From Matrix Inversion to Constraints: Provably Tighter Confidence Regions for Importance Weights in Label Shift

Mushan Li*, **Kihyun Han***, **Yanyuan Ma**
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
{mzl5849, kqh5716, yzm63}@psu.edu

Abstract

Importance weights are essential in domain adaptation under label shift, yet their utility is often undermined by the finite sample uncertainty associated with their estimation. Existing methods typically analyze this uncertainty through Gaussian elimination on interval-valued linear systems, which leads to overly conservative confidence regions and inefficient downstream applications. We propose a paradigm shift from inversion-based inference to a direct matrix constraint framework. We use this framework to define a joint confidence region and extract marginal intervals via linear programming, deriving provably tighter bounds for importance weights while maintaining exact finite-sample validity. Furthermore, we analyze the confidence region's geometry and provide the theoretical results for its diameter bounds. Evaluated across text, image, multimodal benchmarks, including AGNews, MNIST, CIFAR-10, N24News, and a real-world autonomous driving dataset, nuImages, our approach consistently yields shorter confidence intervals and smaller prediction sets than inversion-based methods.

1 Introduction

Label shift is a pervasive challenge in machine learning, occurring when label distributions differ between source and target domains while the conditional distribution of features remains invariant. This scenario arises frequently in real-world applications. For example, it occurs when adapting models trained on historical medical data to new patient populations, or when deploying classifiers across varying geographic regions where class prevalences shift due to demographic or environmental factors. In these cases, the "importance weights", the ratios of target to source label probabilities, act as the essential bridge for model adaptation, enabling tasks like weighted empirical risk minimization and rigorous uncertainty quantification [17, 2, 11, 25, 29].

The confusion matrix approach, exemplified by Black-Box Shift Estimation (BBSE) [17] and Regularized Learning under Label Shifts (RLLS) [2], is the most widely adopted technique because its statistical consistency is independent of classifier calibration. While these methods are effective for point estimation, they generally overlook the inherent statistical uncertainty in the importance weights estimation. When uncertainty is taken into account, it is mostly considered in the asymptotic sense [32]. Notably, Saerens et al. [27] and Alexandari et al. [1] devised maximum likelihood estimators to construct confidence intervals, which rely on the infinite data assumption and, crucially, necessitate well-calibrated classifiers [11]. In finite-sample settings, however, the asymptotic approximations can be unreliable and may fail to account for uncertainties that propagate to downstream tasks.

We are only aware of one work [29] that tries to establish uncertainty quantification in the finite sample setting, where they quantify the importance weights uncertainty by bounding errors in the

*Equal contribution. Correspondence to Mushan Li <mzl5849@psu.edu>

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Gaussian Elimination (GE) procedure. Si et al. [29] considered an interval-version of BBSE, where they replaced every element in the confusion matrix and every element in the target predicted label proportion vector by its corresponding confidence interval, and modified the traditional GE to solve this interval-valued linear system. Thus, this method still builds on the "matrix inversion" mindset. By performing row reduction or inversion on intervals, they introduce unnecessary relaxations and "noise amplification", which leads to loose and suboptimal confidence regions. In practice, this means that it becomes overly conservative. In our numerical analysis, we find that the GE procedure frequently overestimates the uncertainty and offers little practical value for decision-making.

In this work, we propose Matrix Constraints Linear Programming (MaC-LP), a new uncertainty quantification method for the importance weights. It represents a fundamental paradigm shift: from matrix inversion to matrix constraints. Rather than trying to solve the system algebraically or relying on asymptotic approximations, we express the feasible range of importance weights as a system of matrix constraints (MaC) derived from the confidence intervals of the observations. By applying linear programming (LP) to optimize over these constraints, we compute the provably shorter confidence intervals for each one of importance weights.

MaC-LP ensures exact finite-sample validity. The resulting tighter boundaries have a direct, transformative impact on the efficiency of downstream tasks. By sharpening the confidence intervals of the estimated importance weights, we allow for more precise uncertainty-aware predictions. This ensures that even in the label shifted environments, the estimation and inference in the target domain remain valid in finite samples, while avoiding excessive conservatism. Our specific contributions are as follows:

- Constraint-based framework: We introduce a novel matrix-inversion-free approach to formulate confidence regions for importance weights. We propose to construct the confidence regions via the matrix constraints (MaC) and obtain the marginal confidence intervals by the linear programming (LP).
- Optimality and superiority over asymptotic and inversion method: We prove that our confidence region has exact finite-sample validity, in contrast to the asymptotic based methods. We prove that our result achieves the smallest confidence region possible in finite-sample regimes given the elementwise confidence intervals of the confusion matrix and the predicted label proportions. We prove that our linear programming-based confidence intervals are no wider than the ones derived from Gaussian elimination.
- Enhanced downstream efficiency: We prove that the tighter confidence region directly translates to precise and smaller uncertainty-aware prediction sets in the target domain under label shift.
- Empirical validation: Through experiments on diverse benchmarks, including text (AGNews), images (MNIST, CIFAR-10), multimodal data (N24News), and a real-world autonomous driving dataset (nuImages), we show significant reductions in both confidence interval lengths of the importance weights and the size of downstream prediction sets compared to those by Gaussian Elimination.

2 Problem formulation and our goal

In this section, we introduce the basic notation, motivate the role of importance weights in correcting distribution mismatch under label shift, and present our primary goal of this work.

Basic notation. Let $\mathcal{X}$ and $\mathcal{Y} = \{1, \ldots, K\} \equiv [K]$ denote the feature and label spaces, respectively. In the unsupervised label shift setting, we are given a labeled source dataset $\mathcal{D}_{\text{src}} = \{(\boldsymbol{x}_s, y_s)\}_{s=1}^{m}$ drawn i.i.d. from the source distribution P, and an unlabeled target dataset $\mathcal{D}_{\text{tgt}} = \{\boldsymbol{x}_{m+t}\}_{t=1}^{n}$ drawn from the target distribution Q_X, where Q_X is an $\mathbf{X}$-marginal of the target distribution Q, and (m, n) are finite sample sizes. Both P and Q are defined on $\mathcal{X} \times \mathcal{Y}$ with corresponding probability density/mass functions $p(\cdot)$ and $q(\cdot)$. Specifically, $p(\boldsymbol{x}, y)$, $p(\boldsymbol{x}|y)$, $p(\boldsymbol{x})$ and $p(y)$ denote the joint distribution, the class-conditional distribution, the marginal feature and the marginal label distribution, respectively, under P, with analogous notations under Q. In addition, let $\mathbb{P}$ denote probability, with the underlying measure/distribution indicated via subscript when necessary (e.g., $\mathbb{P}_{(\mathbf{X},Y)\sim P}, \mathbb{P}_{\mathbf{X}\sim Q_X}$).

Importance weights under label shift. The *label shift* assumption posits that class-conditional distributions are invariant across domains, i.e., $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$ for all $y \in \mathcal{Y}$, while the marginal label distributions differ, i.e., $p(y) \neq q(y)$. A direct consequence of this assumption is that the joint distributions are related via a ratio of their label proportions:

$$q(\boldsymbol{x}, y) = p(\boldsymbol{x} \mid y)q(y) = p(\boldsymbol{x} \mid y)p(y)\frac{q(y)}{p(y)} = p(\boldsymbol{x}, y)\omega_y,$$

where $\omega_y \equiv q(y)/p(y)$ represents the *importance weights*. These weights, denoted by $\boldsymbol{\omega} = (\omega_1, ..., \omega_K)^\top$, serve as the fundamental bridge between domains. For example, it allows us to express target expectations in terms of the source distribution: $\mathbb{E}_Q\{h(\mathbf{X}, Y)\} = \mathbb{E}_P\{\omega_Y h(\mathbf{X}, Y)\}$, where $\mathbb{E}$ denotes the expectation operator and h is a scalar function defined on $\mathcal{X} \times \mathcal{Y}$. In practice, while the true importance weights $\boldsymbol{\omega}^*$ are unknown, using $\widehat{\boldsymbol{\omega}}$ estimated by finite samples $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tgt}}$ is an essential way for downstream tasks such as *weighted conformal prediction* [25], where $\widehat{\boldsymbol{\omega}}$ is used to recalibrate non-conformity scores to maintain valid coverage guarantees under label shift.

Our goal: improve inference for importance weights. While obtaining a point estimate $\widehat{\boldsymbol{\omega}}$ is the standard approach to addressing label shift, its utility is compromised by its uncertainty. In the finite-sample case, when the source and/or target sample sizes (m, n) are small, the estimate $\widehat{\boldsymbol{\omega}}$ becomes highly sensitive to the data noise.

Using prediction set as a representative example, Figure 1 illustrates that replacing $\boldsymbol{\omega}^*$ by $\widehat{\boldsymbol{\omega}}$ for downstream tasks, such as weighted conformal prediction, can lead to a coverage gap. It is because when the uncertainty in $\widehat{\boldsymbol{\omega}}$ is ignored, the resulting coverage probability can fall strictly below the nominal level 90% when $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tgt}}$ are small, causing prediction sets to be unreliably narrow and miss the ground truth more often than permitted.

This example motivates us to move beyond a point estimate and instead consider a region of importance weights, $\boldsymbol{\Omega}$, such that $\mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}) \geq 1 - \delta$. Meanwhile, to avoid an excessively conservative prediction set, the confidence region $\boldsymbol{\Omega}$ should be as small as possible. Also, a tighter confidence region improves sample efficiency, requiring fewer observations to achieve the same target precision at a fixed confidence level. In the following sections, we introduce the MaC-LP framework to derive these confidence regions through a novel matrix constraint approach.

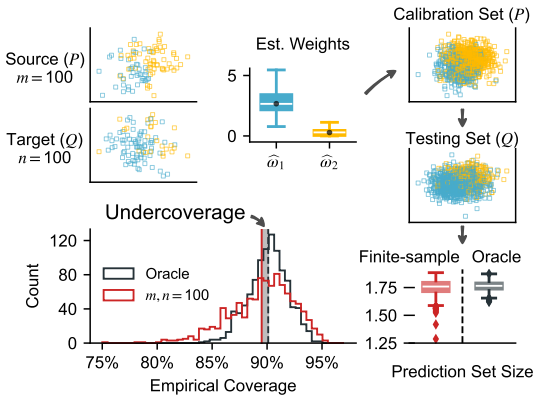


Figure 1: Undercoverage of weighted conformal prediction using $\widehat{\boldsymbol{\omega}}$. Step-histograms show frequency of empirical coverage over 1,000 repetitions. While the Oracle (black) centers on the 90% nominal confidence level, point estimation of importance weights (red) introduces a systematic *coverage gap*, even though the estimation is unbiased (true importance weights are represented by black dots in the boxplots of Est. Weights). At $m, n = 100$, estimation error leads to significant undercoverage. This motivates the need for quantifying the uncertainty in estimated importance weights to ensure its application in the finite-sample setting.

3 Finite-sample uncertainty quantification under label shift

To facilitate the description of our approach, we briefly review the confusion matrix framework for label shift estimation [17, 11]. While this framework was previously proposed for point estimation of the importance weights, we extend it to define a direct matrix constraint system that captures finite-sample uncertainty.

3.1 Preparation: confusion matrix framework

Suppose that we have a classifier g, defined as $g : \mathcal{X} \to [K]$. Define the $K \times K$ (joint) confusion matrix $\mathbf{C}^* = (C^*_{ij})_{i,j\in[K]}$ by $C^*_{ij} \equiv \mathbb{P}_{(\mathbf{X},Y)\sim P}\{g(\mathbf{X}) = i,\ Y = j\}$ for $i, j \in [K]$. In the target domain, define the predicted label proportions $\mathbf{q}^* = (q^*_1, \ldots, q^*_K)^\top$ by $q^*_k \equiv \mathbb{P}_{\mathbf{X}\sim Q_X}\{g(\mathbf{X}) =$

$k\}$ for $k \in [K]$. By the label shift assumption, we have $\mathbb{P}_{(\mathbf{X},Y)\sim P}\{g(\mathbf{X}) = i \mid Y = j\} = \mathbb{P}_{(\mathbf{X},Y)\sim Q}\{g(\mathbf{X}) = i \mid Y = j\}$ for all $i, j \in [K]$. Therefore, for each $i \in [K]$, we obtain that $\sum_{j=1}^{K} C^*_{ij}\omega^*_j = q^*_i$, which implies $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$. We provide the detailed proof in Appendix F.1.

In the source domain, given finite samples $\mathcal{D}_{\text{src}} = \{(\boldsymbol{x}_s, y_s)\}_{s=1}^{m}$, for $i, j \in [K]$, let $M_{ij} = \sum_{s=1}^{m} I\{g(\boldsymbol{x}_s) = i,\ y_s = j\}$, where $I\{\cdot\}$ is the indicator function. Denote $\mathbf{M} = (M_{ij})_{i,j\in[K]}$ which is a $K \times K$ matrix. Note that $\sum_{i=1}^{K}\sum_{j=1}^{K} M_{ij} = m$. In the target domain, given finite samples $\mathcal{D}_{\text{tgt}} = \{\boldsymbol{x}_{m+t}\}_{t=1}^{n}$, for $k \in [K]$, let $N_k = \sum_{t=1}^{n} I\{g(\boldsymbol{x}_{m+t}) = k\}$, and denote $\mathbf{N} = (N_1, \ldots, N_K)^\top$. Similarly, we have $\sum_{k=1}^{K} N_k = n$. Lipton et al. [17] derived the plug-in estimator $\widehat{\boldsymbol{\omega}} = \widehat{\mathbf{C}}^{-1}\widehat{\mathbf{q}}$ by taking $\widehat{\mathbf{C}} = m^{-1}\mathbf{M}$ and $\widehat{\mathbf{q}} = n^{-1}\mathbf{N}$ (see Appendix D for an overview of importance weights estimation). In contrast, to construct the confidence region for $\boldsymbol{\omega}^*$, we marginalize each element of random matrix $\mathbf{M}$ and random vector $\mathbf{N}$ to get $M_{ij} \sim \text{Binomial}(m, C^*_{ij})$ and $N_k \sim \text{Binomial}(n, q^*_k)$ for each $i, j, k \in [K]$. Using standard confidence intervals for binomial probability, such as Clopper-Pearson interval or Hoeffding's inequality-based interval (details in Appendix E), we can obtain the elementwise confidence intervals $[\underline{C}_{ij}, \overline{C}_{ij}]$ for C^*_{ij} and $[\underline{q}_k, \overline{q}_k]$ for q^*_k with coverage levels $1 - \delta_{ij}$ and $1 - \delta_k$, respectively, where $\delta_{ij}, \delta_k \geq 0$.

3.2 MaC-LP: inference via matrix constraints and linear programming

Following the notation from Section 3.1, given the confidence intervals $[\underline{C}_{ij}, \overline{C}_{ij}]$ and $[\underline{q}_k, \overline{q}_k]$ for all $i, j, k \in [K]$, we denote the collections of these endpoints as $\underline{\mathbf{C}} = (\underline{C}_{ij})_{i,j\in[K]}$, $\overline{\mathbf{C}} = (\overline{C}_{ij})_{i,j\in[K]}$, $\underline{\mathbf{q}} = (\underline{q}_k)_{k\in[K]}$, and $\overline{\mathbf{q}} = (\overline{q}_k)_{k\in[K]}$. In addition, for any vectors or matrices $\mathbf{a} = (a_1, \ldots, a_L)$ and $\mathbf{b} = (b_1, \ldots, b_L)$ of the same size, denote $\mathbf{a} \leq \mathbf{b}$ if $a_\ell \leq b_\ell$ for all $\ell \in [L]$, and let the same definitions apply to $\mathbf{a} < \mathbf{b}$, $\mathbf{a} \geq \mathbf{b}$, and $\mathbf{a} > \mathbf{b}$. With these notations, we recast the confidence intervals as $\mathbf{C}^* \in [\underline{\mathbf{C}}, \overline{\mathbf{C}}] \equiv \boldsymbol{\mathcal{C}}$ and $\mathbf{q}^* \in [\underline{\mathbf{q}}, \overline{\mathbf{q}}] \equiv \boldsymbol{\mathcal{Q}}$.

With these confidence intervals in hand, we aim to construct the $(1 - \delta)$-level confidence region for the importance weights $\boldsymbol{\omega}^*$. To this end, we collect all $\boldsymbol{\omega}$ vectors that can be generated by an admissible pair $(\mathbf{C}, \mathbf{q}) \in \boldsymbol{\mathcal{C}} \times \boldsymbol{\mathcal{Q}}$ and define

$$\boldsymbol{\Omega}_0 \equiv \big\{\boldsymbol{\omega} : \exists\, \mathbf{C} \in \boldsymbol{\mathcal{C}},\ \exists\, \mathbf{q} \in \boldsymbol{\mathcal{Q}} \text{ such that } \mathbf{C}\boldsymbol{\omega} = \mathbf{q},\ \boldsymbol{\omega} \geq \mathbf{0}\big\}.$$

By the definition of $\boldsymbol{\Omega}_0$, $\boldsymbol{\omega}^* \in \boldsymbol{\Omega}$ provided that $\mathbf{C}^* \in \boldsymbol{\mathcal{C}}$ and $\mathbf{q}^* \in \boldsymbol{\mathcal{Q}}$.

Let $\delta = \sum_{i,j\in[K]} \delta_{ij} + \sum_{k\in[K]} \delta_k$. If the intervals $[\underline{C}_{ij}, \overline{C}_{ij}]$ and $[\underline{q}_k, \overline{q}_k]$ are valid confidence intervals for C^*_{ij} and q^*_k, respectively, then the union bound imply that $\boldsymbol{\Omega}_0$ is a valid confidence region for $\boldsymbol{\omega}^*$, as summarized in Proposition 3.1. The proof is in Appendix F.2.

Proposition 3.1. *Let $\delta = \sum_{i,j\in[K]} \delta_{ij} + \sum_{k\in[K]} \delta_k$. Suppose $\mathbb{P}(C^*_{ij} \in [\underline{C}_{ij}, \overline{C}_{ij}]) \geq 1 - \delta_{ij}$ for $i, j \in [K]$ and $\mathbb{P}(q^*_k \in [\underline{q}_k, \overline{q}_k]) \geq 1 - \delta_k$ for $k \in [K]$. Then $\boldsymbol{\Omega}_0$ is a valid $(1 - \delta)$-level confidence region for $\boldsymbol{\omega}^*$.*

However, computing $\boldsymbol{\omega}$ by solving the linear system $\mathbf{C}\boldsymbol{\omega} = \mathbf{q}$ for infinitely many pairs $(\mathbf{C}, \mathbf{q}) \in \boldsymbol{\mathcal{C}} \times \boldsymbol{\mathcal{Q}}$ is numerically inconvenient. Instead, we directly impose the linear constraints relying on the linear system $\mathbf{C}\boldsymbol{\omega} = \mathbf{q}$. Given a $\mathbf{C} \in \boldsymbol{\mathcal{C}}$, if $\mathbf{C}\boldsymbol{\omega} = \mathbf{q} \in \boldsymbol{\mathcal{Q}}$, then the quantity $(\mathbf{C}\boldsymbol{\omega})_i = \sum_{j=1}^{K} C_{ij}\omega_j$ must lie in $[\underline{q}_i, \overline{q}_i]$ for each $i \in [K]$. Since $\mathbf{C} \in [\underline{\mathbf{C}}, \overline{\mathbf{C}}]$, a necessary condition for this to be possible is that $[\sum_{j=1}^{K} \underline{C}_{ij}\omega_j,\ \sum_{j=1}^{K} \overline{C}_{ij}\omega_j]$ overlaps with $[\underline{q}_i, \overline{q}_i]$, i.e., for $i \in [K]$,

$$\sum_{j=1}^{K} \overline{C}_{ij}\omega_j \geq \underline{q}_i \text{ and } \sum_{j=1}^{K} \underline{C}_{ij}\omega_j \leq \overline{q}_i.$$

The following theorem shows that the above condition is not only necessary but also sufficient, and hence $\boldsymbol{\Omega}_0$ admits an equivalent characterization via explicit linear inequalities. Throughout this work, we use MaC and $\boldsymbol{\Omega}$ interchangeably to refer to the construction in Theorem 3.2. The proof of Theorem 3.2 is provided in Appendix F.3.

Theorem 3.2. *Let $\boldsymbol{\mathcal{C}} = [\underline{\mathbf{C}}, \overline{\mathbf{C}}]$ and $\boldsymbol{\mathcal{Q}} = [\underline{\mathbf{q}}, \overline{\mathbf{q}}]$. Then*

$$\boldsymbol{\Omega} = \big\{\boldsymbol{\omega} : \overline{\mathbf{C}}\boldsymbol{\omega} \geq \underline{\mathbf{q}},\ \underline{\mathbf{C}}\boldsymbol{\omega} \leq \overline{\mathbf{q}},\ \boldsymbol{\omega} \geq \mathbf{0}\big\}. \tag{1}$$

In particular, $\boldsymbol{\Omega}$ is a convex set.

Remark 3.3 (Connection to the Oettli-Prager Theorem). The explicit solution derived in Theorem 3.2 happens to coincide with the classical Oettli-Prager Theorem [22] restricted to the nonnegative orthant. We detail the connection in Appendix D.1[2].

Although MaC offers a closed-form description of this region, its shape may be irregular. In practice, when a regular shape such as an axis-aligned hyperrectangle is preferred, we can apply the linear programming on the linear-inequality characterization, $\boldsymbol{\Omega}$, in Theorem 3.2. It computes the classwise bounds of $\boldsymbol{\Omega}$ by solving $2K$ linear programs: for each $k \in [K]$,

$$\underline{\omega}_k = \min \omega_k \text{ subject to } \overline{\mathbf{C}}\boldsymbol{\omega} \geq \underline{\mathbf{q}},\ \underline{\mathbf{C}}\boldsymbol{\omega} \leq \overline{\mathbf{q}},\ \boldsymbol{\omega} \geq \mathbf{0}, \tag{2}$$

$$\overline{\omega}_k = \max \omega_k \text{ subject to } \overline{\mathbf{C}}\boldsymbol{\omega} \geq \underline{\mathbf{q}},\ \underline{\mathbf{C}}\boldsymbol{\omega} \leq \overline{\mathbf{q}},\ \boldsymbol{\omega} \geq \mathbf{0}. \tag{3}$$

Collecting these bounds yields the smallest axis-aligned hyperrectangle ("squeezing box") that contains $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega}_{\mathrm{LP}} \equiv \prod_{k=1}^{K} [\underline{\omega}_k, \overline{\omega}_k]. \tag{4}$$

Since the objective ω_k in (2)–(3) is linear in $\boldsymbol{\omega}$ and the constraints are also linear, we can compute $\boldsymbol{\Omega}_{\mathrm{LP}}$ using off-the-shelf linear programming methods (e.g., the simplex algorithm). Figure S1 in Appendix A summarizes the construction of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_{\mathrm{LP}}$ in a binary-classification toy example.

Since $\boldsymbol{\Omega}_{\mathrm{LP}}$ is defined as an axis-aligned hyperrectangle, its side lengths serve as a natural metric for uncertainty, enabling a direct comparison of confidence region sizes across different methodologies. A closely related approach by Si et al. [29] constructs a hyperrectangular region $\boldsymbol{\Omega}_{\mathrm{GE}}$ for $\boldsymbol{\omega}^*$ by evaluating the inversion $\mathbf{C}^{-1}\mathbf{q}$ for all $\mathbf{C} \in \boldsymbol{\mathcal{C}}$ and $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$ using interval-based Gaussian elimination (see Appendix I). In Corollary 3.4, we establish the formal set inclusions between the fundamental region $\boldsymbol{\Omega}$, our proposed $\boldsymbol{\Omega}_{\mathrm{LP}}$, and the conservative $\boldsymbol{\Omega}_{\mathrm{GE}}$, while maintaining $1-\delta$ coverage. The proof of Corollary 3.4 is provided in Appendix F.4. A two-class example illustrating why interval Gaussian elimination is conservative is given in Appendix I.1.

Corollary 3.4. *When $\boldsymbol{\Omega}_{\mathrm{GE}}$ exists, $\boldsymbol{\Omega} \subset \boldsymbol{\Omega}_{\mathrm{LP}} \subset \boldsymbol{\Omega}_{\mathrm{GE}}$. Under the conditions of Proposition 3.1,*

$$1-\delta \leq \mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}) \leq \mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}_{\mathrm{LP}}) \leq \mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}_{\mathrm{GE}}).$$

Note that every optimality statement about $\boldsymbol{\Omega}$ is conditional on the elementwise confidence intervals $\boldsymbol{\mathcal{C}} \times \boldsymbol{\mathcal{Q}}$. Additional simplex or joint multinomial constraints on $(\mathbf{C}, \mathbf{q})$ can shrink the region further, but at the cost of convexity and of the linear characterization (Appendix B).

3.3 Analysis: diameter of the confidence regions

As $\boldsymbol{\Omega}$ represents the tightest confidence region for $\boldsymbol{\omega}^*$ given the uncertainty sets $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{Q}}$, it is essential to characterize its geometric scale. We summarize the size of $\boldsymbol{\Omega} \subset \mathbb{R}^K$ by its diameter under $\|\cdot\|_\infty$ and $\|\cdot\|_2$ norms. For a generic set $\mathcal{S} \subset \mathbb{R}^K$, define $\mathrm{diam}_\infty(\mathcal{S}) \equiv \sup_{\boldsymbol{\omega},\boldsymbol{\omega}' \in \mathcal{S}} \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_\infty$ and $\mathrm{diam}_2(\mathcal{S}) \equiv \sup_{\boldsymbol{\omega},\boldsymbol{\omega}' \in \mathcal{S}} \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2$. Since $\boldsymbol{\Omega}_{\mathrm{LP}}$ is the smallest hyperrectangle containing $\boldsymbol{\Omega}$, we have $\mathrm{diam}_\infty(\boldsymbol{\Omega}) = \mathrm{diam}_\infty(\boldsymbol{\Omega}_{\mathrm{LP}}) = \max_{k\in[K]}(\overline{\omega}_k - \underline{\omega}_k)$.

The next theorem provides upper bounds on $\mathrm{diam}_\infty(\boldsymbol{\Omega})$ and $\mathrm{diam}_2(\boldsymbol{\Omega})$ in terms of $\underline{\mathbf{C}}$, $\overline{\mathbf{C}}$, $\underline{\mathbf{q}}$ and $\overline{\mathbf{q}}$. For a generic matrix $\mathbf{A} = (A_{ij})_{i,j\in[K]}$, let $\|\mathbf{A}\|_\infty \equiv \max_{i\in[K]} \sum_{j=1}^K |A_{ij}|$ and $\|\mathbf{A}\|_2 \equiv \sup_{\|\boldsymbol{v}\|_2=1} \|\mathbf{A}\boldsymbol{v}\|_2$. The proof of Theorem 3.5 is provided in Appendix F.5.

Theorem 3.5. *Let $\underline{\mathbf{C}} \in \mathbb{R}^{K\times K}$ be invertible. If $\|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty < 1$, then*

$$\mathrm{diam}_\infty(\boldsymbol{\Omega}) \leq \frac{\|\underline{\mathbf{C}}^{-1}\|_\infty^2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty \|\overline{\mathbf{q}}\|_\infty}{\left(1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty\right)^2} + \frac{\|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_\infty}{1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty}.$$

If $\|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 < 1$, then

$$\mathrm{diam}_2(\boldsymbol{\Omega}) \leq \frac{\|\underline{\mathbf{C}}^{-1}\|_2^2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 \|\overline{\mathbf{q}}\|_2}{\left(1 - \|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2\right)^2} + \frac{\|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_2}{1 - \|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2}.$$

[2]We thank an anonymous reviewer whose comments led us to unearth the connection to the Oettli-Prager Theorem.

Remark 3.6. The assumptions $\|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty < 1$ and $\|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 < 1$ imply that $\mathbf{C}^{-1}$ exists for all $\mathbf{C} \in \boldsymbol{\mathcal{C}}$, and $\mathbf{C} \mapsto \mathbf{C}^{-1}$ is a smooth function on $\boldsymbol{\mathcal{C}}$. These conditions are not stringent when the confidence intervals for $\mathbf{C}^*$ are sufficiently tight, since $\|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|$ shrinks with sample size m while $\|\underline{\mathbf{C}}^{-1}\|$ is bounded whenever $\underline{\mathbf{C}}$ is well-conditioned (e.g., nearly diagonal or diagonally dominant). We illustrate the usage of Theorem 3.5 by deriving an explicit bound in terms of (m, n, K) in Appendix G.

3.4 Downstream task: finite sample prediction set

Benefiting from the confidence region of $\boldsymbol{\omega}^*$, we propose two ways of constructing a prediction set, conformal prediction [36] and probably approximately correct (PAC) prediction [34]. We provide finite-sample guarantees for these prediction sets. In existing works, finite-sample coverage is typically shown only when the true importance weights $\boldsymbol{\omega}^*$ are known, and when $\boldsymbol{\omega}^*$ is replaced by an estimator $\widehat{\boldsymbol{\omega}}$, results are generally asymptotic. We improve these methods by obtaining prediction sets with a prescribed coverage level in finite samples instead of asymptotically, while allowing unknown $\boldsymbol{\omega}^*$ by incorporating a confidence region for $\boldsymbol{\omega}^*$. We show the conformal prediction results here, while presenting the PAC prediction results in Appendix H.

3.4.1 Conformal prediction

We develop an uncertainty-aware conformal prediction set under label shift. We first review the weighted split conformal construction for a fixed candidate weight vector $\boldsymbol{\omega}$, and then incorporate a confidence region for $\boldsymbol{\omega}^*$ to obtain finite-sample coverage when $\boldsymbol{\omega}^*$ is unknown.

We consider the split conformal prediction setting, where the source data is divided into a calibration set $\mathcal{D}_{\text{cali}} = \{z_i = (\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) : i = 1, \ldots, m_1\}$ and a training set $\mathcal{D}_{\text{train}}$, where the nonconformity score $r(\widetilde{\boldsymbol{x}}, \widetilde{y}) \in [0, 1]$ is trained on. The prediction set is constructed using the calibration set $\mathcal{D}_{\text{cali}}$. Let $z_0 = (\widetilde{\boldsymbol{x}}_0, \widetilde{y}_0)$ denote a new target-domain pair, where $\widetilde{\boldsymbol{x}}_0$ is observed and $\widetilde{y}_0 \in [K]$ is the unknown label. Under the label shift assumption, the calibration data set $\mathcal{D}_{\text{cali}}$ and z_0 satisfy the weighted exchangeability condition of Tibshirani et al. [33],

$$q(z_0, z_1, \ldots, z_{m_1}) = p(z_0, z_1, \ldots, z_{m_1}) \prod_{i=0}^{m_1} \omega^*_{\widetilde{y}_i},$$

where $p(z_0, \ldots, z_{m_1}) = p(z_{\sigma(0)}, \ldots, z_{\sigma(m_1)})$ for any permutation $\sigma : \{0, \ldots, m_1\} \to \{0, \ldots, m_1\}$. This weighted exchangeability motivates calibrating $r(\widetilde{\boldsymbol{x}}, \widetilde{y})$ using the weighted empirical distribution of calibration scores. Let $r_i = r(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)$ for $i = 1, \ldots, m_1$. For any candidate $\boldsymbol{\omega}$, we define the level $(1 - \alpha)$ conformal prediction set

$$F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \boldsymbol{\omega}) = \{\widetilde{y}_0 \in [K] : r(\widetilde{\boldsymbol{x}}_0, \widetilde{y}_0) \le \tau_{\text{CP}}(\widetilde{y}_0; \boldsymbol{\omega})\},$$

where $\tau_{\text{CP}}(\widetilde{y}_0; \boldsymbol{\omega})$ is given by

$$\tau_{\text{CP}}(\widetilde{y}_0; \boldsymbol{\omega}) = Q_{1-\alpha}\left(\sum_{i=1}^{m_1} \kappa_{r_i} \frac{\omega_{\widetilde{y}_i}}{\sum_{j=1}^{m_1} \omega_{\widetilde{y}_j} + \omega_{\widetilde{y}_0}} + \kappa_1 \frac{\omega_{\widetilde{y}_0}}{\sum_{j=1}^{m_1} \omega_{\widetilde{y}_j} + \omega_{\widetilde{y}_0}}\right),$$

with κ_r denoting the Dirac measure at r, and $Q_{1-\alpha}(\mu)$ denoting the $(1-\alpha)$-th quantile of a probability measure μ.

When the true importance weight $\boldsymbol{\omega}^*$ is known, the prediction set $F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \boldsymbol{\omega}^*)$ achieves $1 - \alpha$ coverage by Theorem 2 of Podkopaev and Ramdas [25]. However, $\boldsymbol{\omega}^*$ is unknown in practice. Let $\boldsymbol{\Omega}_0$ be a confidence region for $\boldsymbol{\omega}^*$. To obtain a prediction set that remains valid for all $\boldsymbol{\omega} \in \boldsymbol{\Omega}_0$, we enlarge the threshold by taking a supremum:

$$\tau_{\text{CP}}(\widetilde{y}_0; \boldsymbol{\Omega}_0) = \sup_{\boldsymbol{\omega} \in \boldsymbol{\Omega}_0} \tau_{\text{CP}}(\widetilde{y}_0; \boldsymbol{\omega}), \tag{5}$$

and define the robust conformal prediction set

$$F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \boldsymbol{\Omega}_0) = \{\widetilde{y}_0 \in [K] : r(\widetilde{\boldsymbol{x}}_0, \widetilde{y}_0) \le \tau_{\text{CP}}(\widetilde{y}_0; \boldsymbol{\Omega}_0)\}. \tag{6}$$

Compared to the known $\boldsymbol{\omega}^*$ case, the construction in (6) enlarges the prediction set due to not knowing $\boldsymbol{\omega}^*$. Nevertheless, we can still control the prediction level at $1 - \delta - \alpha$, as established in Theorem 3.7. Intuitively, coverage can fail only if $\boldsymbol{\omega}^* \notin \boldsymbol{\Omega}_0$ (with probability at most δ) or if the conformal event fails (with probability α). See Appendix F.6 for the proof.

Theorem 3.7. *Suppose that an independent training set $\mathcal{D}_{\text{train}}$ is used to train the classifier. Under label shift assumption, suppose $\mathcal{D}_{\text{cali}} \sim P^{m_1}$ i.i.d., $(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q$ independent of $\mathcal{D}_{\text{cali}}$, and $\mathbf{\Omega}_0 = \mathbf{\Omega}_0(\mathcal{D}_{\text{src}}, \mathcal{D}_{\text{tgt}})$ satisfies $\mathbb{P}(\boldsymbol{\omega}^* \in \mathbf{\Omega}_0) \geq 1 - \delta$ given the training set. Then*

$$\mathbb{P}\{\widetilde{Y}_0 \in F_{\text{CP}}(\widetilde{\mathbf{X}}_0; \mathbf{\Omega}_0)\} \geq 1 - \delta - \alpha,$$

where $\mathbb{P}$ is taken jointly over $\mathcal{D}_{\text{cali}}$, $(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0)$, $\mathcal{D}_{\text{src}}$, and $\mathcal{D}_{\text{tgt}}$, conditional on $\mathcal{D}_{\text{train}}$.

Remark 3.8 (Exact guarantee *vs.* Approximation of $\tau_{\text{CP}}(\widetilde{y}_0; \mathbf{\Omega}_0)$). Theorem 3.7 applies to the *exact* supremum $\tau_{\text{CP}}(\widetilde{y}_0; \mathbf{\Omega}_0)$ in (5). Computations of $\tau_{\text{CP}}(\widetilde{y}_0; \mathbf{\Omega}_0)$ for different $\mathbf{\Omega}_0$ require different optimization techniques. When using $\mathbf{\Omega}$ in the experiments, we approximate $\tau_{\text{CP}}(\widetilde{y}_0; \mathbf{\Omega})$ by a Monte Carlo (MC) maximum over points sampled from $\mathbf{\Omega}$. This MC threshold can underestimate the exact supremum, so the MC-based prediction set *does not itself inherit* the guarantee of Theorem 3.7, and the reported results for $\mathbf{\Omega}$ are empirical.

Additionally, the set construction is monotone in the weight region: smaller confidence regions yield smaller prediction sets at the same level. This property, together with the comparison between $\mathbf{\Omega}_{\text{LP}}$ and $\mathbf{\Omega}_{\text{GE}}$ in Corollary 3.4, is summarized in Theorem 3.9. The proof is provided in Appendix F.7.

Theorem 3.9. *If $\mathbf{\Omega}_1 \subset \mathbf{\Omega}_2$, then $F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_1) \subset F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_2)$ for all $\widetilde{\boldsymbol{x}}_0$. In particular, $F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_{\text{LP}}) \subset F_{\text{CP}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_{\text{GE}})$.*

4 Experiments

We evaluate the empirical performance of MaC-LP through standard benchmarks and real-world domain adaptation. Across both settings, we aim to (i) demonstrate the *tightness* of confidence regions for the importance weights established in Section 3.2 in comparison to the interval-based Gaussian elimination approach [29]; (ii) assess the practical impact of MaC-LP on *conformal prediction*, which serves as a representative task of MaC-LP for uncertainty quantification.

4.1 Experimental design

Using standard benchmark datasets, text data AGNews [43], image data MNIST [16] and CIFAR-10 [15], and text-image news classification data N24News [37], we verify our theoretical guarantees while controlling potential confounding factors.

We split each dataset into two parts: one part is *training pool* for classifier development, the other part is *analysis set* for simulating label shift and downstream applications. The precise split ratios are summarized in Table S1. We train 10 classifiers for each dataset, the details are in Appendix J.1.

Next, we simulate label shift by generating the label proportion vectors $\mathbf{v}_p$ and $\mathbf{v}_q$ while maintaining invariant class-conditional distributions. We define an experimental grid of $7 \times 8 = 56$ settings by varying shift intensity and sample size. For each of the 10 classifiers associated with a benchmark, we perform 100 independent trials, resulting in 1,000 repetitions per setting to ensure statistical significance. All samples are shuffled to eliminate ordering bias prior to analysis.

Label shift simulation. We generate the source and target label proportion vectors $\mathbf{v}_p$ and $\mathbf{v}_q$ from a Dirichlet distribution. The source vector is sampled as $\mathbf{v}_p \sim \text{Dir}(\mathbf{1}_K \cdot 10^4)$, providing a near-uniform baseline. The target vector is sampled as $\mathbf{v}_q \sim \text{Dir}(\mathbf{1}_K \cdot 10^a)$, where the concentration parameter $a \in \{-3, -2, -1, 0, 1, 2, 3\}$. Smaller values of a yield higher sparsity across class labels, resulting in more extreme shifts between the source and target distributions.

Data sampling procedure. For each repetition, the synthetic datasets are generated from the analysis set using a multinomial stratified sampling procedure. Specifically, given a label proportion vector $\mathbf{v}$ and a sample size c, we sample a count vector $\mathbf{c}$ from $\text{Multinomial}(c, \mathbf{v})$, then for each $k \in [K]$, we draw c_k observations. Calibration Set ($\mathcal{D}_{\text{cali}}$): We draw a source domain calibration set by the stratified sampling approach without replacement using $c = 5,000$ and $\mathbf{v} = \mathbf{v}_p$ from the analysis set. This calibration set is independent of the following two datasets. Test Set ($\mathcal{D}_{\text{test}}$): We draw a target domain test set by the stratified sampling approach with replacement using $c = 2,500$ and $\mathbf{v} = \mathbf{v}_q$ from the analysis set. Estimation Sets ($\mathcal{D}_{\text{src}}, \mathcal{D}_{\text{tgt}}$): We draw source domain dataset by the same stratified sampling approach with replacement using c from 1000 to 8000, step size of

1000 and $\mathbf{v} = \mathbf{v}_p$ from the analysis set. Further, we draw target domain dataset by the same protocol except using $\mathbf{v} = \mathbf{v}_q$ from the analysis set.

4.2 Results

4.2.1 Objective (i): tightness of confidence regions

Similar to Si et al. [29], we compute the $1 - \delta/(K^2 + K)$ Clopper-Pearson [8] interval for every element in $\mathbf{C}$ and $\mathbf{q}$. Here, we set $\delta = 1/5,000$ (matching $|\mathcal{D}_{\text{cali}}|$). Following Section 3.2, we compute $\mathbf{\Omega}$ and $\mathbf{\Omega}_{\text{LP}}$, and following Appendix I, we compute $\mathbf{\Omega}_{\text{GE}}$.

LP yields shorter intervals than GE. We compute the marginal length of $\mathbf{\Omega}_{\text{LP}}$ as $l_{k,\text{LP}} = \overline{\omega}_k - \underline{\omega}_k$ for all $k \in [K]$, and compute the marginal length of $\mathbf{\Omega}_{\text{GE}}$ analogously, denoted by $l_{k,\text{GE}}$ for all $k \in [K]$. Next, we compute the $\log_2$ length ratio as $r_k = \log_2(l_{k,\text{GE}}/l_{k,\text{LP}})$ for all $k \in [K]$. A one-sample t-test was conducted on $\{r_k\}_{k=1}^K$, where the null hypothesis is that the $\log_2$ ratio is less than or equal to zero. In the left panel of Figure 2, we plot the mean $\log_2$ length ratio, calculated as $K^{-1}\sum_{k=1}^K r_k$, against the statistical significance expressed as $-\log_{10}(p\text{-value})$ for all 56 label shift settings for N24News.

In the left panel of Figure 2, LP consistently yields shorter marginal confidence intervals than the GE baseline. The t-test results remain well above the $-\log_{10}(0.05)$ threshold, indicating that the reduction in uncertainty is statistically significant for N24News. Notably, we observe the highest $-\log_{10}(p\text{-value})$ and length ratios, suggesting that LP is particularly effective when dealing with complex classifiers where joint class dependencies are more pronounced. The similar results are shown in Appendix K.2 for AGNews, MNIST, and CIFAR-10.

MaC captures finer dependency than marginal bounds (LP). We use a Monte Carlo method to quantify the ratio of the volumes of $\mathbf{\Omega}$ and $\mathbf{\Omega}_{\text{LP}}$. Specifically, we sample 10,000 points uniformly from $\mathbf{\Omega}_{\text{LP}}$ to represent $\text{volume}(\mathbf{\Omega}_{\text{LP}})$, and then count the number of points that satisfy the matrix constraints to represent $\text{volume}(\mathbf{\Omega})$. For N24News, we plot a bar chart of the mean ratio $\text{volume}(\mathbf{\Omega})/\text{volume}(\mathbf{\Omega}_{\text{LP}})$ over 1,000 repetitions (presented as a percentage), and the error bars represent the standard deviation in the right panel of Figure 2.

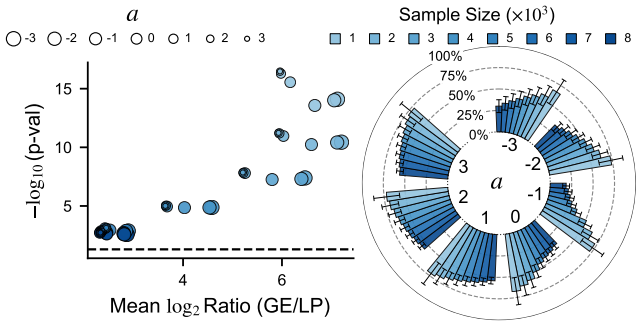


Figure 2: N24News. (Left) Volcano plot of the interval length ratios (GE vs. LP) and their corresponding t-test significance levels. (Right) Volume ratio of $\mathbf{\Omega}$ to $\mathbf{\Omega}_{\text{LP}}$ across varying shift intensities a and sample sizes.

As shown in the bar plots in Figure 2, the volume of $\mathbf{\Omega}$ is consistently a small fraction of the $\mathbf{\Omega}_{\text{LP}}$ across all shift intensities and all sample sizes. The region $\mathbf{\Omega}$, by considering dependent structure of the linear system, inherently excludes vast regions of "infeasible" weight combinations that violate the matrix constraints. The region $\mathbf{\Omega}$ maintains a tight polytope even under extreme label shift (lower a). This volumetric tightness is the primary driver for improved downstream efficiency, as it directly restricts the search space for the weighted quantiles in the conformal prediction phase. The similar results for AGNews, MNIST, and CIFAR-10 are in Appendix K.2.

4.2.2 Objective (ii): assessment of downstream task

To assess the practical utility of MaC-LP, we apply the confidence regions $\mathbf{\Omega}$ and $\mathbf{\Omega}_{\text{LP}}$ to the weighted conformal prediction (WCP) framework [25]. For comparison, we evaluate the WCP procedure using the $\mathbf{\Omega}_{\text{GE}}$, the Oracle vector $\boldsymbol{\omega}^*$, and the point estimate $\widehat{\boldsymbol{\omega}}$ derived from the BBSE method [17]. These experiments are implemented across 56 label shift settings for each benchmark and are repeated 1,000 times to ensure statistical robustness.

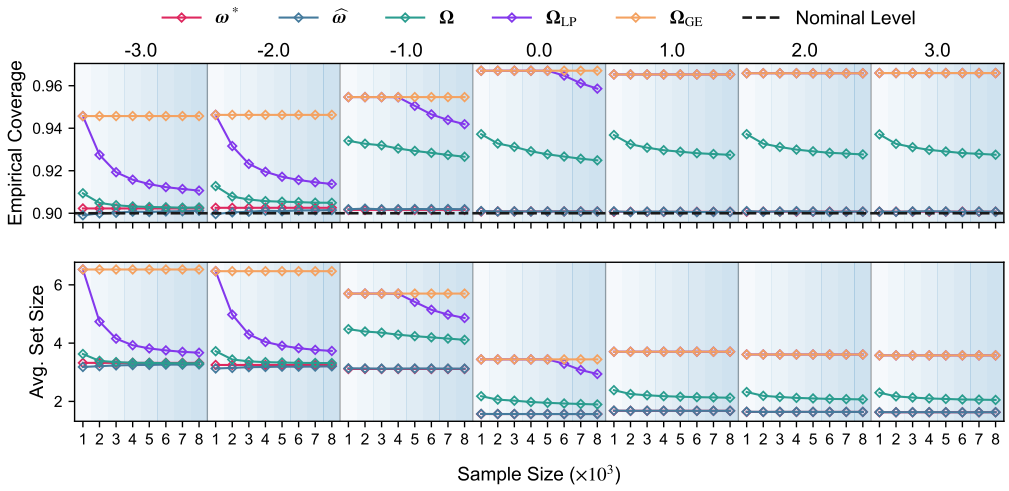


Figure 3: Evaluation of weighted conformal prediction on N24News. The top panel displays empirical coverage across varying shift intensities (a) and sample sizes (n). The bottom panel shows average prediction set sizes.

For each label shift setting, we utilize an independent calibration set $\mathcal{D}_{\text{cali}}$ ($m_1 = 5,000$) to compute non-conformity scores (details in Appendix J.3). We construct prediction sets $F_{\text{CP}}(\boldsymbol{x}_t; \star)$ for the test instances in $\mathcal{D}_{\text{test}}$ by calculating weighted quantiles over the methods $\star \in \{\boldsymbol{\omega}^*, \widehat{\boldsymbol{\omega}}, \mathbf{\Omega}, \mathbf{\Omega}_{\text{LP}}, \mathbf{\Omega}_{\text{GE}}\}$. The confidence levels for $\mathbf{\Omega}$, $\mathbf{\Omega}_{\text{LP}}$, and $\mathbf{\Omega}_{\text{GE}}$ are fixed at $1-\delta$ with $\delta = 1/5,000$. This specific choice of δ ensures that weight estimation error does not compromise the target coverage level $1-\alpha$, where $\alpha = 0.1$. The thresholds for $\mathbf{\Omega}_{\text{LP}}$ and $\mathbf{\Omega}_{\text{GE}}$ are computed exactly, whereas the threshold for $\mathbf{\Omega}$ is approximated by the maximum over 1,000 MC samples from $\mathbf{\Omega}$ (Remark 3.8). We evaluate the impact of importance weights estimation on the trade-off between statistical validity and predictive efficiency using two standard metrics, Empirical Coverage and Average Set Size.

MaC enhances efficiency while maintaining reliability. Figure 3 illustrates the evaluation metrics for the N24News benchmark. Among all methods, $\mathbf{\Omega}$ consistently maintains empirical coverage above the nominal 90% level. While $\mathbf{\Omega}$ typically yields lower coverage than the overly conservative $\mathbf{\Omega}_{\text{LP}}$ and $\mathbf{\Omega}_{\text{GE}}$ regions, it remains empirically above the nominal level. Notably, the point estimate $\widehat{\boldsymbol{\omega}}$ suffers from undercoverage under extreme shifts ($a \in \{-3, -2\}$), thereby losing statistical validity.

Regarding efficiency, the Oracle specification $\boldsymbol{\omega}^*$ provides the theoretical lower bound for prediction set size. $\mathbf{\Omega}$ maintains a stable trajectory that closely tracks the Oracle performance, whereas $\mathbf{\Omega}_{\text{LP}}$ and $\mathbf{\Omega}_{\text{GE}}$ exhibit substantial efficiency loss because their feasible regions include weight combinations prohibited by the matrix constraints. We further observe that while $\widehat{\boldsymbol{\omega}}$ may produce smaller prediction sets in certain label shift settings, these results are unreliable as they coincide with a failure to maintain nominal coverage. Similar trends are observed across all other benchmarks, as detailed in Appendix K.3.

To evaluate real-world robustness, we test MaC-LP on the nuImages dataset [6] by constructing a domain adaptation scenario, treating data from Singapore as the source domain and images captured in Boston as the target domain (details in Appendix J.2). This geographic shift presents a complex environment where the label shift assumption may be violated, allowing us to evaluate the resilience of MaC-LP in the presence of naturalistic distribution shifts.

4.2.3 Evaluation on real-world data (nuImages)

In Figure 4(a), GE still produces much wider confidence intervals for importance weights especially when sample size is small, where the uncertainty is larger. Over the all sample sizes, $\mathbf{\Omega}$ is consistently smaller than $\mathbf{\Omega}_{\text{LP}}$ as shown in Figure 4(b).

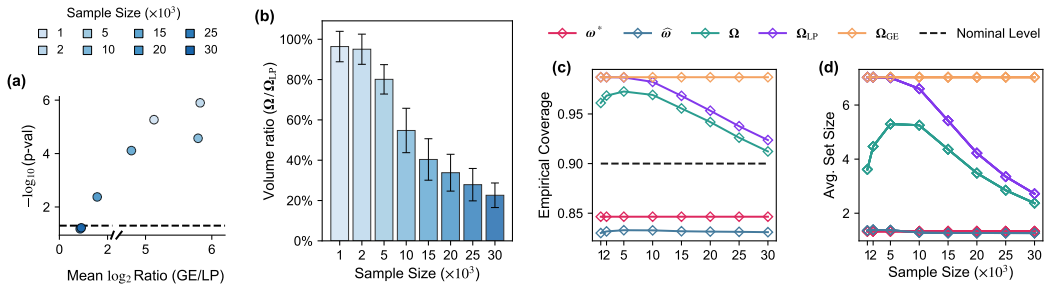


Figure 4: nuImages. (a) Volcano plot of the interval length ratios (GE vs. LP) and their corresponding t-test significance levels. (b) Volume ratio of $\mathbf{\Omega}$ to $\mathbf{\Omega}_{\mathrm{LP}}$ across varying sample sizes. (c) Empirical coverage across varying sample sizes (n). (d) Average prediction set sizes.

For the conformal prediction task, we use the 90% nominal prediction level. From Figure 4(c), we can see that even applying "true" importance weights, the coverage probability is still under 90%, indicating much more complicated data structure and much more noise in the real-world dataset than that in the benchmarks. However, $\mathbf{\Omega}$, $\mathbf{\Omega}_{\mathrm{LP}}$, and $\mathbf{\Omega}_{\mathrm{GE}}$ achieve coverage probability above the nominal level. In Figure 4(d), $\mathbf{\Omega}$ provides the best balance between prespecified coverage level and the prediction set size. Notably, applying the point estimation, $\widehat{\boldsymbol{\omega}}$, remains to have deteriorative performance. Since label shift is likely violated here, Theorem 3.7 does not apply and these results are empirical rather than an exact guarantee: the confidence regions also absorb part of the label shift assumption violation, a buffer that shrinks as the sample size grows, which is illustrated by a controlled experiment in Appendix K.6.

4.3 Runtime and implementation analysis

In Appendix K.4, on synthetic data with K up to 500, $\mathbf{\Omega}_{\mathrm{LP}}$ takes 0.046 seconds at $K = 100$ and 1.72 seconds at $K = 500$ (0.030 seconds and 0.263 seconds for $\mathbf{\Omega}_{\mathrm{GE}}$), while the mean $\log_2$ length ratio of GE to LP intervals grows to about 10.9 for $K \geq 200$. A near-zero target proportion gives $\omega_k^* \to 0$, which is investigated in $a \in \{-3, -2\}$ settings, and the LP solver had no failures. Rare source classes and empty or unbounded regions are discussed in Appendix K.5.

5 Discussion

In this work, we propose *Matrix Constraints Linear Programming* (MaC-LP) for label shift, a novel finite-sample method for uncertainty-aware inference. Specifically, MaC captures the most refined finite-sample valid confidence regions given the elementwise confidence intervals of $\mathbf{C}^*$ and $\mathbf{q}^*$. Our experimental results in conformal prediction highlight that MaC-LP is practically effective in addressing uncertainty quantification problems. Furthermore, the MaC construction in Theorem 3.2 provides an explicit convex representation of the confidence region $\mathbf{\Omega}$ without requiring point estimates of importance weights. This convexity allows MaC to be integrated into domain adaptation tasks as a convex optimization problem, bypassing LP-related overhead while maintaining precision.

As limitations, our guarantees require the exact label shift assumption, the optimality of $\mathbf{\Omega}$ is relative to the input $(\underline{\mathbf{C}}, \overline{\mathbf{C}}, \underline{\mathbf{q}}, \overline{\mathbf{q}})$, and the MC approximation of computing $\tau_{\mathrm{CP}}(\widetilde{y}_0; \mathbf{\Omega})$ does not inherit Theorem 3.7. Nevertheless, MaC-LP enriches the label shift toolbox by providing statistical guarantees. Further discussions regarding optimality, computational complexity, classifier robustness, application scope, and scalability for large number of classes are provided in Appendix B.

Code availability. The python package is available at `https://github.com/carayms/maclp`.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on*

Machine Learning, pages 222–232. PMLR, 2020.

[2] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019.

[3] Helge Blaker. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*, 28(4):783–798, 2000.

[4] Colin R Blyth and Harold A Still. Binomial confidence intervals. *Journal of the American Statistical Association*, 78(381):108–116, 1983.

[5] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, 16(2):101–133, 2001.

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[7] George Casella. Refining binomial confidence intervals. *Canadian Journal of Statistics*, 14(2): 113–129, 1986.

[8] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

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

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[19] Luca Masserano, Alex Shen, Michele Doro, Tommaso Dorigo, Rafael Izbicki, and Ann B Lee. Classification under nuisance parameters and generalized label shift in likelihood-free inference. *arXiv preprint arXiv:2402.05330*, 2024.

[20] Ramon E Moore, R Baker Kearfott, and Michael J Cloud. *Introduction to interval analysis*. SIAM, 2009.

[21] Arnold Neumaier. *Interval methods for systems of equations*. Number 37. Cambridge university press, 1990.

[22] Werner Oettli and William Prager. Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. *Numerische Mathematik*, 6(1): 405–409, 1964.

[23] Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. *arXiv preprint arXiv:2106.09848*, 2021.

[24] 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*, pages 27907–27947. PMLR, 2023.

[25] Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR, 2021.

[26] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020.

[27] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[29] Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. PAC prediction sets under label shift. *arXiv preprint arXiv:2310.12964*, 2023.

[30] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Måns Thulin. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, 8(1), January 2014. ISSN 1935-7524. doi: 10.1214/14-ejs909.

[32] Qinglong Tian, Xin Zhang, and Jiwei Zhao. ELSA: Efficient label shift adaptation through the lens of semiparametric models. In *International Conference on Machine Learning*, pages 34120–34142. PMLR, 2023.

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

[34] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[35] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

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

[37] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24news: A new dataset for multimodal news classification. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.729.

[38] Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online adaptation to label distribution shift. *Advances in Neural Information Processing Systems*, 34:11340–11351, 2021.

[39] Ruihan Wu, Siddhartha Datta, Yi Su, Dheeraj Baby, Yu-Xiang Wang, and Kilian Q Weinberger. Online feature updates improve online (generalized) label shift adaptation. *Advances in Neural Information Processing Systems*, 37:106924–106954, 2024.

[40] Zhiyuan Wu, Changkyu Choi, Xiangcheng Cao, Volkan Cevher, and Ali Ramezani-Kebrya. Addressing label shift in distributed learning via entropy regularization. *arXiv preprint arXiv:2502.02544*, 2025.

[41] Chao Xu, Xijia Tang, and Chenping Hou. Heterogeneous label shift: Theory and algorithm. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=BY3SSEVcjV.

[42] Edvin Listo Zec, Adam Breitholtz, and Fredrik D Johansson. Overcoming label shift in targeted federated learning. *arXiv preprint arXiv:2411.03799*, 2024.

[43] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.

A Workflow of MaC-LP

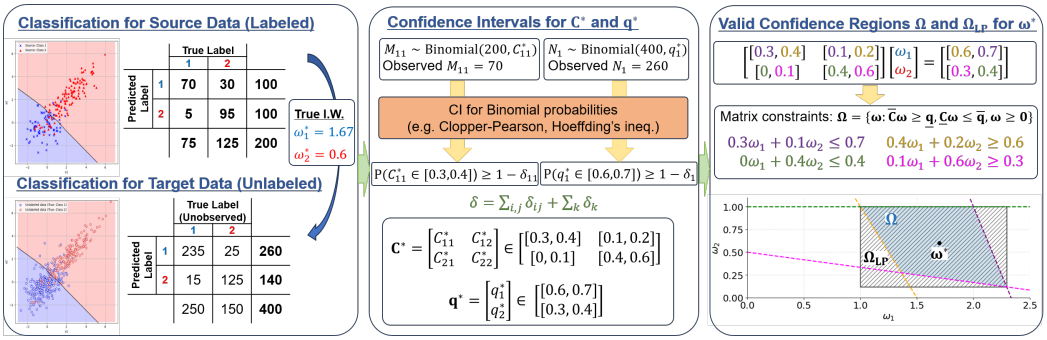


Figure S1: Workflow of MaC-LP in a binary classification toy example: (1) apply a classifier to both source and target datasets; (2) obtain elementwise confidence intervals for $\mathbf{C}^*$ and $\mathbf{q}^*$; (3) construct confidence regions $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_{\text{LP}}$. We can observe that $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_{\text{LP}}$ capture the true importance weigths $\boldsymbol{\omega}^*$.

B Extended discussion

Scope of optimality: simplex and multinomial constraints. The optimality of $\boldsymbol{\Omega}$ holds given the elementwise confidence intervals $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{Q}}$, and it does not exploit the facts that $\sum_{i,j} C_{ij}^* = 1$, $\sum_k q_k^* = 1$, or that $\mathbf{M}$ and $\mathbf{N}$ are jointly multinomial. Imposing such constraints on $(\mathbf{C}, \mathbf{q})$ can make the region smaller or equal (Figure S2). However, because $\boldsymbol{\omega} = \mathbf{C}^{-1}\mathbf{q}$ is a nonlinear map, the resulting region need not be convex and need not admit a linear characterization as in Theorem 3.2: the row-by-row construction in the proof of Theorem 3.2 is no longer available once the rows of $\mathbf{C}$ are coupled. Figure S2 illustrates this on the two-class example of Figure S1: the weights generated by admissible pairs satisfying the simplex constraints form a strict, non-polytope subset of $\boldsymbol{\Omega}$. Characterizing and computing such regions exactly is an interesting open problem; we view the construction in Theorem 3.2 as a balance between tightness and mathematical and computational tractability.

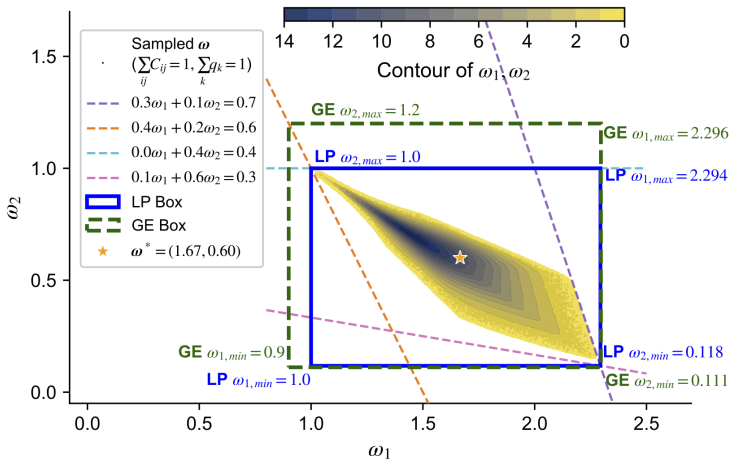


Figure S2: Two-class example of Figure S1. The dashed lines are the four constraints of $\boldsymbol{\Omega}$ in Theorem 3.2; the blue and green boxes are $\boldsymbol{\Omega}_{\text{LP}}$ and $\boldsymbol{\Omega}_{\text{GE}}$. The shaded points are the weights $\boldsymbol{\omega}$ generated by admissible pairs $(\mathbf{C}, \mathbf{q}) \in \boldsymbol{\mathcal{C}} \times \boldsymbol{\mathcal{Q}}$ that additionally satisfy $\sum_{i,j} C_{ij} = 1$ and $\sum_k q_k = 1$, with their density contours.

Computational complexity and parallelization for $\boldsymbol{\Omega}_{\text{LP}}$. While the theoretical complexity for solving the linear program for $\boldsymbol{\Omega}_{\text{LP}}$ is $\mathcal{O}(K^4)$, where K is the number of classes, this does not

represent a practical bottleneck for standard classification tasks because the optimization process is highly amenable to parallel computing. Furthermore, the convexity of $\mathbf{\Omega}$ ensures that it can be handled by a wide range of efficient off-the-shelf solvers. Wall-clock runtimes for K up to 500 are reported in Appendix K.4.

Robustness to classifier quality and calibration. A significant advantage of MaC-LP is its resilience to the performance of the underlying black-box classifier. As illustrated in Figure S3 and Table S4, our framework maintains superior performance regardless of whether the base classifiers are well-calibrated or possess high accuracy. This decoupling suggests that MaC-LP effectively extracts importance weight information even from suboptimal classifiers, providing a reliable framework in practical scenarios where perfect calibration is unattainable.

Coverage guarantees *vs.* point predictions. Standard domain adaptation (DA) techniques primarily focus on improving point prediction accuracy, often neglecting the quantification of predictive uncertainty. In contrast, our work centers on providing $1-\alpha$ coverage guarantees. Even when the predictive model is limited in capacity, our constructed confidence regions ensure that the true label is contained within the set with prespecified probability. This statistical safety guarantee provides more information of the $Y|\mathbf{X}$ distribution in the target domain than simple expectation-based estimates, $\mathbb{E}_Q(Y|\mathbf{X})$. Accordingly, the *downstream efficiency* claimed in this paper refers to set-valued inference, i.e., smaller prediction sets at a guaranteed coverage level. When considering prediction accuracy in the target domain, the confidence region does not automatically improves point-prediction metrics; turning $\mathbf{\Omega}$ into a point prediction requires an additional criterion over $\mathbf{\Omega}$ (e.g., a minimax rule $\min_{\hat{y}} \max_{\boldsymbol{\omega}\in\mathbf{\Omega}} R(\hat{y};\boldsymbol{\omega})$ for a weighted risk), to which the coverage guarantee of $\mathbf{\Omega}$ does not transfer. We leave this direction to future work.

Comparison with asymptotic and bootstrap uncertainty quantification. A direct empirical comparison with other label shift estimators requires each of them to come with an implementable, data-dependent confidence region for $\boldsymbol{\omega}^*$. Most existing estimators are developed for point estimation. BBSE [17] and RLLS [2] provide high-probability error bounds, but these contain unspecified constants and cannot be evaluated as confidence regions from the observed data. ELSA [32] establishes asymptotic normality with an explicit limiting covariance, but does not provide a feasible covariance estimator or an implementation of the corresponding region. A nonparametric bootstrap could produce approximate intervals, but it approximates both the finite-sample problem and its solution, requires hundreds to thousands of refits, and leaves open how to turn bootstrap draws into a multi-dimensional region of prescribed level (e.g., a minimum-volume region would require multivariate density estimation). In all cases the result is an asymptotic or approximate procedure rather than the exact finite-sample guarantee studied here. We regard a systematic comparison with such asymptotic and bootstrap-based regions as future work.

Scalability to large class dimensions. For tasks involving a massive number of labels, the whole model scalability may have negative impact on tractability. In such instances, we recommend a hierarchical approach to decompose the label space, for example, incorporating with a decision tree model. While we treat the optimization of these hierarchies as an extension for future study, the current conclusion of MaC-LP remains highly effective for the experimented datasets.

C Broader impacts

This work establishes a mathematically rigorous framework to achieve operational efficiency under principled safety bounds in machine learning. However, as a purely theoretical framework, MaC-LP has not been verified in complex real-world environments thoroughly. We caution practitioners that these theoretical guarantees may not automatically hold in unverified practical deployments. Comprehensive empirical testing and safety validations are strictly required before applying this method to high-stakes applications.

D Details of related work

Estimation for importance weights under label shift. Although our work focuses on constructing finite-sample confidence regions for the importance weights, we briefly review existing methods for point estimation of $\boldsymbol{\omega}^*$ under label shift. Recall that we are given finite samples $\mathcal{D}_{\mathrm{src}} = \{(\boldsymbol{x}_s, y_s)\}_{s=1}^m$ and $\mathcal{D}_{\mathrm{tgt}} = \{\boldsymbol{x}_{m+t}\}_{t=1}^n$ from source and target domain, respectively. In the source domain, for $i, j \in [K]$, let $M_{ij} = \sum_{s=1}^m I\{g(\boldsymbol{x}_s) = i,\ y_s = j\}$, where $I\{\cdot\}$ is the indicator function. Denote $\mathbf{M} = (M_{ij})_{i,j\in[K]}$ which is a $K \times K$ matrix. Note that $\sum_{i=1}^K \sum_{j=1}^K M_{ij} = m$. Then $\mathrm{vec}(\mathbf{M}) \mid \mathbf{C}^* \sim \mathrm{Multinomial}\{m, \mathrm{vec}(\mathbf{C}^*)\}$, where vec denotes the vectorization of a matrix. In the target domain, for $k \in [K]$, let $N_k = \sum_{t=1}^n I\{g(\boldsymbol{x}_{m+t}) = k\}$, and denote $\mathbf{N} = (N_1, \ldots, N_K)^\top$. Similarly, we have $\sum_{k=1}^K N_k = n$ and $\mathbf{N} \mid \mathbf{q}^* \sim \mathrm{Multinomial}(n, \mathbf{q}^*)$. Motivated by the identity $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$ derived in Section 3.1, a natural approach to estimate $\boldsymbol{\omega}^*$ is via plug-in solutions. When $\mathbf{C}^*$ is invertible, solving the linear system $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$ yields a valid estimator of the importance weights. In practice, since $\mathbf{C}^*$ and $\mathbf{q}^*$ are unknown, one replaces them by their sample analogues $\widehat{\mathbf{C}} = (\widehat{C}_{ij})_{i,j\in[K]}$ and $\widehat{\mathbf{q}} = (\widehat{q}_1, \ldots, \widehat{q}_K)^\top$, where $\widehat{C}_{ij} = M_{ij}/m$ and $\widehat{q}_k = N_k/n$. The resulting estimator $\widehat{\boldsymbol{\omega}}_{\mathrm{BBSE}} = \widehat{\mathbf{C}}^{-1}\widehat{\mathbf{q}}$ is referred to as black box shift estimation (BBSE) [17].

Regularized learning of label shift (RLLS) [2] modifies BBSE by stabilizing the plug-in solution via regularization. Specifically, RLLS computes $\widehat{\boldsymbol{\omega}}_{\mathrm{RLLS}}$ as a minimizer of

$$\widehat{\boldsymbol{\omega}}_{\mathrm{RLLS}} = \arg\min_{\boldsymbol{\omega}} \ \|\widehat{\mathbf{C}}\boldsymbol{\omega} - \widehat{\mathbf{q}}\|_2^2 + \lambda\|\boldsymbol{\omega} - \mathbf{1}_K\|_2^2,$$

where $\mathbf{1}_K$ is the length-K vector of ones. The regularizer $\|\boldsymbol{\omega} - \mathbf{1}_K\|_2^2$ shrinks $\boldsymbol{\omega}$ toward $\mathbf{1}_K$, corresponding to no label shift. In some implementations of BBSE and RLLS, additional constraints such as $\boldsymbol{\omega} \geq 0$ and a normalization condition (ensuring that the implied target label distribution sums to one) may also be imposed.

More generally, moment-based methods can be formulated by matching reweighted source moments to target moments. Efficient label shift adaptation (ELSA) [32] estimates $\boldsymbol{\omega}^*$ by solving the moment-matching equation

$$\frac{1}{m}\sum_{s=1}^m h(\boldsymbol{x}_s)\widehat{\omega}_{\mathrm{ELSA},y_s} = \frac{1}{n}\sum_{t=1}^n h(\boldsymbol{x}_{m+t}),$$

equivalently,

$$\sum_{j=1}^K \frac{1}{m}\sum_{s=1}^m h(\boldsymbol{x}_s)\mathbf{1}\{y_s = j\}\widehat{\omega}_{\mathrm{ELSA},j} = \frac{1}{n}\sum_{t=1}^n h(\boldsymbol{x}_{m+t}).$$

Moreover, ELSA constructs an optimal choice of the moment function h (within an appropriate class) so that the resulting estimator $\widehat{\boldsymbol{\omega}}_{\mathrm{ELSA}}$ attains the semiparametric efficiency bound for $\boldsymbol{\omega}^*$ under label shift.

Complementary to these moment-based approaches, likelihood-based methods estimate the target label distribution directly under the label shift model. Maximum likelihood label shift (MLLS) uses a well-calibrated soft-max classifier to obtain an estimate $\widehat{p}(y)$ of the source label distribution and then estimates the target label distribution $\widehat{q}(y)$ by maximum likelihood, typically via an EM algorithm [1]. The resulting estimator for $\boldsymbol{\omega}^*$ is $\widehat{\boldsymbol{\omega}}_{\mathrm{MLLS}} = (\widehat{\omega}_{\mathrm{MLLS},1}, \ldots, \widehat{\omega}_{\mathrm{MLLS},K})^\top$, where $\widehat{\omega}_{\mathrm{MLLS},y} = \widehat{q}(y)/\widehat{p}(y)$ for $y \in [K]$.

Recent works on label shift in extended settings. The label shift setting continues to receive active attention. In particular, recent work incorporates label shift into online, federated/distributed, and generalized modeling settings. Recent work studies online label shift, where unlabeled target samples arrive sequentially and the learner must update adaptation or calibration procedures over time [38, 39]. Label shift has also been considered in federated learning, where clients may exhibit heterogeneous label distributions and the goal is to improve performance in the target domain [42]. Label shift has further been studied in multi-node distributed learning, where each node may have access to both source and target data under heterogeneous label distributions [40]; for example, Wu et al. [40] use an entropy-based regularizer to stabilize importance-weight estimation across nodes.

Masserano et al. [19] introduce a generalized label shift model of the form $p(\boldsymbol{x} \mid y, \nu) = q(\boldsymbol{x} \mid y, \nu)$, where ν denotes nuisance parameters that affect the class-conditional distribution. Label shift has also been studied under structured heterogeneity, including settings where the source and target feature representations differ (e.g., cross-lingual text) but the label-shift structure is leveraged for adaptation [41]. When these settings can be reduced to (or approximated by) the standard label-shift identity $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$, our framework provides finite-sample confidence regions for $\boldsymbol{\omega}^*$ as an alternative to point estimation in such pipelines.

Uncertainty propagation through linear systems. A common approach to finite-sample uncertainty quantification for parameters defined by linear identities is to treat the estimated coefficients and right-hand side as interval-valued objects and propagate these intervals through interval linear system solvers. In this literature, interval Gaussian elimination is a standard procedure for bounding the solution set of $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ when $(\mathbf{A}, \boldsymbol{b})$ are only known within uncertainty sets, and related alternatives include Krawczyk method [21, 20]. However, interval Gaussian elimination repeatedly applies subtraction and division that reverse endpoints, which yields overly conservative bounds for $\boldsymbol{x}$ [20]. In the label shift setting, Si et al. [29] adopt this interval Gaussian elimination procedure by propagating binomial confidence intervals through the updates of the procedure to obtain confidence intervals for importance weights; details are provided in Appendix I. These developments motivate propagating uncertainty by optimizing directly over the underlying constraints, rather than through interval Gaussian elimination or direct inversion.

Finite-sample prediction under distribution shift. Finite-sample prediction sets are a common downstream objective for importance weights under distribution shift, since uncertainty-aware decisions often require coverage guarantees that hold in the target domain.

Conformal prediction provides a general route to constructing prediction sets with finite-sample coverage guarantees [36], and a growing literature studies how to modify conformal procedures when the test distribution differs from the training or calibration distribution. For example, Tibshirani et al. [33] develop weighted conformal prediction for covariate shift by incorporating importance weights into the conformal calibration step, while subsequent work considers related shift structures such as label shift and associated weighting-based corrections [25, 24], as well as adaptive conformal inference that updates calibration online under sequential distribution shift [12].

Complementary to conformal prediction, probably approximately correct (PAC) prediction sets provide coverage guarantees that hold with high probability conditional on the calibration sample, and have been extended to covariate shift and label shift settings by explicitly incorporating importance weights to ensure target-domain reliability [23, 29]. These connections emphasize that tight, finite-sample-valid uncertainty quantification for importance weights can have a direct impact on the informativeness of prediction sets under distribution shift.

D.1 The connection to Oettli–Prager theorem

The Oettli–Prager theorem [22] provides strong theoretical support for our formulation, and the theorem validates that set-based linear systems under interval uncertainty can be exactly characterized by closed-form linear inequalities, avoiding the existential search over parameters $(\exists \mathbf{C}, \mathbf{q})$.

However, Theorem 3.2 differs fundamentally in its *problem formulation* and *inequality construction*:

- Writing $\mathbf{C}_c = (\underline{\mathbf{C}} + \overline{\mathbf{C}})/2$, $\Delta = (\overline{\mathbf{C}} - \underline{\mathbf{C}})/2$, $\mathbf{q}_c = (\underline{\mathbf{q}} + \overline{\mathbf{q}})/2$ and $\boldsymbol{d} = (\overline{\mathbf{q}} - \underline{\mathbf{q}})/2$, the Oettli–Prager theorem assumes nominal midpoints $(\mathbf{C}_c, \mathbf{q}_c)$ perturbed by tolerances $(\Delta, \boldsymbol{d})$. In contrast, there are no prespecified midpoints $(\mathbf{C}_c, \mathbf{q}_c)$ in our setting; the boundaries $[\underline{\mathbf{C}}, \overline{\mathbf{C}}]$ and $[\underline{\mathbf{q}}, \overline{\mathbf{q}}]$ are the *primary*, data-driven objects obtained directly from statistical inequality bounds. In addition, defining midpoints $(\mathbf{C}_c, \mathbf{q}_c)$ would be a purely artificial construction rather than a reflection of the underlying data-generating process. For example, the Clopper–Pearson interval in our setting is defined through Beta quantiles, which are rarely symmetric about the point estimate, so a midpoint has no natural meaning.
- In one dimension coordinate, two intervals are overlapped if $(x_{\text{start}} - y_{\text{end}})(x_{\text{end}} - y_{\text{start}}) \leq 0$, for $[x_{\text{start}}, x_{\text{end}}]$ and $[y_{\text{start}}, y_{\text{end}}]$. Inspired by this, we characterized $\boldsymbol{\Omega}$ by finding overlaps between $[\underline{\mathbf{C}}\boldsymbol{\omega}, \overline{\mathbf{C}}\boldsymbol{\omega}]$ and $[\underline{\mathbf{q}}, \overline{\mathbf{q}}]$. The proof of Theorem 3.2 uses the intermediate value argument to establish

the overlap, whereas Oettli–Prager theorem is proved by determining whether a hyperplane meets a box.

Thus, while the Oettli–Prager theorem supports the mathematical principle from an algebraic perturbation perspective with nominal centers, Theorem 3.2 adopts an endpoint-first geometric formulation directly tailored to statistical bound settings.

E Confidence intervals for binomial probability

We briefly review standard methods for constructing $(1-\gamma)$ confidence intervals for binomial probabilities. Let $B \sim \mathrm{Binomial}(n, \pi^*)$. An exact $(1-\gamma)$ confidence interval for π^* is obtained by inverting level-γ binomial tests, yielding the Clopper–Pearson interval:

$$\begin{aligned} \underline{\pi} &= \inf\Big\{\pi \in [0,1] : \mathbb{P}\big(\mathrm{Binomial}(n,\pi) \geq B\big) \leq \gamma_1\Big\}, \\ \overline{\pi} &= \sup\Big\{\pi \in [0,1] : \mathbb{P}\big(\mathrm{Binomial}(n,\pi) \leq B\big) \leq \gamma_2\Big\}, \end{aligned}$$

where $\gamma_1 + \gamma_2 = \gamma$. This yields an exact interval $[\underline{\pi}, \overline{\pi}]$ satisfying $\mathbb{P}\{\pi^* \in [\underline{\pi}, \overline{\pi}]\} \geq 1-\gamma$. Equivalently, letting $Q_\alpha(X)$ denote the α-quantile of a random variable X,

$$\begin{aligned} \underline{\pi} &= Q_{\gamma_1}\{\mathrm{Beta}(B, n-B+1)\}, \\ \overline{\pi} &= Q_{1-\gamma_2}\{\mathrm{Beta}(B+1, n-B)\}. \end{aligned}$$

When $1 \leq B \leq n-1$, it is conventional to choose $\gamma_1 = \gamma_2 = \gamma/2$. When $B = 0$, one sets $\gamma_1 = 0$ and $\gamma_2 = \gamma$, and when $B = n$, one sets $\gamma_1 = \gamma$ and $\gamma_2 = 0$.

Beyond the Clopper–Pearson interval, several exact refinements aim to reduce conservativeness while maintaining guaranteed coverage. The Blyth–Still–Casella interval [4, 7] chooses an optimal tail split (γ_1, γ_2) with $\gamma_1 + \gamma_2 = \gamma$ to yield a shorter confidence interval. The Blaker interval [3] constructs an exact interval using an acceptability function and is often shorter than Clopper–Pearson; in some cases, the resulting set can be a union of disjoint intervals.

There are also $(1-\gamma)$ confidence intervals based on concentration inequalities for the sample proportion $\widehat{\pi} = B/n$. These bounds yield algebraically simple intervals, though they may be conservative in finite samples. First, since $\mathrm{Var}(\widehat{\pi}) = \pi^*(1-\pi^*)/n \leq 1/(4n)$, Chebyshev's inequality gives, for any $t > 0$,

$$\mathbb{P}(|\widehat{\pi} - \pi^*| \geq t) \leq \frac{\mathrm{Var}(\widehat{\pi})}{t^2} \leq \frac{1}{4nt^2}.$$

Thus taking $t = 1/\sqrt{4n\gamma}$ yields the interval

$$\pi^* \in [\widehat{\pi} - 1/\sqrt{4n\gamma},\ \widehat{\pi} + 1/\sqrt{4n\gamma}] \cap [0,1]$$

with probability at least $1-\gamma$.

Next, since $\widehat{\pi}$ is the average of n i.i.d. Bernoulli(π^*) variables, Hoeffding's inequality gives, for any $t > 0$,

$$\mathbb{P}(|\widehat{\pi} - \pi^*| \geq t) \leq 2\exp(-2nt^2).$$

Setting $2\exp(-2nt^2) = \gamma$ yields the explicit $(1-\gamma)$ interval

$$\pi^* \in \left[\widehat{\pi} - \sqrt{\frac{\log(2/\gamma)}{2n}},\ \widehat{\pi} + \sqrt{\frac{\log(2/\gamma)}{2n}}\right] \cap [0,1].$$

Lastly, since B is a sum of n Bernoulli random variables, which are bounded, Bernstein's inequality gives a sharper confidence interval than Hoeffding's inequality with a variance term in the bound. To be specific, for any $t > 0$,

$$\mathbb{P}(|\widehat{\pi} - \pi^*| \geq t) \leq 2\exp\Big\{-\frac{nt^2}{2\pi^*(1-\pi^*) + 2t/3}\Big\},$$

One can further obtain a closed-form bound by upper bounding $\pi^*(1-\pi^*) \leq 1/4$, giving

$$\mathbb{P}\{|\widehat{\pi} - \pi^*| \geq t\} \leq 2\exp\Big(-\frac{nt^2}{1/2 + 2t/3}\Big),$$

and then solving for t such that the right-hand side equals γ. For comprehensive studies of interval estimation for binomial proportions, see Brown et al. [5] and Thulin [31].

F Proofs

F.1 Proof of $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$

Proof. For each $i \in [K]$, we obtain that

$$\begin{aligned}
\sum_{j=1}^{K} C_{ij}^* \omega_j^* &= \sum_{j=1}^{K} \mathbb{P}_{(\mathbf{X},Y)\sim P}\{g(\mathbf{X}) = i, Y = j\} \frac{\mathbb{P}_{(\mathbf{X},Y)\sim Q}(Y = j)}{\mathbb{P}_{(\mathbf{X},Y)\sim P}(Y = j)} \\
&= \sum_{j=1}^{K} \mathbb{P}_{(\mathbf{X},Y)\sim P}\{g(\mathbf{X}) = i \mid Y = j\} \mathbb{P}_{(\mathbf{X},Y)\sim Q}(Y = j) \\
&= \sum_{j=1}^{K} \mathbb{P}_{(\mathbf{X},Y)\sim Q}\{g(\mathbf{X}) = i \mid Y = j\} \mathbb{P}_{(\mathbf{X},Y)\sim Q}(Y = j) \\
&= \mathbb{P}_{(\mathbf{X},Y)\sim Q}\{g(\mathbf{X}) = i\} = q_i^*,
\end{aligned}$$

which implies $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$. □

F.2 Proof of Proposition 3.1

Proof. By the union bound, we obtain that

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\omega}^* \notin \boldsymbol{\Omega}_0) &\leq \mathbb{P}(\mathbf{C}^* \notin \boldsymbol{\mathcal{C}} \text{ or } \mathbf{q}^* \notin \boldsymbol{\mathcal{Q}}) \\
&\leq \sum_{i,j\in[K]} \mathbb{P}(C_{ij}^* \notin [\underline{C}_{ij}, \overline{C}_{ij}]) + \sum_{k\in[K]} \mathbb{P}(q_k^* \notin [\underline{q}_k, \overline{q}_k]) \\
&\leq \sum_{i,j\in[K]} \delta_{ij} + \sum_{k\in[K]} \delta_k = \delta.
\end{aligned}$$

□

F.3 Proof of Theorem 3.2

Proof. Suppose that $\mathbf{C} \in \boldsymbol{\mathcal{C}}$, $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$, and that $\boldsymbol{\omega} \in \boldsymbol{\Omega}_0$ satisfies $\mathbf{C}\boldsymbol{\omega} = \mathbf{q}$ and $\boldsymbol{\omega} \geq 0$. The linear equation $\mathbf{C}\boldsymbol{\omega} = \mathbf{q}$ is equivalent to

$$\sum_{j=1}^{K} C_{ij}\boldsymbol{\omega}_j = q_i.$$

Since $\omega_j \geq 0$, we get that

$$\begin{aligned}
\sum_{j=1}^{K} \overline{C}_{ij}\omega_j \geq \sum_{j=1}^{K} C_{ij}\omega_j = q_i \geq \underline{q}_i, \\
\sum_{j=1}^{K} \underline{C}_{ij}\omega_j \leq \sum_{j=1}^{K} C_{ij}\omega_j = q_i \leq \overline{q}_i,
\end{aligned}$$

which implies that $\boldsymbol{\omega} \in \boldsymbol{\Omega}$.

Now, we prove the other direction. Suppose that $\boldsymbol{\omega} \in \boldsymbol{\Omega}$. Then for each $i \in [K]$, we can apply the following procedure. If $\sum_{j=1}^{K} \overline{C}_{ij}\omega_j \leq \overline{q}_i$ for any $i \in [K]$, take $C_i^\top = (\overline{C}_{i1}, \ldots, \overline{C}_{iK})$ and $q_i = \sum_{j=1}^{K} \overline{C}_{ij}\omega_j \in [\underline{q}_i, \overline{q}_i]$. Otherwise, assume $\sum_{j=1}^{K} \overline{C}_{ij}\omega_j > \overline{q}_i$. For $l = 0, 1, ..., K$, define

$$q_i(l) = \sum_{j=1}^{l} \underline{C}_{ij}\omega_j + \sum_{j=l+1}^{K} \overline{C}_{ij}\omega_j.$$

Then $q_i(l)$ is a decreasing function of l. Also, $q_i(K) = \sum_{j=1}^{K} \underline{C}_{ij}\omega_j \leq \overline{q}_i$ by the condition, and $q_i(0) = \sum_{j=1}^{K} \overline{C}_{ij}\omega_j > \overline{q}_i$ by the assumption. Thus, we can find l_0 such that $q_i(l_0 - 1) >$

$\overline{q}_i \geq q_i(l_0)$. Note that $\omega_{l_0} > 0$ since $0 < q_i(l_0 - 1) - q_i(l_0) = (\overline{C}_{i,l_0} - \underline{C}_{i,l_0})\omega_{l_0}$. Let $C_i^\top = (\overline{C}_{i1}, \ldots, \overline{C}_{i,l_0-1}, \widetilde{C}_{i,l_0}, \underline{C}_{i,l_0+1}, \ldots, \underline{C}_{iK})$ and $q_i = \overline{q}_i$, where

$$\widetilde{C}_{i,l_0} = \underline{C}_{i,l_0} + \frac{\overline{q}_i - q_i(l_0)}{\omega_{l_0}}.$$

Then $\widetilde{C}_{i,l_0} \in [\underline{C}_{i,l_0}, \overline{C}_{i,l_0}]$ and $C_i^\top \boldsymbol{\omega} = q_i$. Taking $C_i^\top$ as the i-th row of $\mathbf{C}$ and q_i as the i-th element of $\mathbf{q}$, we get that $\mathbf{C}\boldsymbol{\omega} = \mathbf{q}$, where $\mathbf{C} \in \boldsymbol{\mathcal{C}}$ and $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$, meaning $\boldsymbol{\omega} \in \boldsymbol{\Omega}_0$. □

F.4 Proof of Corollary 3.4

Proof. The first part, $\boldsymbol{\Omega} \subset \boldsymbol{\Omega}_{\mathrm{LP}}$, is trivial from its definition. For the second part, since $\boldsymbol{\Omega}_{\mathrm{GE}}$ exists and contains all $\mathbf{C}^{-1}\mathbf{q}$ for $\mathbf{C} \in \boldsymbol{\mathcal{C}}$ and $\mathbf{q} \in \boldsymbol{\mathcal{Q}}$ by its construction, we have $\boldsymbol{\Omega} \subset \boldsymbol{\Omega}_{\mathrm{GE}}$. Since $\boldsymbol{\Omega}_{\mathrm{LP}}$ is the smallest hyperrectangle that contains $\boldsymbol{\Omega}$, we get $\boldsymbol{\Omega}_{\mathrm{LP}} \subset \boldsymbol{\Omega}_{\mathrm{GE}}$. □

F.5 Proof of Theorem 3.5

Proof. We prove the ℓ_∞ bound; the ℓ_2 bound follows by the same argument with $\|\cdot\|_\infty$ replaced by $\|\cdot\|_2$.

Let $\boldsymbol{\omega} = \mathbf{A}^{-1}\mathbf{q}$ and $\boldsymbol{\omega}' = \mathbf{B}^{-1}\mathbf{q}'$ in $\boldsymbol{\Omega}$, where $\underline{\mathbf{C}} \leq \mathbf{A}, \mathbf{B} \leq \overline{\mathbf{C}}$ and $\underline{\mathbf{q}} \leq \mathbf{q}, \mathbf{q}' \leq \overline{\mathbf{q}}$. Write $\mathbf{A} = \underline{\mathbf{C}} + \Delta$ and $\mathbf{B} = \underline{\mathbf{C}} + \Delta'$ with $\mathbf{0} \leq \Delta, \Delta' \leq \overline{\mathbf{C}} - \underline{\mathbf{C}}$, and write $\mathbf{q} = \overline{\mathbf{q}} - \boldsymbol{d}$ and $\mathbf{q}' = \overline{\mathbf{q}} - \boldsymbol{d}'$ with $\mathbf{0} \leq \boldsymbol{d}, \boldsymbol{d}' \leq \overline{\mathbf{q}} - \underline{\mathbf{q}}$. Then

$$\begin{aligned}
\boldsymbol{\omega} - \boldsymbol{\omega}' &= \mathbf{A}^{-1}(\overline{\mathbf{q}} - \boldsymbol{d}) - \mathbf{B}^{-1}(\overline{\mathbf{q}} - \boldsymbol{d}') \\
&= (\mathbf{A}^{-1} - \mathbf{B}^{-1})(\overline{\mathbf{q}} - \boldsymbol{d}') - \mathbf{A}^{-1}(\boldsymbol{d} - \boldsymbol{d}') \\
&= \mathbf{A}^{-1}(\Delta' - \Delta)\mathbf{B}^{-1}(\overline{\mathbf{q}} - \boldsymbol{d}') - \mathbf{A}^{-1}(\boldsymbol{d} - \boldsymbol{d}'),
\end{aligned}$$

where the last line holds by $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$. Taking $\|\cdot\|_\infty$ gives

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_\infty \leq \|\mathbf{A}^{-1}\|_\infty \|\Delta' - \Delta\|_\infty \|\mathbf{B}^{-1}\|_\infty \|\overline{\mathbf{q}} - \boldsymbol{d}'\|_\infty + \|\mathbf{A}^{-1}\|_\infty \|\boldsymbol{d} - \boldsymbol{d}'\|_\infty.$$

Since $0 \leq \Delta, \Delta' \leq \overline{\mathbf{C}} - \underline{\mathbf{C}}$ and $0 \leq \boldsymbol{d}, \boldsymbol{d}' \leq \overline{\mathbf{q}} - \underline{\mathbf{q}}$, we have $\|\Delta' - \Delta\|_\infty \leq \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty$, $\|\boldsymbol{d} - \boldsymbol{d}'\|_\infty \leq \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_\infty$, and $\|\overline{\mathbf{q}} - \boldsymbol{d}'\|_\infty \leq \|\overline{\mathbf{q}}\|_\infty$. Moreover, for any Δ with $\|\Delta\|_\infty \leq \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty$, $(\underline{\mathbf{C}} + \Delta)^{-1} = (I + \underline{\mathbf{C}}^{-1}\Delta)^{-1}\underline{\mathbf{C}}^{-1}$, and

$$\begin{aligned}
\|(I + \underline{\mathbf{C}}^{-1}\Delta)^{-1}\|_\infty &= \sup_{\boldsymbol{v} \neq 0} \frac{\|(I + \underline{\mathbf{C}}^{-1}\Delta)^{-1}\boldsymbol{v}\|_\infty}{\|\boldsymbol{v}\|_\infty} = \frac{1}{\inf_{\|\boldsymbol{w}\|_\infty = 1} \|(I + \underline{\mathbf{C}}^{-1}\Delta)\boldsymbol{w}\|_\infty} \\
&\leq \frac{1}{1 - \sup_{\|\boldsymbol{w}\|_\infty = 1} \|\underline{\mathbf{C}}^{-1}\Delta\boldsymbol{w}\|_\infty} \leq \frac{1}{1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\Delta\|_\infty},
\end{aligned}$$

where we wrote $\boldsymbol{v} = (I + \underline{\mathbf{C}}^{-1}\Delta)\boldsymbol{w}$ and used $\|\mathbf{C}^{-1}\Delta\|_\infty < 1$. Hence,

$$\|(\underline{\mathbf{C}} + \Delta)^{-1}\|_\infty \leq \frac{\|\underline{\mathbf{C}}^{-1}\|_\infty}{1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\Delta\|_\infty} \leq \frac{\|\underline{\mathbf{C}}^{-1}\|_\infty}{1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty}.$$

Applying this bound to $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ and substituting into the previous inequality yields the stated $\mathrm{diam}_\infty(\boldsymbol{\Omega})$ bound. The $\mathrm{diam}_2(\boldsymbol{\Omega})$ bound follows identically by replacing $\|\cdot\|_\infty$ with $\|\cdot\|_2$ throughout. □

F.6 Proof of Theorem 3.7

Proof. By Theorem 2 of Podkopaev and Ramdas [25], we have

$$\mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\widetilde{Y}_0 \in F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\omega}^*)\} \geq 1 - \alpha.$$

Also, if $\boldsymbol{\omega}^* \in \boldsymbol{\Omega}_0$, we get $F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\omega}^*) \subset F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\Omega}_0)$. Then

$$\begin{aligned}
\mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\widetilde{Y}_0 \notin F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\Omega}_0)\} &= \mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\boldsymbol{\omega}^* \in \boldsymbol{\Omega}_0 \text{ and } \widetilde{Y}_0 \notin F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\Omega}_0)\} \\
&\quad + \mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\boldsymbol{\omega}^* \notin \boldsymbol{\Omega}_0 \text{ and } \widetilde{Y}_0 \notin F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\Omega}_0)\} \\
&\leq \mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\widetilde{Y}_0 \notin F_{\mathrm{CP}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\omega}^*)\} + \mathbb{P}(\boldsymbol{\omega}^* \notin \boldsymbol{\Omega}_0) \\
&\leq \alpha + \delta.
\end{aligned}$$

□

F.7 Proof of Theorem 3.9

Proof. First, (5) implies $\tau_{\mathrm{CP}}(\widetilde{y};\mathbf{\Omega}_1) \le \tau_{\mathrm{CP}}(\widetilde{y};\mathbf{\Omega}_2)$ for all $\widetilde{y}$. Then (6) gives the result. □

G An explicit bound obtained by Theorem 3.5

Theorem 3.5 *Let $\underline{\mathbf{C}} \in \mathbb{R}^{K\times K}$ be invertible. If $\|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty < 1$, then*

$$\mathrm{diam}_\infty(\mathbf{\Omega}) \le \frac{\|\underline{\mathbf{C}}^{-1}\|_\infty^2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty \|\overline{\mathbf{q}}\|_\infty}{\left(1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty\right)^2} + \frac{\|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_\infty}{1 - \|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty}.$$

If $\|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 < 1$, then

$$\mathrm{diam}_2(\mathbf{\Omega}) \le \frac{\|\underline{\mathbf{C}}^{-1}\|_2^2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 \|\overline{\mathbf{q}}\|_2}{\left(1 - \|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2\right)^2} + \frac{\|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_2}{1 - \|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2}.$$

We use Hoeffding's inequality method construct confidence intervals for binomial probability, illustrating the usage of Theorem 3.5. For preparation, let $\delta' = \delta/(K^2+K)$, and for each $i, j, k \in [K]$, let $[\underline{C}_{ij}, \overline{C}_{ij}]$ and $[\underline{q}_k, \overline{q}_k]$ be $(1-\delta')$-level confidence intervals constructed with Hoeffding's inequality (Appendix E): if $B \sim \mathrm{Binomial}(r, \pi)$, then

$$\Pr\{|B/r - \pi| \le \sqrt{\frac{\log(2/\delta')}{2r}}\} \ge 1 - \delta',$$

which gives us

$$\overline{C}_{ij} - \underline{C}_{ij} \le \sqrt{2\log(2/\delta')/m}, \qquad \overline{q}_k - \underline{q}_k \le \sqrt{2\log(2/\delta')/n}. \tag{7}$$

Assume $\underline{\mathbf{C}}$ is invertible and $\sigma_{\min}(\underline{\mathbf{C}}) \ge 1/(LK)$ for some constant $L > 0$. Then $\|\underline{\mathbf{C}}^{-1}\|_2 \le LK$, and subsequently $\|\underline{\mathbf{C}}^{-1}\|_\infty \le LK^{3/2}$.

Now we provide the $\|\cdot\|_\infty$ norm bound. By (7),

$$\|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty \le K\sqrt{2\log(2/\delta')/m}, \qquad \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_\infty \le \sqrt{2\log(2/\delta')/n}.$$

Since $\overline{q}_k \le 1$ for all $k \in [K]$, we have $\|\overline{\mathbf{q}}\|_\infty \le 1$. Assume m is sufficiently large so that $\|\underline{\mathbf{C}}^{-1}\|_\infty \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_\infty \le LK^{5/2}\sqrt{2\log(2/\delta')/m} < 1/2$. Then by Theorem 3.5,

$$\begin{aligned}
\mathrm{diam}_\infty(\mathbf{\Omega}) &\le \frac{L^2K^4\sqrt{2\log(2/\delta')/m}}{\left(1 - LK^{5/2}\sqrt{2\log(2/\delta')/m}\right)^2} + \frac{LK^{3/2}\sqrt{2\log(2/\delta')/n}}{1 - LK^{5/2}\sqrt{2\log(2/\delta')/m}} \\
&\le 4L^2 m^{-1/2} K^4 \sqrt{2\log(4K^2/\delta)} + 2Ln^{-1/2}K^{3/2}\sqrt{2\log(4K^2/\delta)}.
\end{aligned}$$

In a similar manner, we get the $\|\cdot\|_2$ norm bound under mild conditions (Appendix G.1): when $LK^2\sqrt{2\log(2/\delta')/m} < 1/2$, we have

$$\mathrm{diam}_2(\mathbf{\Omega}) \le 4L^2 m^{-1/2} K^{7/2} \sqrt{2\log(4K^2/\delta)} + 2Ln^{-1/2}K^{3/2}\sqrt{2\log(4K^2/\delta)}.$$

Overall, we can observe that both $\mathrm{diam}_\infty(\mathbf{\Omega})$ and $\mathrm{diam}_2(\mathbf{\Omega})$ have $O\{(m^{-1/2}K^4 + n^{-1/2}K^{3/2})\sqrt{\log(K/\delta)}\}$ and $O\{(m^{-1/2}K^{7/2} + n^{-1/2}K^{3/2})\sqrt{\log(K/\delta)}\}$ rate, respectively.

G.1 Proof of $\|\cdot\|_2$ bound

Proof. By (7), we have

$$\|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 \le K\sqrt{\frac{2\log(2/\delta')}{m}}, \qquad \|\overline{\mathbf{q}} - \underline{\mathbf{q}}\|_2 \le \sqrt{K}\sqrt{\frac{2\log(2/\delta')}{n}}.$$

Since $\overline{q}_k \le 1$ for $k \in [K]$, we have $\|\overline{\mathbf{q}}\|_2 \le \sqrt{K}$. Assume in addition that m is sufficiently large so that $\|\underline{\mathbf{C}}^{-1}\|_2 \|\overline{\mathbf{C}} - \underline{\mathbf{C}}\|_2 \le LK^2\sqrt{2\log(2/\delta')/m} < 1/2$. Then by Theorem 3.5,

$$\begin{aligned}\mathrm{diam}_2(\mathbf{\Omega}) &\le \frac{L^2K^{7/2}\sqrt{\frac{2\log(2/\delta')}{m}}}{\big(1 - LK^2\sqrt{\frac{2\log(2/\delta')}{m}}\big)^2} + \frac{LK^{3/2}\sqrt{\frac{2\log(2/\delta')}{n}}}{1 - LK^2\sqrt{\frac{2\log(2/\delta')}{m}}} \\ &\le 4L^2 m^{-1/2}K^{7/2}\sqrt{2\log(4K^2/\delta)} + 2Ln^{-1/2}K^{3/2}\sqrt{2\log(4K^2/\delta)}.\end{aligned}$$

□

H Details of PAC prediction

Let the calibration set $\mathcal{D}_{\mathrm{cali}} = \{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)\}_{i=1}^{m_1}$ follow the source distribution and denote by $r(\widetilde{\boldsymbol{x}}, \widetilde{y})$ the nonconformity score trained on an independent training set. The PAC prediction set $F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}; \boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}})$ under label shift [35, 23, 29] is defined by

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cali}} \sim P^{m_1}}[\mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\widetilde{Y}_0 \in F_{\mathrm{PAC}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}})\} \ge 1 - \epsilon] \ge 1 - \eta.$$

Si et al. [29] constructed a PAC prediction set based on the confidence region $\mathbf{\Omega}_{\mathrm{GE}}$. For a fixed candidate weight vector $\boldsymbol{\omega}$, calibration set $\mathcal{D}_{\mathrm{cali}}$, auxiliary random variable $V \sim \mathrm{Uniform}([0,1])^{m_1}$, and a large constant $b \ge \max_{k\in[K]} \omega_k$, their construction yields a prediction set $F_{\mathrm{PAC}}(\cdot; \mathbf{\Omega}_0, \mathcal{D}_{\mathrm{cali}}, V, b)$ satisfying the following PAC guarantee:

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cali}} \sim P^{m_1}, V}[\mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\widetilde{Y}_0 \in F_{\mathrm{PAC}}(\widetilde{\mathbf{X}}_0; \boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\} \ge 1 - \epsilon] \ge 1 - \eta, \tag{8}$$

Then, the PAC prediction set $F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}; \boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)$ is in the form of

$$F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}; \boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b) = \{\widetilde{y} \in [K] : r(\widetilde{\boldsymbol{x}}, \widetilde{y}) \le \tau_{\mathrm{PAC}}\{T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\}\},$$

where $T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b) = \{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) \in \mathcal{D}_{\mathrm{cali}} : V_i \le \omega_{\widetilde{y}_i}/b\}$ is a target sample generated by rejection-sampling from $\mathcal{D}_{\mathrm{cali}}$. Let $m_0 = |T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)|$. Here, $\tau_{\mathrm{PAC}}\{T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\}$ is chosen to satisfy

$$\begin{aligned}&\sum_{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) \in T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)} I\{\widetilde{y}_i \notin F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}_i; \boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\} \\ &= \sum_{(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) \in T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)} I[r(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) > \tau_{\mathrm{PAC}}\{T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\}] \le k(m_0, \epsilon, \eta),\end{aligned}$$

where

$$k(m_0, \epsilon, \eta) = \max\{k : F_{\mathrm{Binomial}(m_0, \epsilon)}(k) \le \eta\}.$$

That is, $\tau_{\mathrm{PAC}}\{T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\}$ is the $k(m_0, \epsilon, \eta)$-th largest value of $\{r(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) : (\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i) \in T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\}$. Note that $F_{\mathrm{Binomial}(m_0,\epsilon)}(\cdot)$ is the CDF of Binomial(m_0, ϵ). If the true importance weight $\boldsymbol{\omega}^*$ replaces $\boldsymbol{\omega}$, then the modified PAC condition (8) is satisfied. When the confidence region $\mathbf{\Omega}_0$ with $\mathbb{P}(\boldsymbol{\omega}^* \in \mathbf{\Omega}_0) \ge 1 - \delta$ is provided, we can define

$$\tau_{\mathrm{PAC}}\{T(\mathbf{\Omega}_0, \mathcal{D}_{\mathrm{cali}}, V, b)\} = \sup_{\boldsymbol{\omega} \in \mathbf{\Omega}_0} \tau_{\mathrm{PAC}}\{T(\boldsymbol{\omega}, \mathcal{D}_{\mathrm{cali}}, V, b)\} \tag{9}$$

and

$$F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}; \mathbf{\Omega}_0, \mathcal{D}_{\mathrm{cali}}, V, b) = \{\widetilde{y} \in [K] : r(\widetilde{\boldsymbol{x}}, \widetilde{y}) \le \tau_{\mathrm{PAC}}\{T(\mathbf{\Omega}_0, \mathcal{D}_{\mathrm{cali}}, V, b)\}\}. \tag{10}$$

Then $F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}; \mathbf{\Omega}_0, \mathcal{D}_{\mathrm{cali}}, V, b)$ satisfies the modified PAC condition (8) with $1-\eta$ replaced by $1-\eta-\delta$.

As in Section 3.4.1, the resulting prediction set is monotone in the weight region. Therefore, replacing $\mathbf{\Omega}_{\mathrm{GE}}$ by the tighter region $\mathbf{\Omega}_{\mathrm{LP}}$ yields a smaller prediction set while preserving the same PAC guarantee [23].

Theorem H.1. *If $\mathbf{\Omega}_1 \subset \mathbf{\Omega}_2$, then $F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_1, \mathcal{D}_{\mathrm{cali}}, V, b) \subset F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_2, \mathcal{D}_{\mathrm{cali}}, V, b)$ for all $\widetilde{\boldsymbol{x}}_0$. In particular, $F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_{\mathrm{LP}}, \mathcal{D}_{\mathrm{cali}}, V, b) \subset F_{\mathrm{PAC}}(\widetilde{\boldsymbol{x}}_0; \mathbf{\Omega}_{\mathrm{GE}}, \mathcal{D}_{\mathrm{cali}}, V, b)$.*

Proof. First, (9) implies $\tau_{\mathrm{PAC}}\{T(\boldsymbol{\Omega}_1, \mathcal{D}_{\mathrm{cali}}, V, b)\} \leq \tau_{\mathrm{PAC}}\{T(\boldsymbol{\Omega}_2, \mathcal{D}_{\mathrm{cali}}, V, b)\}$ for all $\widetilde{y}$. Then (10) gives the result. □

Theorem H.2. *Suppose that* $\mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}_0) \geq 1 - \delta$. *Then*

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cali}} \sim P^{m_1}, V}[\mathbb{P}_{(\widetilde{\mathbf{X}}_0, \widetilde{Y}_0) \sim Q}\{\widetilde{Y}_0 \in F_{PAC}(\widetilde{\mathbf{X}}_0; \boldsymbol{\Omega}_0, \mathcal{D}_{\mathrm{cali}}, V, b)\} \geq 1 - \epsilon] \geq 1 - \eta - \delta.$$

Proof. The proof follows from Theorem 3 of [29]. □

I Gaussian elimination with intervals

Given that $\mathbf{C}^* \in \boldsymbol{\mathcal{C}} = [\underline{\mathbf{C}}, \overline{\mathbf{C}}]$ and $\mathbf{q}^* \in \boldsymbol{\mathcal{Q}} = [\underline{\mathbf{q}}, \overline{\mathbf{q}}]$, Si et al. [29] introduce an intuitive way, which they named Gaussian elimination with intervals, of finding $\boldsymbol{\Omega}$ that contains $\boldsymbol{\omega}^* = \mathbf{C}^{*-1}\mathbf{q}^*$. Suppose that $\underline{C}_{ij} \geq 0$, $\underline{q}_i > 0$, and $\omega_i^* > 0$ for $i, j \in [K]$. They follow two phases of Gaussian elimination when solving a system of linear equations $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$ and derive the classwise interval for ω_i^*. First, set $\underline{C}_{ij}^1 = \underline{C}_{ij}$, $\overline{C}_{ij}^1 = \overline{C}_{ij}$, $\underline{q}_i^1 = \underline{q}_i$, and $\overline{q}_i^1 = \overline{q}_i$. In the first phase (forward elimination), the elementary row operations are applied sequentially for $k = 1, ..., K-1$ to delete the (i, k) element in the matrix for $i > k$ by adding the multiple of the k-th row. Then the lower bound $\underline{C}_{ij}^{k+1}$ and the upper bound $\overline{C}_{ij}^{k+1}$ are derived from the interval $[\underline{\mathbf{C}}^k, \overline{\mathbf{C}}^k]$ at the k-th step as

$$\underline{C}_{ij}^{k+1} = \begin{cases} 0, & \text{if } i > k, j \leq k, \\ \underline{C}_{ij}^k - \frac{\overline{C}_{ik}^k \overline{C}_{kj}^k}{\underline{C}_{kk}^k}, & \text{if } i, j > k, \\ \underline{C}_{ij}^k, & \text{otherwise.} \end{cases}$$

$$\overline{C}_{ij}^{k+1} = \begin{cases} 0, & \text{if } i > k, j \leq k, \\ \overline{C}_{ij}^k - \frac{\underline{C}_{ik}^k \underline{C}_{kj}^k}{\overline{C}_{kk}^k}, & \text{if } i, j > k, \\ \overline{C}_{ij}^k, & \text{otherwise.} \end{cases}$$

Simultaneously, $\underline{q}_i^{k+1}$ and $\overline{q}_i^{k+1}$ are obtained from the same row operations to be

$$\underline{q}_i^{k+1} = \begin{cases} \underline{q}_i^k - \frac{\overline{C}_{ik}^k \overline{q}_k^k}{\underline{C}_{kk}^k}, & \text{if } i > k, \\ \underline{q}_i^k, & \text{otherwise.} \end{cases}$$

$$\overline{q}_i^{k+1} = \begin{cases} \overline{q}_i^k - \frac{\underline{C}_{ik}^k \underline{q}_k^k}{\overline{C}_{kk}^k}, & \text{if } i > k, \\ \overline{q}_i^k, & \text{otherwise.} \end{cases}$$

Then $C_{ij}^{*,k+1}$ and $q_i^{*,k+1}$, which would have been obtained in the forward elimination step solving $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$, always lie in $[\underline{C}_{ij}^{k+1}, \overline{C}_{ij}^{k+1}]$ and $[\underline{q}_i^{k+1}, \overline{q}_i^{k+1}]$. In the second phase (back substitution), they compute $\underline{\omega}_i$ and $\overline{\omega}_i$, iteratively for $i = K, ..., 1$, replacing the truth with intervals as in the first phase.

$$\underline{s}_i = \sum_{j=i+1}^{K} \underline{C}_{ij}^K \underline{\omega}_j \quad \text{and} \quad \overline{s}_i = \sum_{j=i+1}^{K} \overline{C}_{ij}^K \overline{\omega}_j,$$
$$\underline{\omega}_i = \frac{\underline{q}_i - \overline{s}_i}{\overline{C}_{ii}^K} \quad \text{and} \quad \overline{\omega}_i = \frac{\overline{q}_i - \underline{s}_i}{\underline{C}_{ii}^K}.$$

Then $\boldsymbol{\Omega}_{\mathrm{GE}}$ is defined as the K-dimensional hypercube $\prod_{i=1}^K [\underline{\omega}_i, \overline{\omega}_i]$. Si et al. [29] provide a theoretical result that their method yields $\omega^* \in \boldsymbol{\Omega}_{\mathrm{GE}}$ if $\underline{C}_{ij}^k \geq 0$, $\underline{C}_{ii}^k > 0$, and $\underline{q}_i^k \geq 0$ for all $i, j, k \in [K]$. The basic assumption in order to satisfy the condition is that $\overline{C}_{ik} \ll \underline{C}_{kk}$. This is ensured when the classifier $g(X)$ is accurate, that is, when the diagonal terms C_{kk}^* in $\mathbf{C}^*$ dominate non-diagonal

terms. If the assumption is violated, we may encounter a possibility that $C^{*,k}_{kk} \approx 0$, which may lead to $\underline{C}^k_{kk} \leq 0$. Then in the forward elimination phase, $\underline{C}^{k+1}_{ij}$ for all $i, j > k$ will be $-\infty$, which may make the algorithm impractical. Furthermore, if $\underline{q}_i \leq \overline{s}_i$ or $\underline{C}^K_{ii} \leq 0$ for some i, then the back substitution phase would lead to $\underline{\omega}_i \leq 0$ or $\overline{\omega}_i = \infty$, which does not provide any information about the interval of ω^*_i. In order to deal with the nonpositive bounds, they mention that choosing a wider margin, which would, however, make $\mathbf{\Omega}_{\mathrm{GE}}$ larger than its optimal size.

I.1 An example to illustrate why GE is conservative

Interval Gaussian elimination enlarges the region because for some (i, j), both $\overline{C}_{ij}$ and $\underline{C}_{ij}$ are used within the same computation. LP never does this, since each constraint uses one consistent choice of endpoints.

For example, apply interval Gaussian elimination to the matrix equation in Appendix A, where $C_{11} \in [0.3, 0.4]$, $C_{12} \in [0.1, 0.2]$, $C_{21} \in [0, 0.1]$, $C_{22} \in [0.4, 0.6]$, $q_1 \in [0.6, 0.7]$, and $q_2 \in [0.3, 0.4]$. Then $\underline{\omega}_2 = \underline{q}^{(2)}_2 / \overline{C}^{(2)}_{22}$, where

$$\overline{C}^{(2)}_{22} = \overline{C}_{22} - \frac{\underline{C}_{21}\,\underline{C}_{12}}{\overline{C}_{11}} = 0.6 - \frac{0 \cdot 0.1}{0.4} = 0.6, \qquad \underline{q}^{(2)}_2 = \underline{q}_2 - \frac{\overline{C}_{21}\,\overline{q}_1}{\underline{C}_{11}} = 0.3 - \frac{0.1 \cdot 0.7}{0.3} = \frac{1}{15},$$

giving $\underline{\omega}_2 = 1/9 \approx 0.111$. Here C_{21} enters at its upper endpoint in $\underline{q}^{(2)}_2$ and at its lower endpoint in $\overline{C}^{(2)}_{22}$, and C_{11} likewise appears as $\underline{C}_{11}$ in one place and $\overline{C}_{11}$ in the other, so no admissible matrix attains this bound.

On the other hand, $\underline{\omega}_2$ in $\mathbf{\Omega}$ from MaC is attained by an admissible pair $(\mathbf{C}, \mathbf{q}) \in \mathcal{C} \times \mathcal{Q}$ satisfying

$$\begin{pmatrix} \underline{C}_{11} & \underline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} \overline{q}_1 \\ \underline{q}_2 \end{pmatrix},$$

which yields $\underline{\omega}_2 = 2/17 \approx 0.118$, a tighter lower bound. This value is attained at $\boldsymbol{\omega} = (2.294,\, 0.118)^\top \geq \mathbf{0}$ with $\mathbf{C} = \left(\begin{smallmatrix} 0.3 & 0.1 \\ 0.1 & 0.6 \end{smallmatrix}\right)$ and $\mathbf{q} = (0.7,\, 0.3)^\top$, as guaranteed by the sufficiency direction of Theorem 3.2. The resulting boxes are shown in Figure S2.

J Detailed experimental setup

The experiment was implemented using `Python 3.13` on MacBook Pro with the Apple M3 Pro chip, which has a 11-core CPU (5 performance cores and 6 efficiency cores), an 14-core GPU, and 18 GB of unified memory. The classifiers were all trained using Google Colab with NVIDIA H100 GPU.

J.1 Benchmark data organization and classifiers

Table S1: Summary of dataset splits and architectures. *VGG16, ResNet18, PreActResNet18, MobileNetV2, and RegNetX.

Dataset	Architecture	Training Pool (Train + Test)	Analysis Set
AGNews	RoBERTa	60,000 + 7,600	60,000
MNIST	4-Layer CNN	30,000 + 10,000	30,000
CIFAR-10	Multiple*	25,000 + 10,000	25,000
N24News	BERT + ViT	24,487 + 6,122	30,609

To obtain a classifier g, we did the following. For AGNews, we fine-tuned 10 RoBERTa [18] models using distinct random seeds. For MNIST, we trained 10 four-layer CNNs with varying initializations. For CIFAR-10, we utilized five distinct architectures, VGG16 [30], ResNet18 [13], PreActResNet18 [14], MobileNetV2 [28], and RegNetX [26], each trained with two different seeds. For N24News, we employed a multi-modal late-fusion architecture combining BERT [9] and ViT [10], training 10 models with different seeds.

Text (AGNews). We created a stratified analysis set of 60,000 training examples (15,000 per class) using a fixed random seed. The remaining 60,000 examples were used for training. We fine-tuned `RoBERTa` for 3 epochs. Hyperparameters: learning rate 3×10^{-5}, batch sizes of 32 (train) and 64 (eval), AdamW optimizer (0.01 weight decay), and 500 warmup steps. Mixed-precision (fp16) was enabled.

Vision benchmarks (MNIST & CIFAR-10). For both datasets, we use the half of total training samples with the 10,000 standard test samples.

- **MNIST**: We trained 10 classifiers (10 epochs each) using a CNN with two convolutional layers (5×5 kernels, 10 and 20 filters) and two fully connected layers (50 hidden units). The remaining 30,000 original training observations were reserved as the analysis set.
- **CIFAR-10**: We trained 10 classifiers (two instances each of `VGG`, `ResNet18`, `PreActResNet18` `MobileNetV2`, and `RegNetX`). Each model was trained for 200 epochs. The remaining 25,000 training samples served as the analysis set.

Multimodal (N24News). We implemented a late-fusion model concatenating 768-d features from `BERT` and `ViT` followed by a linear layer to 24 classes. Using a 50/50 stratified split ($\approx$ 30,609 samples each), models were trained for 3 epochs (lr 3×10^{-5}, batch size 16) across 10 replications.

J.2 Real-world data (nuImages) organization and classifiers

J.2.1 Dataset and preprocessing

We evaluate the proposed MaC-LP framework using the **nuImages**[3] dataset [6], a large-scale autonomous driving dataset featuring high-resolution images (1600×900 pixels) captured across diverse urban environments, which have 23 annotated classes. Our study utilizes the full metadata of the nuImages v1.0 release, encompassing a total of 50,786 training images and 13,394 validation(test;

[3]Licenses: CC BY-NC-SA 4.0

in our context for training a classifier) images from Singapore, alongside 16,493 training images and 3,051 validation images from Boston.

To transition from raw street-scene images to a structured classification task, we perform object-centric cropping. For each annotated instance, we extract the image content within the ground-truth bounding box. To ensure the reliability of the classifier $g(x)$ and mitigate label noise from low-resolution distant objects, we apply a visibility filter: only crops with a large resolution (larger than 64×64 pixels) are retained. Following extraction, all crops are resized to 224×224 pixels using bicubic interpolation and normalized according to ImageNet statistics to facilitate the use of pre-trained deep convolutional neural networks.

J.2.2 Data partitioning, classifiers, and setting

Consistent with the methodology required for the MaC-LP framework, we partition the Singapore and Boston data into two functionally distinct subsets:

1. **Training Pool (train + test):** We combine the original training and validation splits from the nuImages metadata. From this unified pool, we randomly sample 50% of the Singapore train data (stratified by category) with the whole Singapore validation dataset and fine-tune the classification model $g(x)$. This phase optimizes the model to capture the feature representations of the source domain.
2. **Analysis Set:** The remaining 50% of the Singapore train data and the whole Boston data serve as the Analysis Set. As specified in the MaC-LP framework, this set remains unseen during the training phase. It is utilized exclusively for label shift estimation and downstream application tasks.
3. **Classifier:** We used 10 different random seeds fine-tuned ImageNet-pretrained backbone for our classifier $g(x)$ directly on the long-tailed source distribution without artificial rebalancing solely on Singapore data. The classifier was tuned for 5 epochs given the robust feature representations provided by the ImageNet-pretrained V1 model.

The resulting data distribution for the Training Pool and Analysis sets is summarized in Table S2 after filtering the classes which have less than 100 samples. This results in 15 classes retained in our real-world data application. In Table S3, only Singapore data are used in training classifiers process. We denote the number of samples in the source domain as m and in the target domain as n. This stratified splitting strategy ensures that our estimation of $\mathbf{C}$ is unbiased, as it is derived from data independent of the training process, thereby preserving the mathematical integrity of the linear programming bounds.

In our experimental design, *Singapore* serves as the source domain and *Boston* as the target domain. The "true" label proportions in both domain and the importance weights (Ratio) are in Table S2.

Table S2: Statistics of the original Training Set of nuImages (larger than 64×64 pixels). Singapore and Boston represent the Source and Target domains, respectively.

Category	**Source Singapore**	**Target Boston**	$p(y)$	$q(y)$	**Ratio** $w(y)$
human.pedestrian.adult	8,493	3,148	0.0889	0.0599	0.6735
human.pedestrian.construction_worker	1,469	107	0.0154	0.0020	0.1325
human.pedestrian.personal_mobility	389	1	0.0041	1.9×10^{-5}	0.0047
movable_object.barrier	17,256	3,990	0.1807	0.0759	0.4202
movable_object.debris	782	111	0.0082	0.0021	0.2579
movable_object.pushable_pullable	988	93	0.0103	0.0018	0.1710
movable_object.trafficcone	4,024	2,703	0.0421	0.0514	1.2206
static_object.bicycle_rack	732	560	0.0077	0.0107	1.3901
vehicle.bicycle	5,222	942	0.0547	0.0179	0.3278
vehicle.bus.rigid	1,925	909	0.0202	0.0173	0.8580
vehicle.car	38,293	29,511	0.4009	0.5614	1.4004
vehicle.construction	2,197	752	0.0230	0.0143	0.6220
vehicle.motorcycle	5,943	610	0.0622	0.0116	0.1865
vehicle.trailer	139	1,983	0.0015	0.0377	25.9230
vehicle.truck	7,660	7,143	0.0802	0.1359	1.6945
TOTAL	**95,512**	**52,563**	**1.0000**	**1.0000**	–

Table S3: Statistics of the original Validation Set of nuImages (larger than 64×64 pixels). Used alongside Training Set for category-wise performance monitoring.

Category	**Source Singapore**	**Target Boston**	$p(y)$	$q(y)$	**Ratio** $w(y)$
human.pedestrian.adult	1,851	647	0.0762	0.0584	0.7665
human.pedestrian.construction_worker	421	14	0.0173	0.0013	0.0729
human.pedestrian.personal_mobility	87	1	0.0036	0.0001	0.0252
movable_object.barrier	4,903	749	0.2018	0.0676	0.3350
movable_object.debris	183	20	0.0075	0.0018	0.2397
movable_object.pushable_pullable	227	12	0.0093	0.0011	0.1159
movable_object.trafficcone	1,112	640	0.0458	0.0578	1.2621
static_object.bicycle_rack	177	119	0.0073	0.0107	1.4744
vehicle.bicycle	1,223	246	0.0503	0.0222	0.4411
vehicle.bus.rigid	580	242	0.0239	0.0218	0.9150
vehicle.car	9,383	6,547	0.3861	0.5908	1.5301
vehicle.construction	665	151	0.0274	0.0136	0.4979
vehicle.motorcycle	1,329	138	0.0547	0.0125	0.2277
vehicle.trailer	48	294	0.0020	0.0265	13.4318
vehicle.truck	2,111	1,261	0.0869	0.1138	1.3099
TOTAL	**24,300**	**11,081**	**1.0000**	**1.0000**	–

J.3 Computation of τ_{CP}

J.3.1 Over a marginal confidence region

The maximum of $\tau_{CP}(y_0; \mathbf{\Omega}_{\mathrm{LP}})$ is achieved at one of the vertices because the property of $\tau_{CP}(y_0, \boldsymbol{\omega})$ and its monotonicity in each ω_k, resulting a coordinate-wise procedure to compute its maximum. In our study, we found the exact maximum for $\tau_{CP}(y_0; \mathbf{\Omega}_{\mathrm{LP}})$ and for $\tau_{CP}(y_0; \mathbf{\Omega}_{\mathrm{GE}})$.

J.3.2 Over a joint confidence region

The supremum over $\mathbf{\Omega}$ can be computed exactly using fractional programming iteratively. Denote the score as r_i and the label as $\widetilde{y}_i$ in the calibration set for $i = 1, ..., m_1$. Given the weight $\boldsymbol{\omega}^{(t)}$ at time t, let $\tau^{(t)} = \tau_{\mathrm{CP}}(\widetilde{y}_0; \boldsymbol{\omega}^{(t)})$. By definition, $\tau^{(t)}$ is a $(1-\alpha)$-quantile of a CDF $F(x; \boldsymbol{\omega}) = \frac{\sum_{i:r_i \le x} \boldsymbol{\omega}_{\widetilde{y}_i}}{\sum_{i=1}^{m_1} \boldsymbol{\omega}_{\widetilde{y}_i}}$. Then we update $\boldsymbol{\omega}^{(t)}$ with $\boldsymbol{\omega}^{(t+1)} = \arg\min_{\boldsymbol{\omega} \in \mathbf{\Omega}} F(\tau^{(t)}; \boldsymbol{\omega}) = \arg\min_{\boldsymbol{\omega} \in \mathbf{\Omega}} \frac{\sum_{i:r_i \le \tau^{(t)}} \boldsymbol{\omega}_{\widetilde{y}_i}}{\sum_{i=1}^{m_1} \boldsymbol{\omega}_{\widetilde{y}_i}}$, which is obtained via fractional programming. If $F(\tau^{(t)}; \boldsymbol{\omega}^{(t+1)}) < 1-\alpha$, then $\tau^{(t)} < \tau_{\mathrm{CP}}(\widetilde{y}_0; \boldsymbol{\omega}^{(t+1)})$, so the supremum is not reached, and we proceed to the next iteration with t. Otherwise, $F(\tau^{(t)}; \boldsymbol{\omega}) \ge 1-\alpha$ for every $\boldsymbol{\omega} \in \mathbf{\Omega}$, so $\tau^{(t)}$ cannot be improved and is the exact supremum. In the worst case, the iteration has to be repeated m_1 times, which may be computationally intractable. Hence, we implemented an alternative Monte Carlo (MC) method.

Validity of the MC approximation. Let $\widehat{\tau}_{\mathrm{MC}}(\widetilde{y}_0; \mathbf{\Omega}) \equiv \max_{1 \le j \le n_{MC}} \tau_{\mathrm{CP}}(\widetilde{y}_0; \boldsymbol{\omega}_j)$ with $\boldsymbol{\omega}_j \in \mathbf{\Omega}$. Since $\mathbf{\Omega} \subset \mathbf{\Omega}_{\mathrm{LP}}$,

$$\widehat{\tau}_{\mathrm{MC}}(\widetilde{y}_0; \mathbf{\Omega}) \le \tau_{\mathrm{CP}}(\widetilde{y}_0; \mathbf{\Omega}) \le \tau_{\mathrm{CP}}(\widetilde{y}_0; \mathbf{\Omega}_{\mathrm{LP}}).$$

The finite-sample guarantee of Theorem 3.7 holds for the two exact thresholds on the right, but not for $\widehat{\tau}_{\mathrm{MC}}$, which may underestimate the exact supremum. The MC-based results for $\mathbf{\Omega}$ in Section 4 are therefore an empirical approximation of the exact procedure, while $F_{\mathrm{CP}}(\cdot; \mathbf{\Omega}_{\mathrm{LP}})$ is a conservative alternative that is computed exactly and retains the guarantee. The following argument shows that the MC approximation error vanishes as $n_{MC} \to \infty$.

When we compute the maximum over 1,000 Monte Carlo samples from $\mathbf{\Omega}$, we explain the estimation error as below. $\mathbb{P}[\max_{1 \le j \le n_{MC}} \tau_{CP}(y_0; \boldsymbol{\omega}_j) \le \sup_{\boldsymbol{\omega}} \tau_{CP}(y_0; \boldsymbol{\omega}) - \epsilon] \approx \exp[-f(\sup_{\boldsymbol{\omega}} \tau_{CP}(y_0; \boldsymbol{\omega}))\epsilon n_{MC}]$, where f denotes the PDF of $\tau_{CP}(y_0; \boldsymbol{\omega})$ over uniformly sampled $\boldsymbol{\omega} \in \mathbf{\Omega}$. For any positive $f(\cdot)$ and ϵ, this probability goes to zero when $n_{MC} \to \infty$. This implies the difference between the Monte Carlo estimated $\tau_{CP}(y_0; \mathbf{\Omega})$ and the true $\tau_{CP}(y_0; \mathbf{\Omega})$ goes to zero in probability as $n_{MC} \to \infty$.

K Additional experimental results

K.1 Accuracy and expected calibration error (ECE) of classifiers

Table S4: Summary of accuracy for 10 classifiers trained on each dataset.

Dataset	AGNews	CIFAR-10	MNIST	N24News	nuImages
Accuracy mean(std)	94.79%(0.14)	91.63%(0.92)	97.49%(0.37)	85.99%(0.43)	95.08%(0.27)

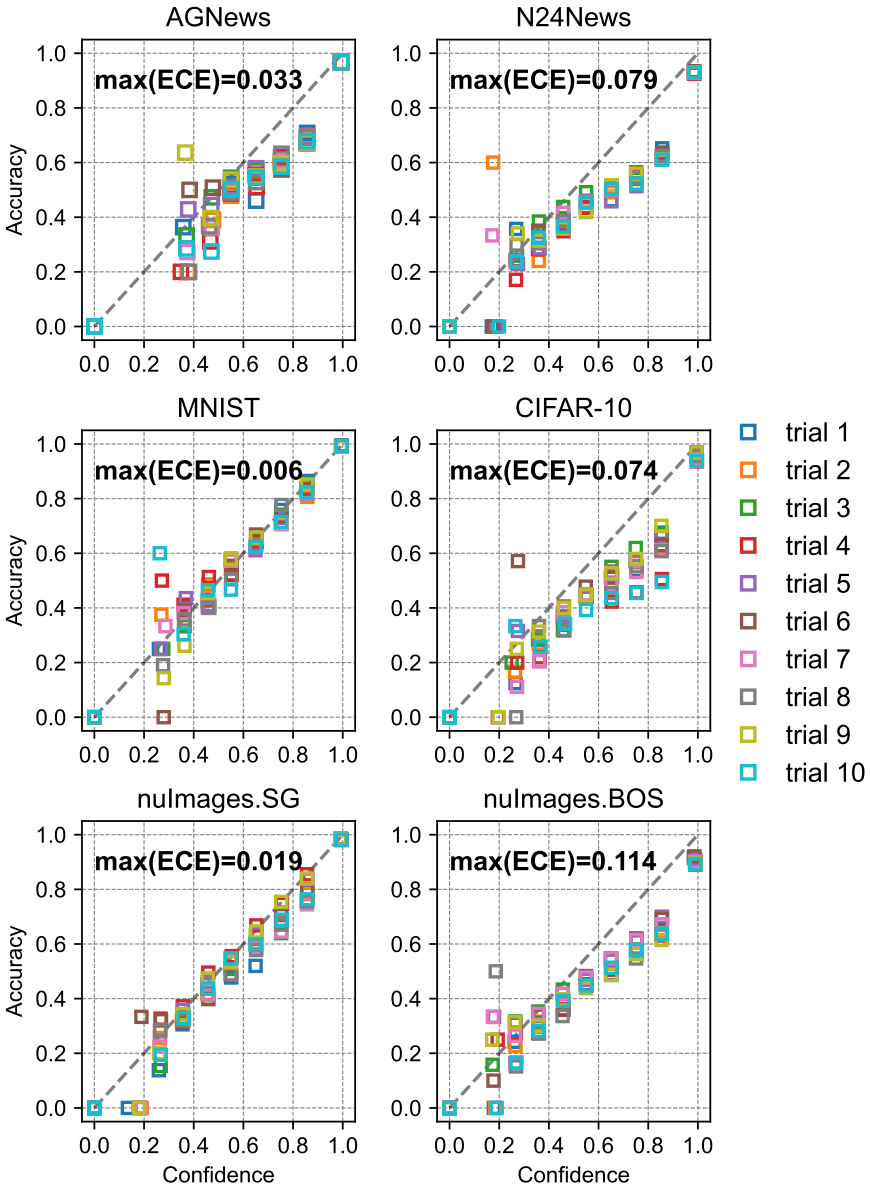


Figure S3: Expected calibration error (ECE) for each classifier on each dataset. The first column looks all classifiers are well-calibrated, whereas the classifiers on the second column datasets show different levels of uncalibrated performance.

K.2 Tightness of confidence regions

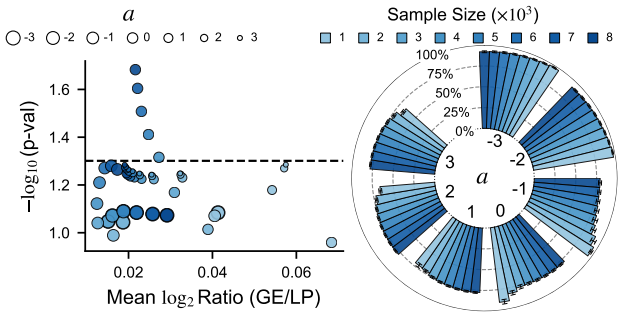


Figure S4: AGNews. (Left) Volcano plot of the interval length ratios (GE vs. LP) and their corresponding t-test significance levels. (Right) Volume ratio of $\mathbf{\Omega}$ to $\mathbf{\Omega}_{\mathrm{LP}}$ across varying shift intensities a and sample sizes.

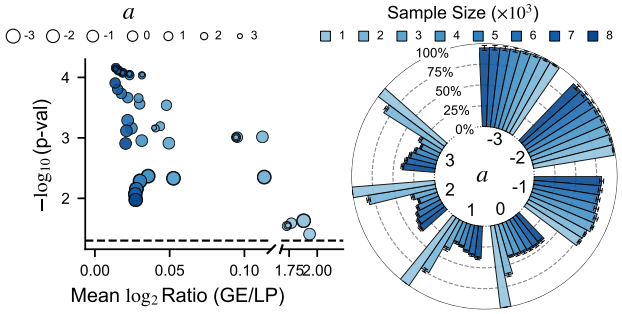


Figure S5: MNIST. (Left) Volcano plot of the interval length ratios (GE vs. LP) and their corresponding t-test significance levels. (Right) Volume ratio of $\mathbf{\Omega}$ to $\mathbf{\Omega}_{\mathrm{LP}}$ across varying shift intensities a and sample sizes.

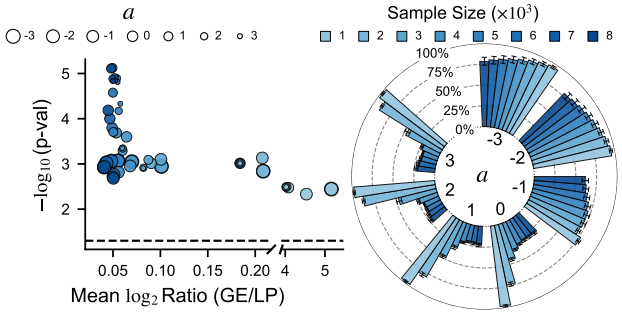


Figure S6: CIFAR-10. (Left) Volcano plot of the interval length ratios (GE vs. LP) and their corresponding t-test significance levels. (Right) Volume ratio of $\mathbf{\Omega}$ to $\mathbf{\Omega}_{\mathrm{LP}}$ across varying shift intensities a and sample sizes.

K.3 Assessment of downstream task

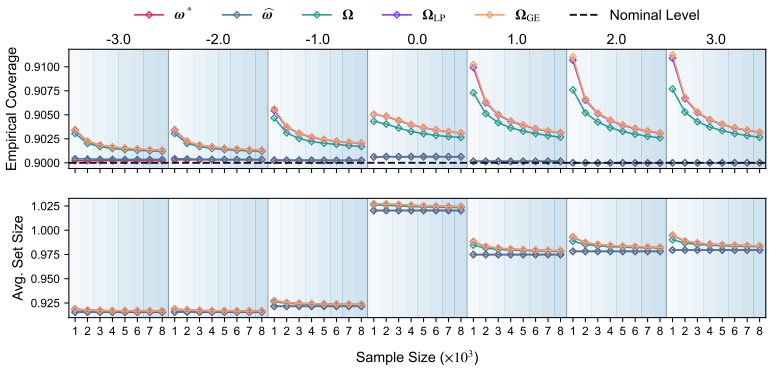


Figure S7: Evaluation of weighted conformal prediction on AGNews. The top panel displays empirical coverage across varying shift intensities (a) and sample sizes (n). The bottom panel shows average prediction set sizes.

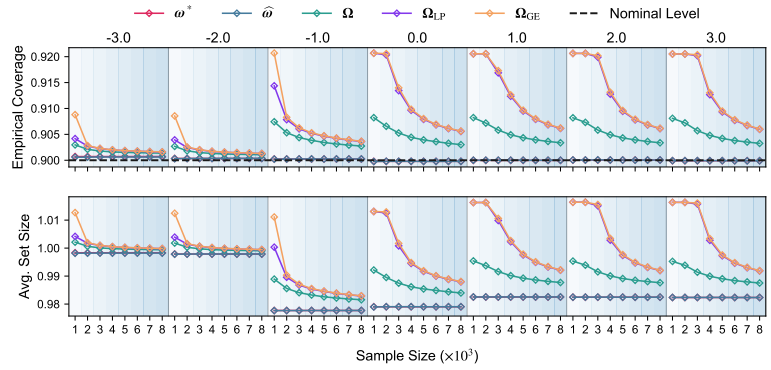


Figure S8: Evaluation of weighted conformal prediction on MNIST. The top panel displays empirical coverage across varying shift intensities (a) and sample sizes (n). The bottom panel shows average prediction set sizes.

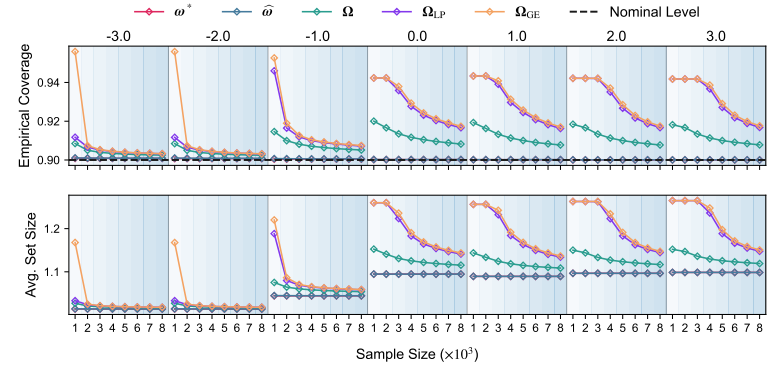


Figure S9: Evaluation of weighted conformal prediction on CIFAR-10. The top panel displays empirical coverage across varying shift intensities (a) and sample sizes (n). The bottom panel shows average prediction set sizes.

K.4 Runtime and scalability across the number of classes

Data generating process. We generate synthetic data from a two-dimensional Gaussian mixture model with $K \in \{2, 5, 10, 20, 50, 100, 200, 500\}$ classes. Each class shares the covariance matrix $\mathbf{\Sigma}_k = 0.5 \cdot \mathbf{I}_2$. To keep the classes reasonably separated as K grows, the first mean is placed at the origin and the remaining means on concentric circles of radii $1, 2, 3, \ldots$, with at most eight equally spaced points (separated by $45°$) per circle. We set $p(y) = 1/K$ and draw $q(y)$ from $\mathrm{Dir}(\mathbf{1}_K \cdot 10^{-1})$. For each trial, a Naive Bayes classifier g is trained on $2000K$ samples, and $1000K$ samples are drawn for each of the source and target domains with the stratified sampling procedure of Section 4.1. Each K is repeated 50 times.

Implementation. Since the constraints are identical across the $2K$ linear programs in (2)–(3), we solve them with the HiGHS dual simplex method, warm-starting each program from the previous basis. The solver is called from C++ through `pybind11` with zero-copy access to the constraint arrays, and the programs are distributed across CPU cores with OpenMP after releasing the Python global interpreter lock. The GE baseline is compiled with `numba 0.60.0`. All runs use the machine described in Appendix J with 5 CPU cores for LP.

Results. Table S5 reports the wall-clock time per run. $\mathbf{\Omega}_{\mathrm{LP}}$ is computed within a fraction of a second for $K \leq 200$ and in 1.72 seconds at $K = 500$. GE is faster for large K, but Table S6 shows that its marginal intervals are much longer: the mean $\log_2$ length ratio $K^{-1} \sum_k \log_2(l_{k,\mathrm{GE}}/l_{k,\mathrm{LP}})$ increases from 0.30 at $K = 2$ to about 10.9 for $K \geq 200$. Thus $\mathbf{\Omega}_{\mathrm{LP}}$ trades a moderate amount of computation for substantially tighter intervals.

Table S5: Runtime comparison (repetition $= 50$, unit: seconds per run).

K	2	5	10	20	50	100	200	500
LP mean	0.00023	0.00028	0.00040	0.00089	0.01160	0.04606	0.17202	1.72289
LP std	0.00005	0.00002	0.00005	0.00009	0.00145	0.00498	0.01689	0.24284
GE mean	0.00046	0.00086	0.00180	0.00331	0.01500	0.02974	0.03896	0.26335
GE std	0.00014	0.00009	0.00027	0.00023	0.00084	0.00192	0.00111	0.02558

Table S6: Mean $\log_2$ length ratio of GE to LP marginal intervals (repetition $= 50$).

K	2	5	10	20	50	100	200	500
mean	0.30	7.15	8.25	9.04	10.70	10.75	10.89	10.89
std	0.16	0.75	0.88	0.85	0.65	0.41	0.27	0.18

K.5 Numerical stability under extreme shift and rare classes

Empty regions. $\boldsymbol{\Omega}$ is not empty whenever the BBSE estimate $\widehat{\boldsymbol{\omega}} = \widehat{\mathbf{C}}^{-1}\widehat{\mathbf{q}}$ exists and satisfies $\widehat{\boldsymbol{\omega}} \geq \mathbf{0}$. It can be empty when $\{\boldsymbol{\omega} : \underline{\mathbf{C}}\boldsymbol{\omega} \leq \overline{\mathbf{q}},\ \overline{\mathbf{C}}\boldsymbol{\omega} \geq \underline{\mathbf{q}}\}$ does not intersect the nonnegative orthant. If this happens for a class j, we follow the strategy of BBSE and project onto the nonnegative orthant, setting $[\underline{\omega}_j, \overline{\omega}_j] = [0, 0]$.

Unbounded regions and rare source classes. If $\underline{C}_{ij} > 0$ for some i, then $\boldsymbol{\omega} \geq \mathbf{0}$ and $\sum_l \underline{C}_{il}\omega_l \leq \overline{q}_i$ imply $\omega_j \leq \overline{q}_i/\underline{C}_{ij}$. Hence $\boldsymbol{\Omega}$ can be unbounded only if $\underline{C}_{ij} = 0$ for all i in some column j, which typically reflects a very rare source class. In that case we replace the zero endpoints by a tiny positive value (e.g., 10^{-16}), and we additionally cap each upper bound at 500. Unbounded regions did not occur in our experiments, and we did not observe any adverse effect of this cap. More generally, a rare source class mainly affects the training of the classifier, which can be mitigated by standard imbalanced-learning techniques such as resampling or class-weighted losses; within MaC-LP, as long as the confusion matrix is full rank, its only effect is a looser upper bound for the corresponding weight.

Near-zero target proportions. When a target class proportion approaches zero, $\omega_k^* = q(k)/p(k)$ approaches zero rather than becoming large. Our benchmark design already covers this regime: Table S7 shows the target proportions generated by $\mathrm{Dir}(\mathbf{1}_K \cdot 10^a)$ with $K = 10$ (sorted within each draw, 100 repetitions). For $a \in \{-3, -2\}$, most target proportions are essentially zero. Across all benchmark settings, which combine these shifts with source and target sample sizes from 1,000 to 8,000, the LP solver had no failures.

Table S7: Target label proportions under the Dirichlet shift design ($K = 10$, repetition $= 100$).

a		0	1	2	3	4	5	6	7	8	9
-3	mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.989
	std	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.056	0.056
-2	mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.060	0.937
	std	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.014	0.119	0.124
-1	mean	0.000	0.000	0.000	0.000	0.003	0.012	0.034	0.085	0.238	0.628
	std	0.000	0.000	0.000	0.001	0.007	0.023	0.040	0.070	0.127	0.194
0	mean	0.009	0.021	0.034	0.049	0.064	0.085	0.107	0.141	0.196	0.294
	std	0.008	0.012	0.016	0.019	0.019	0.022	0.024	0.029	0.036	0.076
1	mean	0.057	0.070	0.078	0.085	0.093	0.100	0.109	0.119	0.133	0.157
	std	0.011	0.009	0.008	0.008	0.007	0.007	0.007	0.008	0.013	0.020
2	mean	0.085	0.090	0.093	0.096	0.098	0.101	0.103	0.107	0.111	0.116
	std	0.005	0.003	0.003	0.002	0.002	0.002	0.003	0.003	0.004	0.004
3	mean	0.095	0.097	0.098	0.099	0.100	0.100	0.101	0.102	0.103	0.105
	std	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002

K.6 Controlled experiment under label shift violation

When $p(\boldsymbol{x} \mid y) \neq q(\boldsymbol{x} \mid y)$, the identity $\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*$ no longer holds, so neither the point estimate $\widehat{\boldsymbol{\omega}}$ nor Theorem 3.7 has a formal guarantee. To understand the empirical behavior of the confidence regions in this case, we generate two-dimensional Gaussian mixture data with $K = 5$, means $(0, 0)$, $(2, 2)$, $(-1, 2)$, $(4, 0)$, $(-2, -2)$, covariance $\mathbf{I}_2$, $p(y) = (0.1, 0.2, 0.2, 0.15, 0.35)^\top$ and $q(y) = (0.37, 0.23, 0.19, 0.2, 0.01)^\top$. In the violated setting, the target covariance is inflated to $1.2\,\mathbf{I}_2$. Tables S8 and S9 and Figure S10 report the conformal prediction results (mean and standard error), with "Oracle" using $\boldsymbol{\omega}^* = q(y)/p(y)$.

When label shift holds, all methods attain the nominal 90% coverage, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_{\mathrm{LP}}$ approach the oracle as the sample size grows. When it is violated, even the oracle weights and the point estimate fall below the nominal level at every sample size. The coverage of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_{\mathrm{LP}}$ stays above 90% for small and moderate sample sizes, because the confidence regions are wide enough to absorb the misspecification, but it declines toward the point-estimate level once the sampling uncertainty becomes small relative to the systematic error. This behavior is an empirical observation for this design and should not be read as a validity guarantee outside the label shift model.

Table S8: Conformal prediction when label shift holds (mean (standard error)).

	Sample size	50	500	5000	50000	100000
Coverage	Oracle	0.901 (0.002)	0.901 (0.002)	0.900 (0.002)	0.901 (0.002)	0.901 (0.002)
	$\widehat{\boldsymbol{\omega}}$ (BBSE)	0.897 (0.003)	0.901 (0.002)	0.901 (0.002)	0.901 (0.002)	0.900 (0.002)
	$\boldsymbol{\Omega}$	0.936 (0.002)	0.935 (0.002)	0.919 (0.002)	0.907 (0.002)	0.905 (0.002)
	$\boldsymbol{\Omega}_{\mathrm{LP}}$	0.939 (0.002)	0.940 (0.002)	0.928 (0.002)	0.909 (0.002)	0.907 (0.002)
Set size	Oracle	1.621 (0.008)	1.621 (0.008)	1.619 (0.008)	1.620 (0.008)	1.621 (0.008)
	$\widehat{\boldsymbol{\omega}}$ (BBSE)	1.611 (0.013)	1.624 (0.008)	1.618 (0.008)	1.619 (0.008)	1.620 (0.008)
	$\boldsymbol{\Omega}$	1.846 (0.015)	1.836 (0.014)	1.721 (0.010)	1.652 (0.008)	1.644 (0.008)
	$\boldsymbol{\Omega}_{\mathrm{LP}}$	1.872 (0.016)	1.872 (0.016)	1.777 (0.012)	1.663 (0.009)	1.651 (0.008)

Table S9: Conformal prediction when label shift is violated (mean (standard error)).

	Sample size	50	500	5000	50000	100000
Coverage	Oracle	0.872 (0.002)	0.873 (0.002)	0.872 (0.002)	0.873 (0.002)	0.872 (0.002)
	$\widehat{\boldsymbol{\omega}}$ (BBSE)	0.866 (0.003)	0.871 (0.002)	0.870 (0.002)	0.870 (0.002)	0.870 (0.002)
	$\boldsymbol{\Omega}$	0.913 (0.003)	0.911 (0.002)	0.891 (0.002)	0.878 (0.002)	0.875 (0.002)
	$\boldsymbol{\Omega}_{\mathrm{LP}}$	0.917 (0.002)	0.917 (0.002)	0.901 (0.002)	0.880 (0.002)	0.877 (0.002)
Set size	Oracle	1.619 (0.008)	1.619 (0.008)	1.617 (0.008)	1.618 (0.008)	1.619 (0.008)
	$\widehat{\boldsymbol{\omega}}$ (BBSE)	1.597 (0.012)	1.611 (0.008)	1.605 (0.007)	1.606 (0.007)	1.607 (0.007)
	$\boldsymbol{\Omega}$	1.840 (0.015)	1.827 (0.014)	1.705 (0.010)	1.639 (0.008)	1.631 (0.008)
	$\boldsymbol{\Omega}_{\mathrm{LP}}$	1.865 (0.015)	1.865 (0.015)	1.759 (0.012)	1.649 (0.008)	1.637 (0.008)

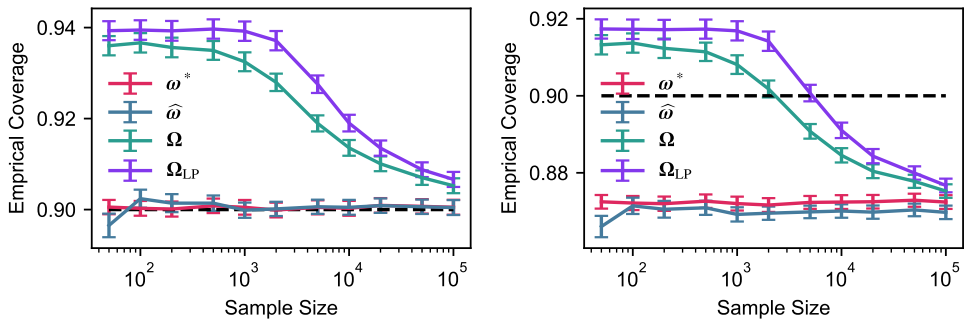


Figure S10: Empirical coverage (mean $\pm$ standard error) against sample size when label shift holds (left) and when the target covariance is inflated to $1.2\,\mathbf{I}_2$ (right).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions are (1) constraint-based framework of finding a confidence region of importance weights under label shift, (2) finite-sample valid optimality of the constructed confidence region, (3) downstream efficiency in prediction task using the confidence region, and (4) an extensive empirical validation on diverse benchmarks. The paper's scope is to find a confidence region under label shift rather than applying point estimation of the importance weights. These are reflected in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed limitations of the proposed method in the Discussion and Extended Discussion sections. These include the reliance on the label shift assumption (Section 4.2.3 and Appendix K.6), the scope of optimality under elementwise confidence intervals (Appendix B), and the Monte Carlo approximation of the conformal threshold (Remark 3.8).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proofs of Propositions, Theorems, and other theoretical claims are provided in Section F.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendix E, I, J contain all components to reproduce the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used in experiments are publicly available. The code is available at `https://github.com/carayms/maclp`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendix J contain all information to help understand the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We provide error bars for the volume ratio plots. Volcano plots contain the one-sided t-test p-value, which summarizes statistical significance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computation resources are described in Appendix J, and wall-clock runtimes of the LP and GE computations are reported in Section 4.3 and Appendix K.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The societal impacts of our paper is discussed in Appendix C.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper is a methodological one and we do not release data and model with potential risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The used data and methods are publicly available and properly credited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The proposed method is well-documented in the paper with licenses CC BY-NC 4.0.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.