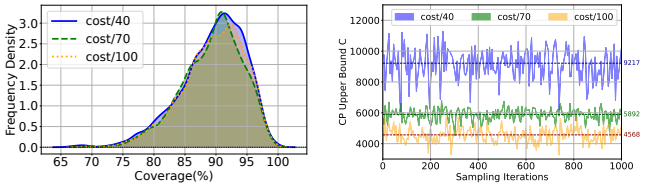
Additional Experiments on Paper ID 1106: "Conformal Prediction for Verifiable Learned Query Optimization"

Anonymous Author(s)

1 Evaluation

1.1 Hyper-Parameter Micro-benchmarking.

Impact of Normalization. The third hyper-parameter is the normalization selection. In some cases, models do not numerically predict the cost $\hat{c}$ well-aligned with the actual latency t. In these situations, we apply a normalization function f to obtain the normalized cost $f(\hat{c})$. We perform $f(\hat{c}) = \hat{c}/40$, $f(\hat{c}) = \hat{c}/70$ and $f(\hat{c}) = \hat{c}/100$ on RTOS. Figure 1 (a) shows that multiple normalization methods exhibit similar trends; this indicates that our CP-based latency guarantee method can effectively handle different normalizations while maintaining the same $1-\delta$ guarantee. This also indicates that proper normalization has a low impact on the overall CP Theory, confirming the robustness of our approach against ill-aligned normalizations. Figure 1 (b) shows changing the normalization function primarily affects the non-conformity scores R, which in turn influences the values of corresponding upper bound C. Lower values in R and C lead to more accurate latency guarantees. Based on this consideration, we select $f(\hat{c}) = \hat{c}/40$ for Lero prototype and $f(\hat{c}) = \hat{c}/100$ for RTOS prototype.

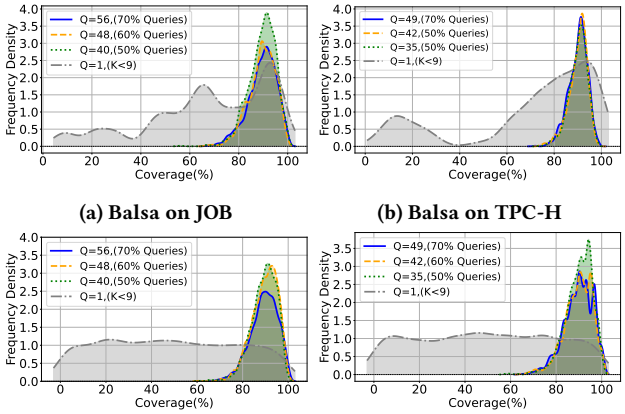


(a) Coverage **(b) CP Upper Bound Distribution**

Figure 1: Impact on Normalization Function f.

Changing Calibration Size. Another hyper-parameter is the size of the calibration set. In Lemma 1, we discuss that the lower bound of the calibration set size is $K \geq \frac{1-\delta}{\delta}$. In this experiment, we aim to observe how the size of the calibration set influences coverage. Setting $\delta = 0.1$, we obtain a threshold for the calibration size of $K^* = 9$ based on the previous equation.

Figure 2 illustrates the influence of different calibration set sizes. For JOB, we use $Q = 56$ (70% of queries as calibration queries Q^{Cal} among all calibration-test queries), $Q = 48$ (60% queries), $Q = 40$ (50% queries), and $Q = 1$ ($K < 9$). For TPC-H, we use $Q = 49$, $Q = 42$, $Q = 35$, and $Q = 1$ ($K < 9$) correspondingly. Since each query contains multiple operators that contribute to the total K, in the first three cases for each workload, K greatly exceeds K^*. In the fourth case, only one query is selected, representing the scenario where $K < K^*$. We observe that if the calibration size exceeds the threshold K^*, the curve shows similar coverage as expected. However, when the calibration size is below K^*, the curves exhibit a relatively flat trend, indicating that the current C cannot reliably provide the expected guarantee of $1 - \delta = 0.9$.



(a) Balsa on JOB **(b) Balsa on TPC-H**

(c) Lero on JOB **(d) Lero on TPC-H**

Figure 2: Impact of Changing Calibration Size.