present further experimental findings from Subsection 6.4 with α values of 0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 on Figure 11, 12, 13, 14, 15, 16, 17, and 18, respectively. Clearly, WR-CP demonstrates the ability to generate more tightly concentrated coverages near $1 - \alpha$ compared to vanilla CP and IW-CP. Additionally, it yields smaller prediction set sizes than the state-of-the-art method WC-CP. These figures also reveal a trend where as the α value increases, WR-CP requires a smaller β to achieve acceptable coverages around $1 - \alpha$, so the prediction set sizes produced by WR-CP are closer with those of vanilla CP and IW-CP, as evidenced by the results on the PeMSD4 in Figure 11 and Figure 18. This phenomenon could be attributed to the trade-off between conformal prediction accuracy and efficiency under joint distribution shift. The Wasserstein regularization term in Eq. (19) tends to prioritize aligning smaller conformal scores initially, as it reduces the Wasserstein penalty with a lesser increase in the empirical risk minimization term. Hence, as the hyperparameter β increases, the model gradually aligns larger conformal scores from two different distributions, which will adversely impact the risk-driven term more. When considering a higher α value, the focus is on ensuring that the coverages on test data are close to the smaller $1 - \alpha$, indicating the importance of aligning small conformal scores. Consequently, a high β value is not necessary in this case, leading to smaller prediction set sizes being achieved.

F.4 Experiment setups in ablation study

To visualize a comprehensive and evenly-distributed set of optimal solutions on Pareto fronts, we utilized WR-CP with varying values of β to produce the results depicted in Figure 5. As mentioned in Section 5, it is worth noting that when $\beta = 0$, WR-CP reverts to IW-CP. The selected β values for the results of Figure 5 are shown in Table 2.

Table 2: β values of WR-CP in ablation study

Dataset	β values
Airfoil	1, 1.5, 2, 2.5, 3, 3.5, 4.5, 6, 8, 9, 13, 20.
PeMSD4	1, 1.5, 2, 2.5, 3, 5, 7, 9, 11, 15, 20.
PeMSD8	1, 1.5, 2, 2.5, 3, 4, 5, 7, 9, 17.
Seattle	1, 2, 3, 4, 4.5, 5, 5.5, 6, 7, 8, 10, 13, 15, 20.
U.S.	1, 1.5, 2, 2.5, 3, 5, 6, 8, 13.
Japan	1, 2, 3, 4, 6, 8, 10, 13, 20.

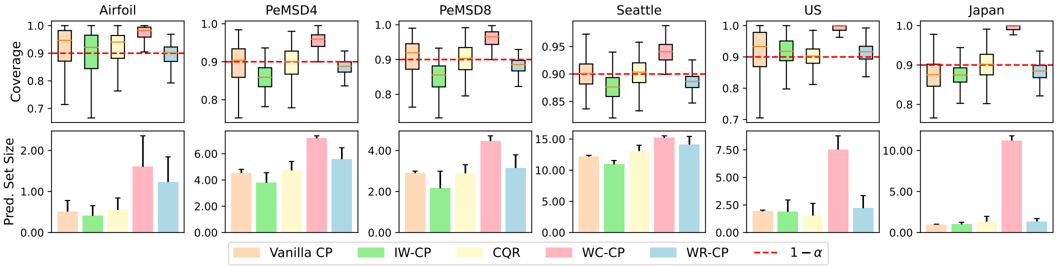

Figure 11: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.1$:** The β values for the WR-CP method are 9, 11, 9, 8, 13, and 20, respectively.

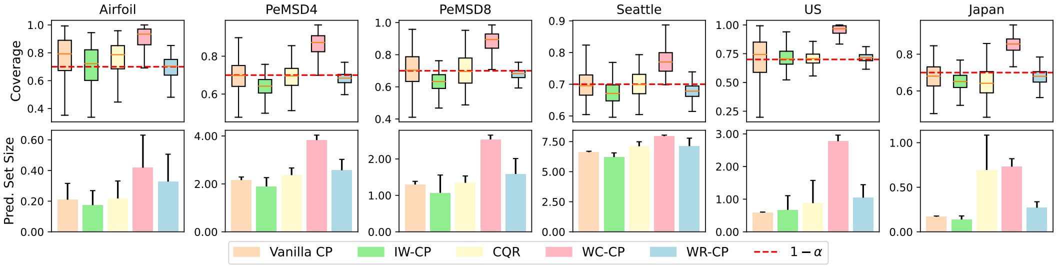

Figure 12: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.3$:** The β values for the WR-CP method are 3, 5, 5, 5, 8, and 13, respectively.

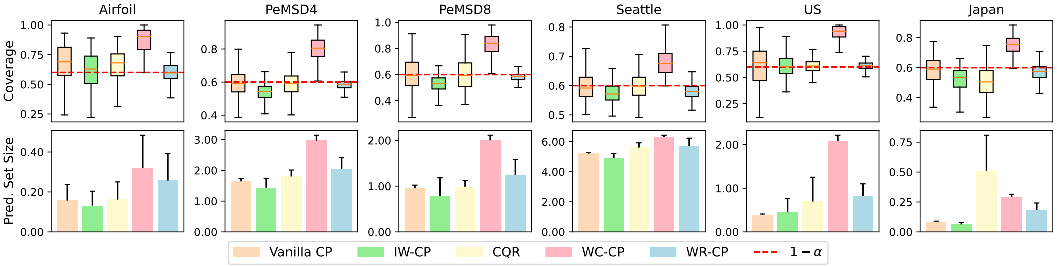

Figure 13: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.4$:** The β values for the WR-CP method are 3, 5, 5, 5, 8, and 13, respectively.

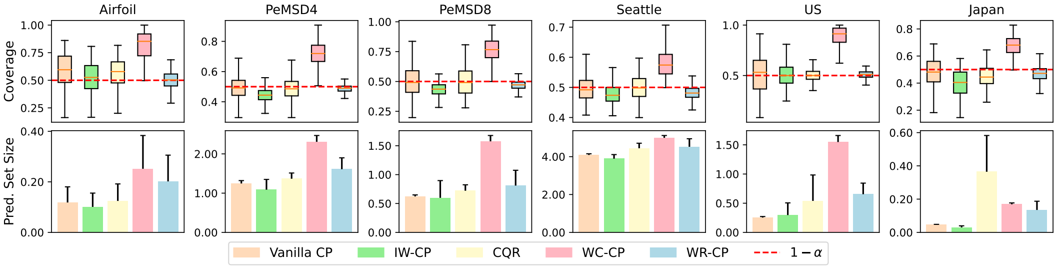

Figure 14: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.5$:** The β values for the WR-CP method are 3, 5, 3, 5, 8, and 13, respectively.

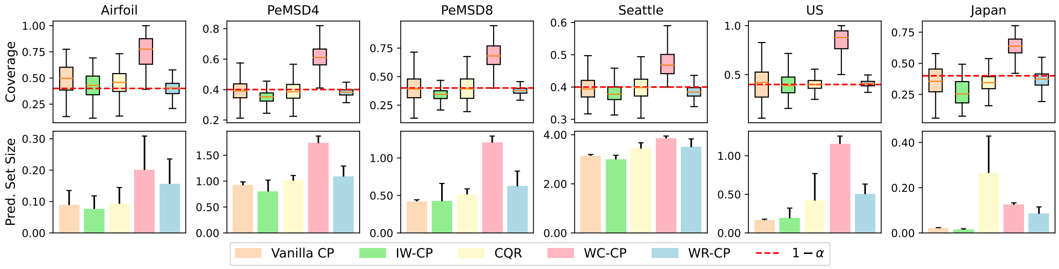

Figure 15: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.6$:** The β values for the WR-CP method are 3, 5, 3, 5, 8, and 13, respectively.

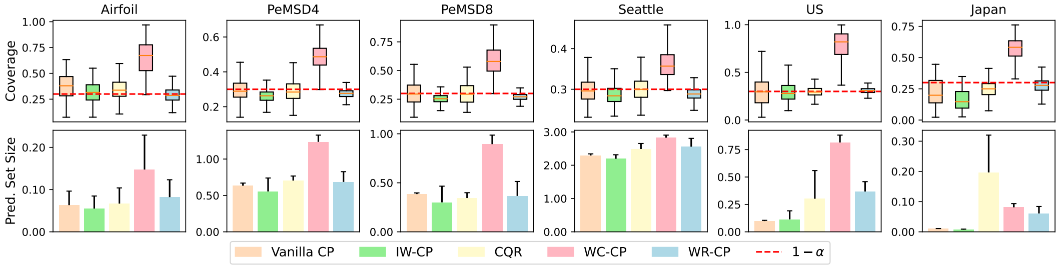

Figure 16: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.7$:** The β values for the WR-CP method are 2, 2, 2, 5, 8, and 10, respectively.

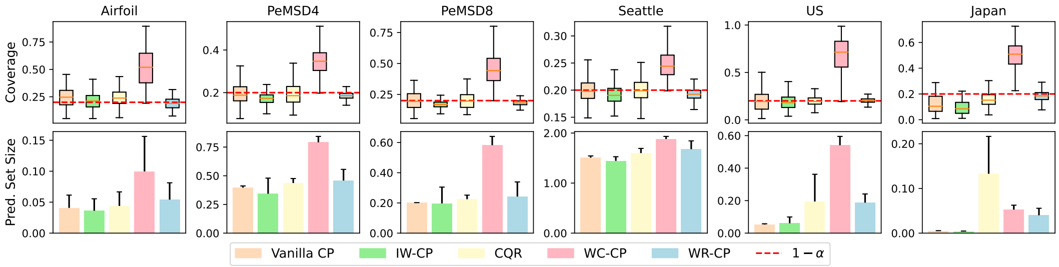

Figure 17: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.8$:** The β values for the WR-CP method are 2, 2, 2, 5, 5, and 10, respectively.

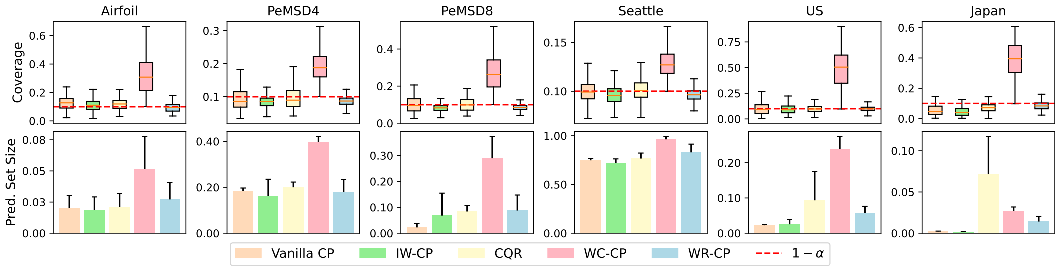

Figure 18: **Coverages and set sizes of WR-CP and baselines with $\alpha = 0.9$:** The β values for the WR-CP method are 2, 1, 1, 5, 2, and 6, respectively.

G Prediction efficiency with coverage guarantee

Although Wasserstein-regularized loss in Eq. (19) offers a controllable trade-off with significantly improved prediction efficiency and a mild coverage loss, it is worth investigating if this efficiency can be achieved with a coverage guarantee. In this section, we first derive a coverage lower bound of WR-CP via the multi-source setup in Appendix G.1. Then, we show that the combination of WC-CP and the Wasserstein-regularized loss can not achieve small prediction sets with ensured coverage in Appendix G.2.

G.1 Coverage guarantee from multi-source setup

Under the setup of multi-source conformal prediction, with τ as the $1-\alpha$ quantile of the weighted calibration conformal score distribution Q_{V,s_P}, we can derive the coverage gap upper bound by

$$\begin{aligned} |F_{Q_{V,s_P}}(\tau) - F_{Q_V}(\tau)| &= \left| \sum_{i=1}^{k} w_i F_{D^{(i)}_{V,s_P}}(\tau) - \sum_{i=1}^{k} w_i F_{D^{(i)}_V}(\tau) \right| \\ &\leq \sum_{i=1}^{k} w_i |F_{D^{(i)}_{V,s_P}}(\tau) - F_{D^{(i)}_V}(\tau)| \\ &\leq \sup_{i\in\{1,\ldots,k\}} |F_{D^{(i)}_{V,s_P}}(\tau) - F_{D^{(i)}_V}(\tau)|. \end{aligned} \tag{38}$$

In other words, the coverage gap on a test distribution must be less or equal to the largest gap at τ among multiple training distributions. Denoting $\alpha_D = \sup_{i\in\{1,\ldots,k\}} |F_{D^{(i)}_{V,s_P}}(\tau) - F_{D^{(i)}_V}(\tau)|$, we have a coverage guarantee $\Pr(Y_{n+1} \in X_{n+1}) \geq 1-\alpha-\alpha_D$.

The regularization term $\sum_{i=1}^{k} W(D^{(i)}_{V,s_P}, D^{(i)}_V)$ in Eq. (19) can minimize α_D, and thus making $1-\alpha-\alpha_D$ closer to the desired $1-\alpha$. It is important to highlight that α_D is adaptive to variations in test distribution Q_V, as evident from Eq. (38). This adaptivity ensures that the lower bound dynamically adjusts to different Q_V. To evaluate the prediction efficiency of WR-CP under this guarantee, we set $\alpha = 0.1$ and computed the corresponding α_D for various test distributions. Additionally, we calculated the coverage and prediction set size of WC-CP on each test distribution, using the corresponding guarantee at $1-\alpha-\alpha_D$ for comparison.

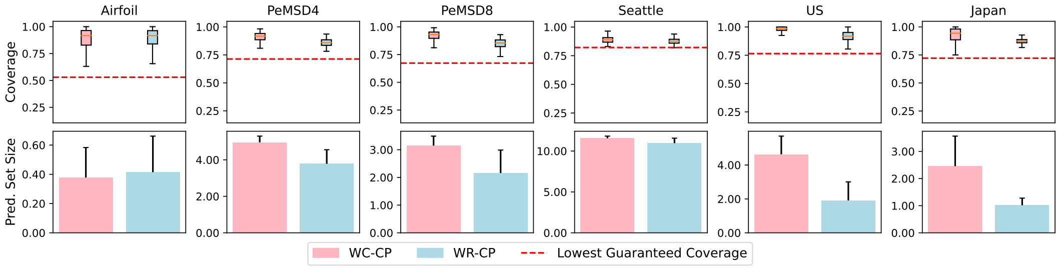

Figure 19: **Coverages and set sizes of WC-CP and WR-CP with coverage guarantee at $1-\alpha-\alpha_D$.**

The experiment results are depicted in Figure 19, demonstrating improved prediction efficiency on the PeMSD4, PeMSD8, U.S.-States, and Japan-Prefectures datasets. However, the efficiency remains almost unchanged on the Seattle-loop dataset and even declines on the airfoil self-noise dataset. This phenomenon can be attributed to the regularization mechanism. While WR-CP enhances prediction efficiency by leveraging the calibration distribution to generate prediction sets, regularization inevitably increases prediction residuals, leading to larger prediction sets. These two opposing effects can interact differently depending on the dataset characteristics. When the efficiency gains outweigh the drawbacks of regularization, we observe reduced prediction set size. Conversely, in datasets like the Seattle-loop and airfoil self-noise, the benefits of regularization are outweighed by the increased prediction residuals, resulting in unchanged or diminished efficiency. The averaged prediction set size reduction across the six datasets is 26.9%.

G.2 Poor compatibility between Wasserstein-regularized loss and WC-CP

Since the WC-CP is a conservative *post-hoc* uncertainty quantification method but the proposed regularized loss in Eq. (19) is applied during *training*, one may consider applying WC-CP upon the model trained by the regularized loss to obtain guaranteed coverage. However, WC-CP and the model are not suitable for complementing each other. While regularization enhances the reliability of calibration distributions, the worst-case approach depends exclusively on the upper bound of $1 - \alpha$ test conformal score quantile, rendering it unable to benefit from regularization. In contrast, the WC-CP may result in larger prediction sets under this condition, as the regularization inevitably increases the prediction residuals, which in turn increases the upper bound of the test conformal score quantile. Experiment results in Figure 20 demonstrate the analysis, where WC-CP is the worst-case method based on a residual-driven model (same as the WC-CP method in Section 6.4), and Hybrid WC-WR represents applying WC-CP to a model trained by Eq. (19).

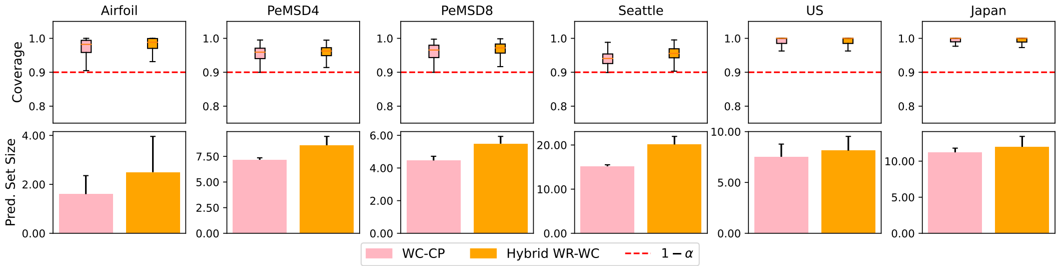

Figure 20: **Coverages and set sizes of WC-CP and Hybrid WC-WR with coverage guarantee** $1 - \alpha = 0.9$**.**

H Limitations

H.1 Susceptibility to density estimation errors

Given that Wasserstein regularization relies on importance-weighted conformal scores, its performance is greatly influenced by the accuracy of the estimated likelihood ratio obtained through KDE. Inaccurate estimation can significantly impact the effectiveness of WR-CP. For instance, in Figure 4, WR-CP yields larger prediction set sizes with less concentrated coverages on the airfoil self-noise dataset compared to other datasets. This can be attributed to the airfoil self-noise dataset having the highest feature dimension (5) and the smallest size of the sampled $\mathcal{S}_{XY}^{P}$ (500). These challenges in KDE lead to suboptimal performance of WR-CP on the airfoil self-noise dataset when compared to its performance on others.

The main reason for KDE error is numerical instability, which can arise from several factors. A poor choice of kernel is a critical contributor; for instance, kernels with sharp edges or discontinuities, such as rectangular or triangular kernels, can result in jagged density estimates and amplify errors near boundaries. Fat-tailed kernels, such as the Cauchy kernel, may assign excessive weight to distant data points, leading to inaccuracies in density estimates and numerical precision challenges. Additionally, the lack of feature normalization can exacerbate the effects of extreme values, skewing the density estimation process and reducing computational stability. Lastly, inappropriate bandwidth selection, either too small (overfitting) or too large (underfitting), can disrupt the balance between bias and variance, further contributing to instability in the estimation.

In our work, we first adopted the Gaussian kernel, valued for its smoothness and numerical stability. To mitigate the influence of extreme values, we applied feature normalization, ensuring a more stable density estimation process. Additionally, we conducted a comprehensive grid search to fine-tune the bandwidth, achieving an optimal balance between bias and variance for robust and accurate results. The bandwidth candidates were selected from a logarithmically spaced range between 10^{-2} and $10^{0.5}$, consisting of 20 evenly distributed values on a logarithmic scale.

H.2 Computational challenges in KDE

We applied a grid search approach to identify the optimal bandwidth for KDE, which ensures an effective balance between bias and variance in density estimation. However, this method often involves extensive computational effort, particularly when working with high-dimensional datasets, as it requires repeated calculations over a range of bandwidth values. To address this challenge, Bernacchia–Pigolotti KDE (Bernacchia & Pigolotti, 2011) introduces an innovative framework that combines a Fourier-based filter with a systematic approach for simultaneously determining both the kernel shape and bandwidth. This method not only reduces subjectivity in kernel selection but also offers a more efficient computational pathway. Building on this foundation, FastKDE (O'Brien et al., 2016) adapts and extends the Bernacchia–Pigolotti approach for high-dimensional scenarios, incorporating optimizations that significantly improve computational speed and scalability. These advancements represent promising directions for mitigating the computational overhead in our own work, where similar strategies could be leveraged to streamline the bandwidth selection process and enhance the overall efficiency of KDE in complex datasets.

H.3 Other choices of the calibration distribution

In the experiments conducted in Section 6, we specifically examine the scenario where the calibration data follows a mixture distribution of $D_{XY}^{(i)}$ for $i = 1, ..., k$ with equal weights. However, this may not always be the case in real-world situations. Given that the calibration distribution plays a crucial role in determining the difficulty of minimizing Eq. (19) during training, it is valuable to investigate the performance of WR-CP with a calibration distribution different from a mixture of training distributions.