From Matrix Inversion to Constraints

Provably Tighter Confidence Regions for Importance Weights in Label Shift

Mushan Li*, Kihyun Han*, Yanyuan Ma

NeurIPS 2026 · Code: `github.com/carayms/maclp`

Department of Statistics, The Pennsylvania State University

*Equal contribution

Label shift and importance weights

Setting. Labeled source data $\sim P$, unlabeled target data $\sim Q$.

- Label shift: $q(\mathbf{x} \mid y) = p(\mathbf{x} \mid y)$, but $q(y) \neq p(y)$.
- Examples: a disease rare in training hospitals but common in a new clinic; more pedestrians in a city than on highways.

Importance weights bridge the two domains:

$$\omega_y = \frac{q(y)}{p(y)}, \qquad \mathbb{E}_Q\{h(\mathbf{X}, Y)\} = \mathbb{E}_P\{\omega_Y\, h(\mathbf{X}, Y)\}.$$

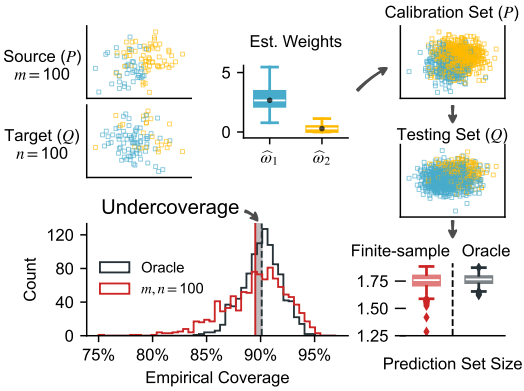
Used for weighted risk minimization and weighted conformal prediction.

The question

ω^* is unknown and estimated from *finite* samples.

How uncertain is $\widehat{\omega}$, and how do we account for it *without* being overly conservative?

Ignoring uncertainty in $\widehat{\omega}$ breaks coverage



- Weighted conformal prediction with the *true* weights: coverage centers on 90%.
- Plugging in an unbiased point estimate $\widehat{\omega}$ $(m, n = 100)$: systematic undercoverage.

Goal. A confidence region $\boldsymbol{\Omega}$ with

$$\mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}) \geq 1 - \delta,$$

valid in finite samples, and as small as possible.

Confusion matrix framework

Fix a black-box classifier $g : \mathcal{X} \to [K]$.

$$C^*_{ij} = \mathbb{P}_P\{g(\mathbf{X}) = i,\ Y = j\}, \qquad q^*_k = \mathbb{P}_Q\{g(\mathbf{X}) = k\} \quad \Longrightarrow \quad \boxed{\mathbf{C}^*\boldsymbol{\omega}^* = \mathbf{q}^*}$$

- Point estimate (BBSE): $\widehat{\boldsymbol{\omega}} = \widehat{\mathbf{C}}^{-1}\widehat{\mathbf{q}}$.
- Counts are binomial: $M_{ij} \sim \mathrm{Bin}(m, C^*_{ij})$, $N_k \sim \mathrm{Bin}(n, q^*_k)$.
- Elementwise intervals (Clopper–Pearson / Hoeffding): $\mathbf{C}^* \in [\underline{\mathbf{C}}, \overline{\mathbf{C}}] \equiv \mathcal{C}$, $\mathbf{q}^* \in [\underline{\mathbf{q}}, \overline{\mathbf{q}}] \equiv \mathcal{Q}$.
- Union bound: each of the $K^2 + K$ intervals at level $1 - \delta/(K^2 + K) \Rightarrow$ joint level $1 - \delta$.

Natural target

$$\boldsymbol{\Omega} = \big\{\boldsymbol{\omega} \geq \mathbf{0} :\ \exists\, \mathbf{C} \in \mathcal{C},\ \mathbf{q} \in \mathcal{Q} \text{ such that } \mathbf{C}\boldsymbol{\omega} = \mathbf{q}\big\}, \qquad \mathbb{P}(\boldsymbol{\omega}^* \in \boldsymbol{\Omega}) \geq 1 - \delta.$$

Prior work: interval Gaussian elimination is conservative

Si et al. (2023): solve the interval system $[\underline{\mathbf{C}}, \overline{\mathbf{C}}]\,\boldsymbol{\omega} = [\underline{\mathbf{q}}, \overline{\mathbf{q}}]$ by Gaussian elimination (GE) on intervals $\Rightarrow$ a box $\boldsymbol{\Omega}_{\mathrm{GE}}$.

- Still a matrix inversion mindset: row operations on intervals add relaxation at every step.
- The same entry can enter one bound at its *upper* end and at its *lower* end at the same time, so the bound is attained by no admissible matrix.
- Can break down ($\underline{C}^k_{kk} \le 0$) when the classifier is not diagonally dominant.

Two-class example

$C_{11} \in [0.3, 0.4]$, $C_{12} \in [0.1, 0.2]$, $C_{21} \in [0, 0.1]$, $C_{22} \in [0.4, 0.6]$, $q_1 \in [0.6, 0.7]$, $q_2 \in [0.3, 0.4]$.

GE: $\underline{\omega}_2 = 1/9 \approx 0.111$ (uses $\overline{C}_{21}$ *and* $\underline{C}_{21}$) MaC: $\underline{\omega}_2 = 2/17 \approx 0.118$, attained by a real $(\mathbf{C}, \mathbf{q})$.

MaC: from inversion to matrix constraints

Theorem (exact characterization)

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

In particular, Ω is a convex polytope described by $2K$ linear inequalities.

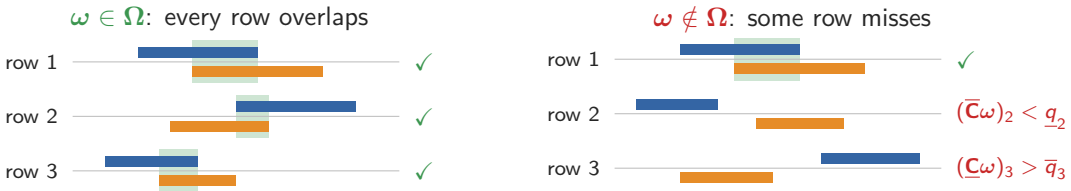
- No inversion, no point estimate: we only ask which ω are *compatible* with the intervals.
- Necessity is immediate; sufficiency builds an admissible $(\mathbf{C}, \mathbf{q})$ row by row.
- Coincides with the classical Oettli–Prager theorem on the nonnegative orthant; our contribution is the statistical construction and analysis built on it.
- Ω is the smallest region given the box uncertainty sets $\mathcal{C} \times \mathcal{Q}$.

Why the constraints are exact: row-wise interval overlap

Fix a candidate $\omega \geq \mathbf{0}$. As $\mathbf{C}$ ranges over the box $\mathcal{C}$, row i of $\mathbf{C}\omega$ sweeps an interval:

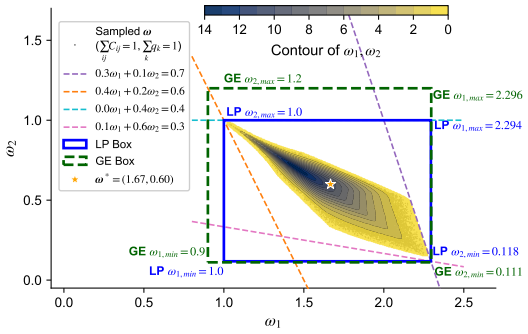
$$\{(\mathbf{C}\omega)_i : \mathbf{C} \in \mathcal{C}\} = \left[(\underline{\mathbf{C}}\omega)_i,\ (\overline{\mathbf{C}}\omega)_i\right].$$

reachable $(\mathbf{C}\omega)_i$ $[\underline{q}_i, \overline{q}_i]$



- Overlap in row i $\iff$ $(\overline{\mathbf{C}}\omega)_i \geq \underline{q}_i$ and $(\underline{\mathbf{C}}\omega)_i \leq \overline{q}_i$: the two inequalities of the theorem.
- Also sufficient: rows of $\mathbf{C}$ vary *independently* in the box, so each row can hit its own q_i in the overlap $\Rightarrow$ an admissible $(\mathbf{C}, \mathbf{q})$ with $\mathbf{C}\omega = \mathbf{q}$.

Scope of optimality: adding simplex constraints



Two-class example. Dashed lines: the 4 constraints of $\mathbf{\Omega}$. Blue / green: $\mathbf{\Omega}_{\mathrm{LP}}$ / $\mathbf{\Omega}_{\mathrm{GE}}$. Shaded: weights from admissible $(\mathbf{C}, \mathbf{q})$ that also satisfy $\sum_{i,j} C_{ij} = 1$, $\sum_k q_k = 1$.

- Given the box sets $\mathcal{C} \times \mathcal{Q}$, $\mathbf{\Omega}$ is optimal: every point of $\mathbf{\Omega}$ is attained by an admissible $(\mathbf{C}, \mathbf{q})$, so no valid region built from the same intervals can be smaller.
- Using more structure ($\sum_{i,j} C_{ij} = 1$, $\sum_k q_k = 1$, joint multinomial counts) can only shrink the region further (shaded points).
- The price: rows of $\mathbf{C}$ become coupled, and the region is non-convex with no linear description, so no LP.
- The box formulation balances tightness and tractability, and it is already much tighter than GE.

LP: marginal intervals, provably no wider than GE

For each class k, solve two linear programs over the same constraints:

$$\underline{\omega}_k = \min_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \omega_k, \qquad \overline{\omega}_k = \max_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \omega_k, \qquad \boldsymbol{\Omega}_{\mathrm{LP}} = \prod_{k=1}^{K} [\underline{\omega}_k, \overline{\omega}_k].$$

$\boldsymbol{\Omega}_{\mathrm{LP}}$ is the smallest axis-aligned box containing $\boldsymbol{\Omega}$ ("squeezing box").

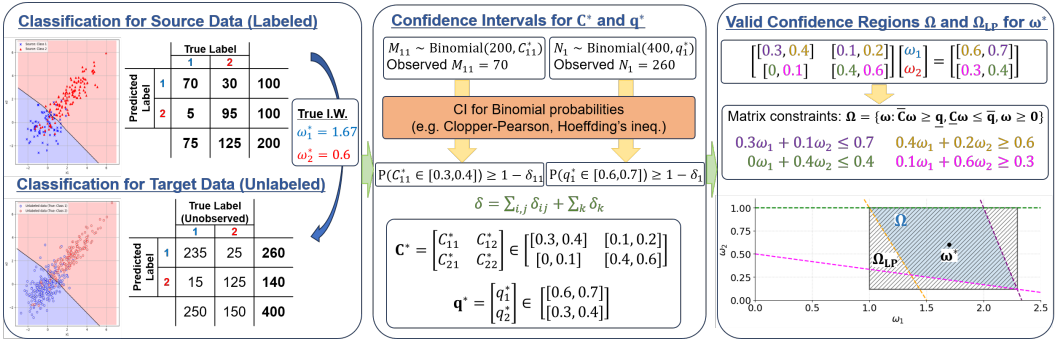
Corollary

When $\boldsymbol{\Omega}_{\mathrm{GE}}$ exists, $\boldsymbol{\Omega} \subset \boldsymbol{\Omega}_{\mathrm{LP}} \subset \boldsymbol{\Omega}_{\mathrm{GE}}$, and

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

- Same validity, tighter region. Diameter bounds for $\boldsymbol{\Omega}$ in $\|\cdot\|_\infty$ and $\|\cdot\|_2$ are given in the paper.

MaC-LP at a glance (two-class toy example)



(1) classify source and target data; (2) elementwise CIs for $\mathbf{C}^*$ and $\mathbf{q}^*$; (3) constraints give $\boldsymbol{\Omega}$ and the box $\boldsymbol{\Omega}_{\text{LP}}$, both containing $\boldsymbol{\omega}^*$.

Downstream: finite-sample conformal prediction

Weighted split conformal set for a candidate ω: $F_{\mathrm{CP}}(\tilde{\mathbf{x}}_0; \omega) = \{y : r(\tilde{\mathbf{x}}_0, y) \le \tau_{\mathrm{CP}}(y; \omega)\}$.

Robust version: take the worst case over the region,

$$\tau_{\mathrm{CP}}(y; \mathbf{\Omega}) = \sup_{\omega \in \mathbf{\Omega}} \tau_{\mathrm{CP}}(y; \omega).$$

Theorem (coverage)

If $\mathbb{P}(\omega^* \in \mathbf{\Omega}) \ge 1 - \delta$, then

$$\mathbb{P}\{\tilde{Y}_0 \in F_{\mathrm{CP}}(\tilde{\mathbf{X}}_0; \mathbf{\Omega})\} \ge 1 - \delta - \alpha$$

in finite samples.

Theorem (monotonicity)

$\mathbf{\Omega}_1 \subset \mathbf{\Omega}_2 \Rightarrow F_{\mathrm{CP}}(\cdot; \mathbf{\Omega}_1) \subset F_{\mathrm{CP}}(\cdot; \mathbf{\Omega}_2)$.

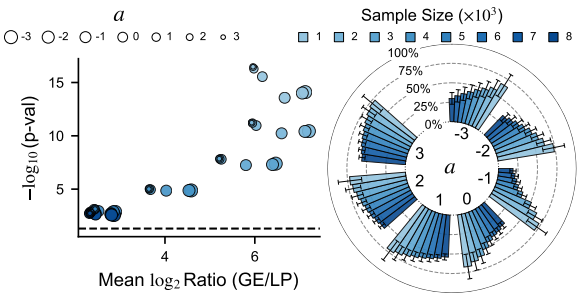
Hence $F_{\mathrm{CP}}(\cdot; \mathbf{\Omega}_{\mathrm{LP}}) \subset F_{\mathrm{CP}}(\cdot; \mathbf{\Omega}_{\mathrm{GE}})$: tighter weights give smaller sets.

For boxes ($\mathbf{\Omega}_{\mathrm{LP}}$, $\mathbf{\Omega}_{\mathrm{GE}}$) the supremum is computed exactly at a vertex. For the polytope $\mathbf{\Omega}$ we use a Monte Carlo maximum in experiments (empirical, not covered by the theorem).

Experimental design

- **Benchmarks:** AGNews (text), MNIST and CIFAR-10 (images), N24News (text + image); 10 classifiers each.
- **Real world:** nuImages, Singapore $\to$ Boston (natural geographic shift).
- **Simulated shift:** $\mathbf{v}_p \sim \mathrm{Dir}(\mathbf{1} \cdot 10^4)$, $\mathbf{v}_q \sim \mathrm{Dir}(\mathbf{1} \cdot 10^a)$, $a \in \{-3, \ldots, 3\}$ (smaller a = more extreme shift).
- Sample sizes 1,000 to 8,000 $\Rightarrow$ $7 \times 8 = 56$ settings, 1,000 repetitions each.
- Clopper–Pearson intervals at level $1 - \delta/(K^2 + K)$, $\delta = 1/5{,}000$; conformal level $1 - \alpha = 90\%$.
- **Compared:** oracle $\boldsymbol{\omega}^*$, point estimate $\widehat{\boldsymbol{\omega}}$ (BBSE), $\boldsymbol{\Omega}$ (MaC), $\boldsymbol{\Omega}_{\mathrm{LP}}$, $\boldsymbol{\Omega}_{\mathrm{GE}}$.

Tighter confidence regions (N24News)



Left: mean $\log_2$ length ratio GE/LP vs. t-test significance over 56 settings.

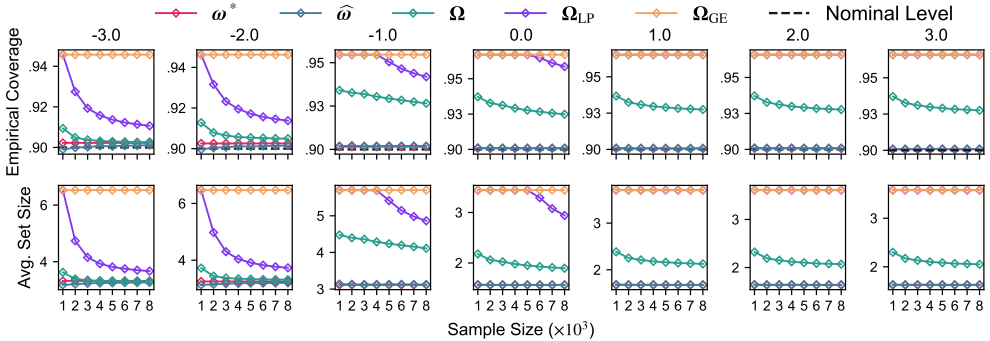
- LP intervals are consistently shorter than GE, with strong significance.

Right: volume of $\boldsymbol{\Omega}$ as a share of $\boldsymbol{\Omega}_{\mathrm{LP}}$.

- $\boldsymbol{\Omega}$ is a small fraction of its box, even under extreme shift: the constraints rule out incompatible weight combinations.

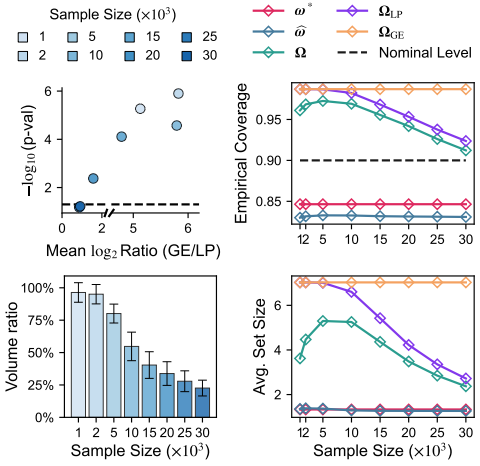
Similar results for AGNews, MNIST, CIFAR-10.

Smaller prediction sets at the nominal level (N24News)



- Ω stays above 90% coverage while its set size closely tracks the oracle.
- Ω_{LP}, Ω_{GE}: valid but larger sets. Point estimate $\widehat{\omega}$: undercovers under extreme shift.

Real-world shift: nuImages (Singapore → Boston)



- GE intervals are much wider, especially at small sample sizes; $\boldsymbol{\Omega}$ stays well inside $\boldsymbol{\Omega}_{\mathrm{LP}}$.
- Even the "true" weights undercover here: real data is noisier and label shift is likely violated.
- $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_{\mathrm{LP}}$, $\boldsymbol{\Omega}_{\mathrm{GE}}$ stay above 90%; $\boldsymbol{\Omega}$ gives the best balance of coverage and set size.
- Point estimate $\widehat{\boldsymbol{\omega}}$ undercovers throughout.

Results here are empirical; a controlled violation experiment in the paper shows the regions absorb part of the misspecification, a buffer that shrinks with sample size.

Scalability in the number of classes

Synthetic Gaussian mixtures, 50 runs per K (seconds per run).

K	2	5	10	20	50	100	200	500
LP time	0.0002	0.0003	0.0004	0.0009	0.012	0.046	0.172	1.72
GE time	0.0005	0.0009	0.0018	0.0033	0.015	0.030	0.039	0.263
mean $\log_2(l_{\mathrm{GE}}/l_{\mathrm{LP}})$	0.30	7.15	8.25	9.04	10.70	10.75	10.89	10.89

- LP stays under a second for $K \leq 200$ (HiGHS dual simplex, warm starts, parallel over the $2K$ programs).
- GE intervals become up to $2^{10.9} \approx 1{,}900\times$ longer as K grows.

Takeaways

- **Paradigm shift:** describe the importance weights by matrix constraints instead of inverting an interval system.
- **MaC:** exact, convex, finite-sample valid region; the smallest one given elementwise intervals.
- **LP:** marginal intervals provably no wider than interval Gaussian elimination.
- **Downstream:** finite-sample conformal coverage $\geq 1 - \delta - \alpha$, with smaller prediction sets.
- **Evidence:** text, image, multimodal benchmarks and real autonomous driving data.

Scope. Guarantees assume exact label shift; optimality is relative to box uncertainty sets; the Monte Carlo computation over $\mathbf{\Omega}$ is empirical.

Python package: `github.com/carayms/maclp`