Proceedings of Machine Learning Research 1–3

# A Gibbs-Boltzmann Approach to Conformal Prediction under Distribution Shift

**Shivaprasad Channapura Sudhakara** SC3280@SCARLETMAIL.RUTGERS.EDU
*Rutgers University New Brunswick*

**Ilia Nouretdinov** I.R.NOURETDINOV@CS.RHUL.AC.UK
*RHUL, United Kingdom*

## 1. Introduction

Building on the foundation of standard Conformal Prediction and its known vulnerabilities to covariate shift, we formalize our Thermodynamic DRO framework. While recent methods such as Aolaritei et al. (2025) address score shift via topological metrics like the Lévy–Prokhorov distance, our approach introduces a statistical mechanics perspective. By applying a physics-inspired thermodynamic base measure directly to 1D non-conformity scores within a KL-divergence ambiguity set, we yield a smooth, closed-form partition function that expands conformal thresholds to absorb out-of-distribution errors with minimal computational overhead.

Our robust calibration procedure, which can be implemented using standard optimization libraries (Pedregosa et al., 2011), is grounded in established finite-sample concentration inequalities (Massart, 1990; van der Vaart and Wellner, 1996).

Assume a finite calibration set of size $N$.

- Let $s_i$ be the non-conformity score of the $i$-th calibration set.
- Let $\pi_i$ be the possible probability mass assumed to the $i$-th point.
- Let $k_i$ be the normalized Gibbs-Boltzmann kernel weight for the $i$-th point from

$$K_T(x, y) = \exp\left(-\frac{d(x, y)}{T}\right)$$

- Let the binary indicator be $l_i = \mathbb{I}[s_i \leq q]$.

## 2. Methodology: Primal Problem and Solving for Optimal Distribution

To find the worst-case distribution that minimizes the expected coverage, which is constrained by a KL-divergence ball of radius $\eta$ around a Gibbs-Boltzmann kernel:

$$\inf_{\pi} \sum_{i=1}^{N} \pi_i l_i \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i \log\left(\frac{\pi_i}{k_i}\right) \leq \eta \quad \text{and} \quad \sum_{i=1}^{N} \pi_i = 1 \tag{1}$$

We convert the constrained optimization into a single function using Lagrangian Multipliers $\lambda \geq 0$ for the KL-ball constraint and $\nu$ for the normalization constant:

$$L(\pi, \lambda, \nu) = \sum_{i=1}^{N} \pi_i l_i + \lambda \left( \sum_{i=1}^{N} \pi_i \log \left( \frac{\pi_i}{k_i} \right) - \eta \right) + \nu \left( \sum_{i=1}^{N} \pi_i - 1 \right) \tag{2}$$

Taking the partial derivative of the Lagrangian w.r.t $\pi_i$ and setting it to zero yields the optimal probability mass $\pi_i^*$:

$$\frac{\partial L}{\partial \pi_i} = l_i + \lambda \left( \log \left( \frac{\pi_i}{k_i} \right) + 1 \right) + \nu = 0 \implies \pi_i^* = \frac{k_i}{Z} \exp \left( -\frac{l_i}{\lambda} \right) \tag{3}$$

where $Z = \sum_{i=1}^{N} k_i \exp(-l_i/\lambda)$. Substituting $\pi_i^*$ back into the Lagrangian yields $L(\pi^*, \lambda, \nu) = -\lambda \log Z - \lambda\eta$. To find the worst-case coverage, we take the supremum over $\lambda > 0$, which is equivalent to the infimum of the negative objective:

$$\inf_{\lambda > 0} \{\lambda \log Z + \lambda\eta\} \tag{4}$$

## 3. Theoretical Guarantees

**Theorem 1 (Finite-Sample Concentration)** *Let $n$ be the calibration set size. Let $\hat{S}_n(q)$ denote the empirical CDF of calibration scores, and $S(q)$ the true population distribution. For any error probability $\delta \in (0, 1)$, if the robust threshold $q_{robust}$ is chosen such that the empirical worst-case coverage is $1-\alpha$, then with probability at least $1-\delta$, the true population worst-case coverage is bounded by:*

$$Coverage_{true} \geq 1 - \alpha - \sqrt{\frac{\ln(2/\delta)}{2n}} \tag{5}$$

*Proof Sketch.* The result follows directly from the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Massart, 1990), which guarantees that $\sup_q |\hat{S}_n(q) - S(q)| \leq \sqrt{\ln(2/\delta)/(2n)}$ with probability $1 - \delta$.

## 4. Experimental Results

We evaluated our framework on the California Housing dataset (Pace and Barry, 1997) ($N = 3000$) using a Random Forest regressor ($\alpha = 0.10, \eta = 0.05$). Covariate shift was simulated via Gaussian noise $\mathcal{N}(0, \sigma^2_{\text{noise}})$ applied to test features. As shown in Table 1, standard CP fails under perturbation as widening errors escape fixed bounds. Conversely, Thermodynamic DRO proactively expands the threshold, successfully preserving robust coverage without requiring model retraining.

Table 1: Coverage under Covariate Shift ($\alpha = 0.10$)

| **Noise Level** | **Standard CP** | **Robust DRO** |
|---|---|---|
| 0% (Baseline) | 90.00% | 90.00% |
| 75% Std. Dev. | 64.93% | 80.80% |
| 85% Std. Dev. | 63.87% | 79.47% |
| 95% Std. Dev. | 62.00% | 77.33% |

## References

Liviu Aolaritei, Zheyu Oliver Wang, Julie Zhu, Michael I. Jordan, and Youssef Marzouk. Conformal prediction under Lévy–Prokhorov distribution shifts: Robustness to local and global perturbations. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2502.14105.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.