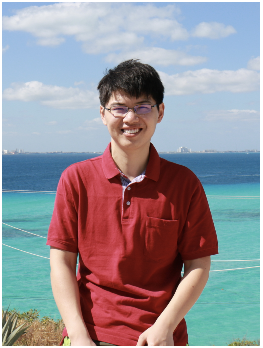
Distributionally robust risk evaluation with shape constraints

Yu Gui

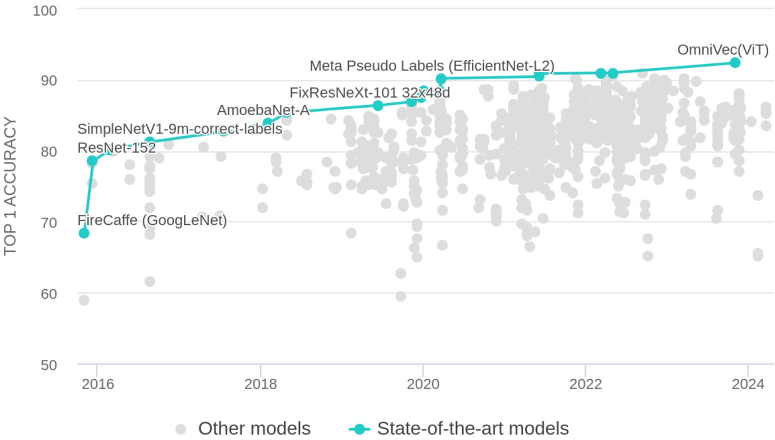
Department of Statistics and Data Science, the Wharton School

Rina Foygel Barber @UChicago



Cong Ma @UChicago

ImageNet Dataset



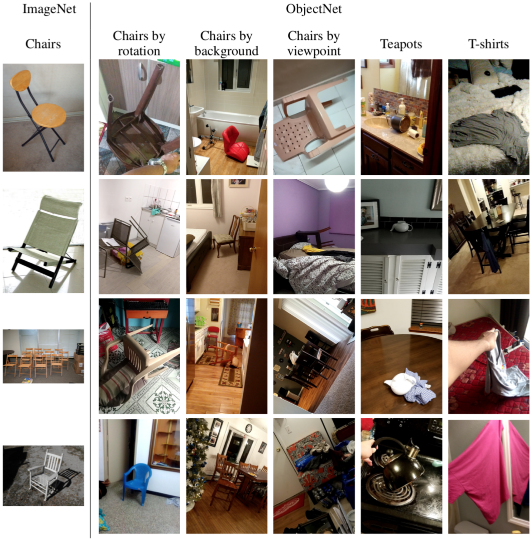
Leaderboard: image classification on ImageNet*

*Deng et al. (2009)

Question

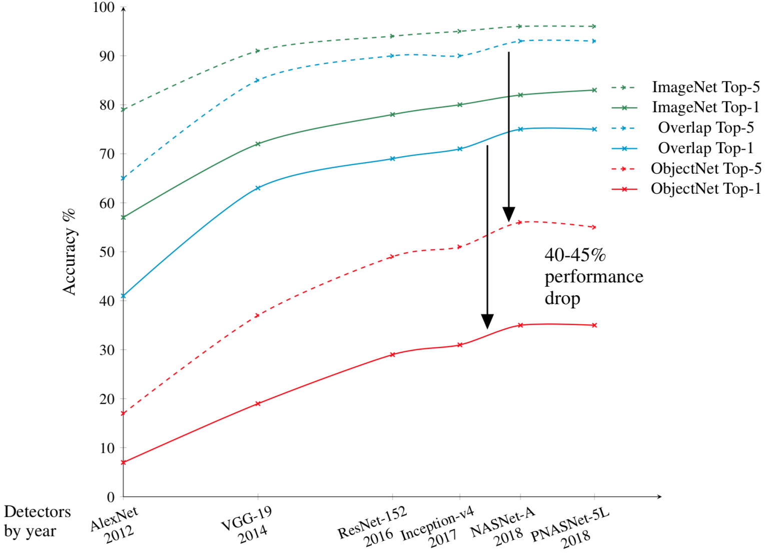
What if test distribution $\neq$ training distribution?

An example: ObjectNet[†]



[†]Barbu et al. (2019)

Performance on ObjectNet



Question

How to quantify the out-of-sample performance?

Statistical inference with distribution shift

$$\mathbb{E}_P[R_\alpha(X)] \leq \alpha \quad \overset{P^{\text{test}} \neq P}{\Longrightarrow} \quad \mathbb{E}_{P^{\text{test}}}[R_\alpha(X)] = \,?$$

Statistical inference with distribution shift

$$\mathbb{E}_P[R_\alpha(X)] \leq \alpha \quad \overset{P^{\text{test}} \neq P}{\Longrightarrow} \quad \mathbb{E}_{P^{\text{test}}}[R_\alpha(X)] = ?$$

Example: hypothesis test for $P \in \mathcal{H}_0$ with data from P^{test} (Thams et al., 2023)

- Risk function

$$R_\alpha(X) = \phi_\alpha(X)$$

- Valid type-I error control with data from P

$$\mathbb{P}_P(\phi_\alpha(X) = 1) \leq \alpha \quad \Longleftrightarrow \quad \mathbb{E}_P[R_\alpha(X)] \leq \alpha$$

Statistical inference with distribution shift

$$\mathbb{E}_P[R_\alpha(X)] \le \alpha \qquad \overset{P^{\text{test}} \neq P}{\Longrightarrow} \qquad \mathbb{E}_{P^{\text{test}}}[R_\alpha(X)] = ?$$

A concrete example: predictive inference under covariate shift[‡]

[‡](Vovk et al., 2005; Tibshirani et al., 2019).

Statistical inference with distribution shift

$$\mathbb{E}_P[R_\alpha(X)] \le \alpha \quad \overset{P^{\text{test}} \neq P}{\Longrightarrow} \quad \mathbb{E}_{P^{\text{test}}}[R_\alpha(X)] = ?$$

A concrete example: predictive inference under covariate shift[‡]

- Prediction set $\widehat{C}_{1-\alpha}$ constructed with a dataset $\mathcal{D}$ drawn from P
- Risk function
$$R_\alpha(X) = \mathbb{P}\left(Y \notin \widehat{C}_{1-\alpha}(X) \mid X\right)$$
- *Conformal prediction* CP: validity when $\{(X, Y)\} \cup \mathcal{D}$ is exchangeable (implies $X \sim P$)

$$\text{for any } \alpha \in (0,1) \qquad \mathbb{P}(Y \notin \widehat{C}_{1-\alpha}(X)) \le \alpha \quad \Longleftrightarrow \quad \mathbb{E}_P[R_\alpha(X)] \le \alpha$$

[‡](Vovk et al., 2005; Tibshirani et al., 2019).

Reweighting methods

“Estimable” distribution shift

“Estimable” distribution shift

$$\widehat{\mathbf{w}} \approx \frac{\mathrm{d}P^{\text{test}}}{\mathrm{d}P}$$

"Estimable" distribution shift

- Covariate shift[†]: choose β

$$\mathbb{E}_{P^{\text{test}}}[R_\beta(X)] = \mathbb{E}_P\left[\frac{\mathsf{d}P^{\text{test}}}{\mathsf{d}P}(X)R_\beta(X)\right] \approx \mathbb{E}_P[\widehat{\mathbf{w}}(X)R_\beta(X)] \leq \alpha$$

[†] (Sugiyama, 2011)

Reweighting methods

"Estimable" distribution shift

- An example: missing at random (MAR)[†]

$$\mathbf{M} \in \mathbb{R}^{d_1 \times d_2} \qquad M_{ij} \text{ is observed independently with probability } p_{ij} \in (0,1)$$

- $\mathcal{S} = \{(i,j) : \; M_{i,j} \text{ is observed}\}$ and $(i_*, j_*) \mid \mathcal{S} \sim \mathrm{Unif}(\mathcal{S}^c)$

$$\mathbb{P}\Big((i_*, j_*) = (i_k, j_k) \mid \mathcal{S} \,\cup\, \{(i_*, j_*)\} = \{(i_l, j_l)\}_{l \le n+1} \Big) = \frac{(1 - p_{i_k j_k})/p_{i_k j_k}}{\sum_{l \le n+1} (1 - p_{i_l j_l})/p_{i_l j_l}}$$

$$\text{importance sampling with "density ratio"} \quad = \quad \frac{1 - p_{i,j}}{p_{i,j}}$$

missingness $\approx$ distribution shift between sampled and unsampled populations

[†]Gui, Yu, Rina Barber, and Cong Ma. "Conformalized matrix completion." Advances in Neural Information Processing Systems 36 (2023): 4820-4844.

Reweighting methods

"Estimable" distribution shift

- Covariate shift: choose β

$$\mathbb{E}_{P^{\text{test}}}[R_\beta(X)] = \mathbb{E}_P\left[\frac{\mathsf{d}P^{\text{test}}}{\mathsf{d}P}(X)R_\beta(X)\right] \approx \mathbb{E}_P[\widehat{\mathbf{w}}(X)R_\beta(X)] \le \alpha$$

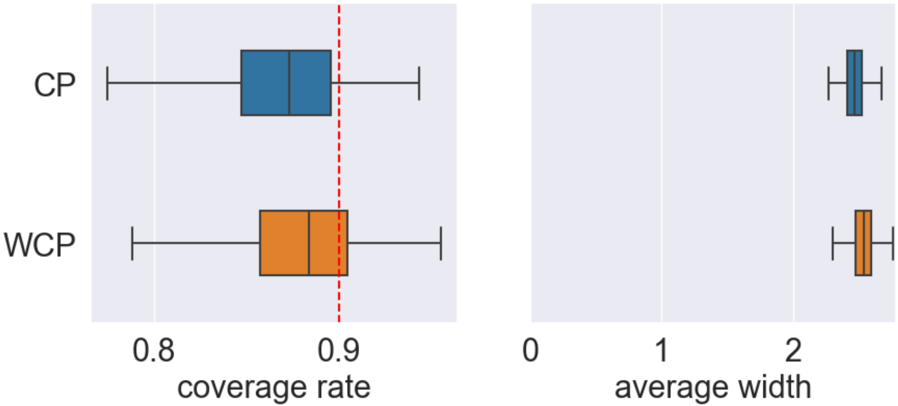
"Estimable" distribution shift

- Covariate shift: choose β

$$\mathbb{E}_{P^{\text{test}}}[R_\beta(X)] = \mathbb{E}_P\left[\frac{\mathrm{d}P^{\text{test}}}{\mathrm{d}P}(X)R_\beta(X)\right] \approx \mathbb{E}_P[\widehat{\mathbf{w}}(X)R_\beta(X)] \leq \alpha$$

An inevitable error term $\|\mathbf{w} - \widehat{\mathbf{w}}\|_1$!

An example with a wine quality dataset[§]: white wine (4898) vs red wine (1599)



[§] Cortez et al. (2009), https://archive.ics.uci.edu/dataset/186/wine+quality.

Distributionally robust learning (DRL)[†]

Worst-case control: choose β

$$\mathbb{E}_{P^{\text{test}}}[R_\beta(X)] \leq \sup_{Q' \in \mathcal{Q}} \mathbb{E}_{Q'}[R_\beta(X)] \leq \alpha \qquad \text{if } P^{\text{test}} \in \mathcal{Q} \qquad \text{(DRL)}$$

[†] El Ghaoui and Lebret (1997); Ben-Tal and Nemirovski (1998); Lam (2016); Duchi and Namkoong (2019); Blanchet et al. (2019)

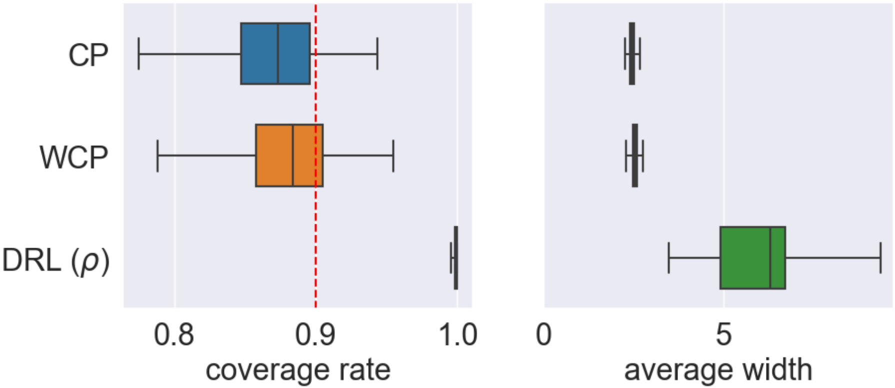
Distributionally robust learning (DRL)[†]

Worst-case control: choose β

$$\mathbb{E}_{P^{\text{test}}}[R_\beta(X)] \leq \sup_{Q' \in \mathcal{Q}} \mathbb{E}_{Q'}[R_\beta(X)] \leq \alpha \qquad \text{if } P^{\text{test}} \in \mathcal{Q} \qquad \text{(DRL)}$$

Too conservative/pessimistic!

[†] El Ghaoui and Lebret (1997); Ben-Tal and Nemirovski (1998); Lam (2016); Duchi and Namkoong (2019); Blanchet et al. (2019)



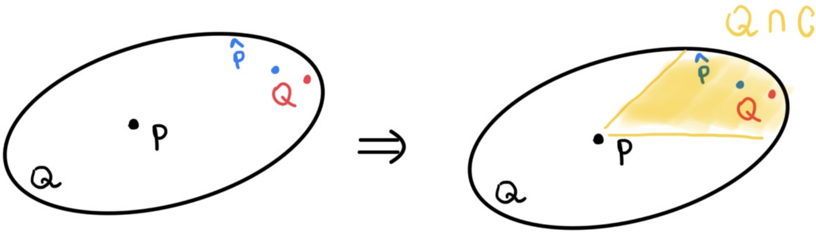
$$\rho \approx D_{\mathrm{KL}}(P^{\mathrm{test}} \mid\mid P)$$

A middle ground?

Misspecification of reweighting methods VS *Overly pessimism* of (DRL)

A middle ground?

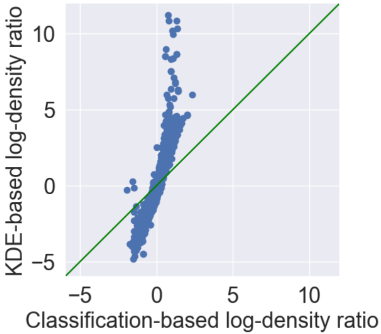
Misspecification of reweighting methods VS *Overly pessimism* of (DRL)



A middle ground?

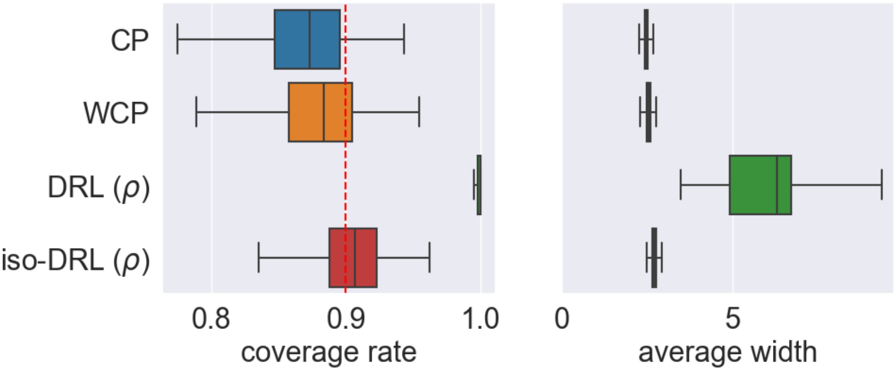
Misspecification of reweighting methods VS *Overly pessimism* of (DRL)

Fitted density ratio $\widehat{w}$ vs $\frac{\mathrm{d}P^{\mathrm{test}}}{\mathrm{d}P}$ proxy: an illustrative example with a wine quality dataset



- Biased but exhibits an approximately isotonic trend
- Under(Over)-represented regions in P^{test} are revealed by the under(over)-represented regions in $\widehat{P}$
- Use the side information to construct an additional cone constraint

$$\mathcal{Q}^{\mathrm{iso}}_{\widehat{w}} = \{Q' : \ \mathrm{d}Q'/\mathrm{d}P \text{ is isotonic in } \widehat{w}\}$$



$$\rho \approx D_{\mathrm{KL}}(P^{\mathrm{test}} \mid\mid P)$$

ISO-DRL UNDER GENERAL PARTIAL ORDERS

- Under any fixed partial order $\preccurlyeq$ on $\mathcal{X} \subseteq \mathbb{R}^d$

$$\mathcal{Q}^{\text{iso}}_{\preccurlyeq} = \{Q' : \ \mathrm{d}Q'/\mathrm{d}P \text{ is isotonic under } \preccurlyeq\}$$

iso-DRL under general partial orders

- Under any fixed partial order $\preccurlyeq$ on $\mathcal{X} \subseteq \mathbb{R}^d$

$$\mathcal{Q}_{\preccurlyeq}^{\mathrm{iso}} = \{Q' : \; \mathrm{d}Q'/\mathrm{d}P \text{ is isotonic under } \preccurlyeq\}$$

- iso-DRL chooses β such that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}_{\preccurlyeq}^{\mathrm{iso}}} \mathbb{E}_Q \left[R_\beta(X)\right] \leq \alpha \qquad \text{(iso-DRL)}$$

iso-DRL under general partial orders

Question

How to solve the cone-constrained optimization problem (iso-DRL)?

Question

How to solve the cone-constrained optimization problem (iso-DRL)?

- At the population level: a cone-constrained optimization problem in function space?
- With a finite sample: efficient computation? consistent estimate?

ISO-DRL UNDER GENERAL PARTIAL ORDERS

Question

How to solve the cone-constrained optimization problem (iso-DRL)?

- At the population level: a cone-constrained optimization problem in function space?
- With a finite sample: efficient computation? consistent estimate?

Improvements over DRL?

An equivalent formulation

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X) \right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R^{\text{iso}}_\beta(X) \right] \tag{Equiv}$$

$$R^{\text{iso}}_\beta(X) = \operatorname*{argmin}_{a \in \mathcal{C}^{\text{iso}}_{\preccurlyeq}} \int (a - R_\beta)^2 \, \mathrm{d}P$$

$\mathcal{C}^{\text{iso}}_{\preccurlyeq}$ = cone of isotonic functions under $\preccurlyeq$

An equivalent formulation

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X) \right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R^{\text{iso}}_\beta(X) \right] \quad \text{(Equiv)}$$

- **Examples of $\mathcal{Q}$**
  - Γ-marginal selection model in sensitivity analysis (Rosenbaum, 1987; Tan, 2006)

$$\mathcal{Q} = \left\{ Q : \Gamma^{-1} \le \frac{\mathrm{d}Q}{\mathrm{d}P}(X) \le \Gamma \text{ almost surely} \right\} \quad (\Gamma\text{-MS})$$

  - f-divergence constrained distribution shift (Ben-Tal and Nemirovski, 1998; El Ghaoui and Lebret, 1997; Duchi and Namkoong, 2019)

$$\mathcal{Q} = \{ Q : D_f(Q \,||\, P) \le \rho \} \quad (f\text{-Div})$$

An equivalent formulation

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\mathrm{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X) \right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R^{\mathrm{iso}}_\beta(X) \right] \tag{Equiv}$$

- **Examples of $\mathcal{Q}$**
  - Γ-marginal selection model in sensitivity analysis
  - f-divergence constrained distribution shift
- **Two sources of computational costs are separated**:

An equivalent formulation

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\mathrm{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X)\right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R_\beta^{\mathrm{iso}}(X)\right] \tag{Equiv}$$

- **Examples of $\mathcal{Q}$**
  - Γ-marginal selection model in sensitivity analysis
  - f-divergence constrained distribution shift
- **Two sources of computational costs are separated**:
  - $\mathcal{Q} \longrightarrow$ computational cost in solving (DRL) with R_β^{iso}

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X)\right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R^{\text{iso}}_\beta(X)\right] \tag{Equiv}$$

- **Examples of $\mathcal{Q}$**
  - Γ-marginal selection model in sensitivity analysis
  - f-divergence constrained distribution shift
- **Two sources of computational costs are separated**:
  - $\mathcal{Q} \longrightarrow$ computational cost in solving (DRL) with R^{iso}_β
  - $\mathcal{Q}^{\text{iso}}_{\preccurlyeq} \longrightarrow$ isotonic projection of R

An equivalent formulation

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X) \right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R^{\text{iso}}_\beta(X) \right] \tag{Equiv}$$

- **Examples of $\mathcal{Q}$**
  - Γ-marginal selection model in sensitivity analysis
  - f-divergence constrained distribution shift
- **Two sources of computational costs are separated**:
  - $\mathcal{Q} \longrightarrow$ computational cost in solving (DRL) with R^{iso}_β
  - $\mathcal{Q}^{\text{iso}}_{\preccurlyeq} \longrightarrow$ isotonic projection of R
- (Equiv) holds at **both population and sample levels**: *reference measure can be P or $\widehat{P}_n$*

An equivalent formulation

Gui et al, 2024 (Theorem 3.1)

Under regularity conditions on $\mathcal{Q}$, it holds that

$$\sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q \left[R_\beta(X) \right] = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[R^{\text{iso}}_\beta(X) \right] \tag{Equiv}$$

Shape constraints protect against "nonsmooth" or adversarial distribution shifts

Finite-sample estimate

$$\Delta^{\text{iso}}(R;\mathcal{Q}) = \sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q\left[R_\beta(X)\right] = \sup_{w_{\#}P \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_P\left[w(X) \cdot R_\beta(X)\right]$$

Finite-sample estimate

$$\Delta^{\text{iso}}(R;\mathcal{Q}) = \sup_{Q \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q\left[R_\beta(X)\right] = \sup_{w_\# P \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_P\left[w(X) \cdot R_\beta(X)\right]$$

$$\Downarrow$$

Finite-sample estimate

$$\Delta^{\text{iso}}(R;\mathcal{Q}) = \sup_{Q\in\mathcal{Q}\cap\mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q\left[R_\beta(X)\right] = \sup_{w_\# P\in\mathcal{Q}\cap\mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_P\left[w(X)\cdot R_\beta(X)\right]$$

$$\Downarrow$$

$$\widehat{\Delta}^{\text{iso}}(\mathcal{Q}) = \sup_{w_\# P\in\mathcal{Q}\cap\mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_{\widehat{P}_n}\left[w(X)\cdot r_\beta(X)\right]$$

$$\overset{\text{(Equiv)}}{=} \sup_{w_\# P\in\mathcal{Q}} \mathbb{E}_{\widehat{P}_n}\left[w(X)\cdot \widehat{r}^{\,\text{iso}}_\beta(X)\right]$$

Finite-sample estimate

$$\Delta^{\text{iso}}(R;\mathcal{Q}) = \sup_{Q\in\mathcal{Q}\cap\mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_Q\left[R_\beta(X)\right] = \sup_{w_\# P\in\mathcal{Q}\cap\mathcal{Q}^{\text{iso}}_{\preccurlyeq}} \mathbb{E}_P\left[w(X)\cdot R_\beta(X)\right]$$

$$\Downarrow$$

$$\widehat{\Delta}^{\text{iso}}(\mathcal{Q}) = \sup_{w_\# P\in\mathcal{Q}\cap\mathcal{Q}^{\text{iso}}_{\preccurlyeq},\|w\|_\infty\le\Omega} \mathbb{E}_{\widehat{P}_n}\left[w(X)\cdot r_\beta(X)\right]$$

$$\overset{\text{(Equiv)}}{=} \sup_{w_\# P\in\mathcal{Q},\|w\|_\infty\le\Omega} \mathbb{E}_{\widehat{P}_n}\left[w(X)\cdot \widehat{r}^{\,\text{iso}}_\beta(X)\right]$$

Finite-sample estimate

$$\widehat{\Delta}^{\text{iso}}(\mathcal{Q}) = \sup_{w_{\#}P \in \mathcal{Q} \cap \mathcal{Q}^{\text{iso}}_{\preccurlyeq}, \|w\|_\infty \leq \Omega} \mathbb{E}_{\widehat{P}_n}\left[w(X) \cdot r_\beta(X)\right]$$

$$\overset{\text{(Equiv)}}{=} \sup_{w_{\#}P \in \mathcal{Q}, \|w\|_\infty \leq \Omega} \mathbb{E}_{\widehat{P}_n}\left[w(X) \cdot \widehat{r}^{\text{iso}}_\beta(X)\right]$$

- $r_\beta(X)$ is a noisy observation of $R_\beta(X)$
- $\widehat{r}^{\text{iso}}_\beta(X)$ is the isotonic projection of $r_\beta(X)$ w.r.t. $\widehat{P}_n$

Finite-sample estimate

Gui et al, 2024 (Theorem 4.4, informal)

For both (Γ-MS) and (f-Div) with adequately large Ω,

$$\left|\Delta^{\text{iso}}(R;\mathcal{Q}) - \widehat{\Delta}^{\text{iso}}(\mathcal{Q})\right| \lesssim \mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preccurlyeq,\Omega}) + \sqrt{\frac{\log n}{n}}$$

Finite-sample estimate

Gui et al, 2024 (Theorem 4.4, informal)

For both (Γ-MS) and (f-Div) with adequately large Ω,

$$\left|\Delta^{\text{iso}}(R;\mathcal{Q}) - \widehat{\Delta}^{\text{iso}}(\mathcal{Q})\right| \lesssim \mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preccurlyeq,\Omega}) + \sqrt{\frac{\log n}{n}}$$

Bounding the Rademacher complexity

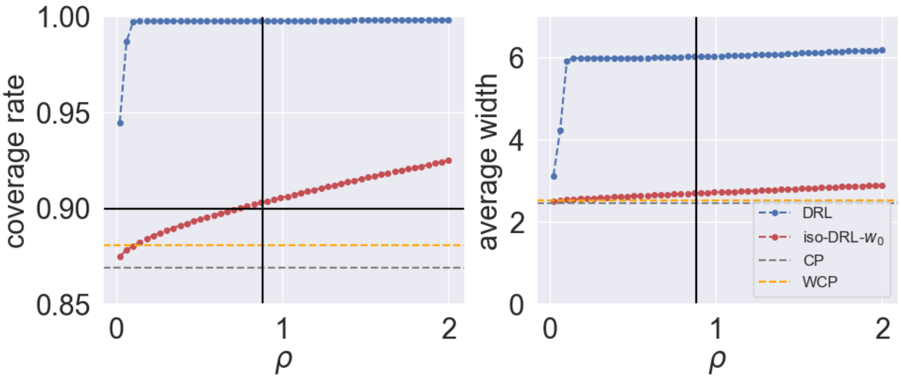
- $d = 1$ (Chatterjee and Lafferty, 2019)

$$\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preccurlyeq,\Omega}) \lesssim n^{-1/2}$$

- $d \geq 2$ with componentwise order, i.e. $\mathbf{x} \preccurlyeq \mathbf{z}$ iff $x_i \leq z_i$ for all $i \in [d]$ (Han et al., 2019)

$$\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preccurlyeq,\Omega}) \lesssim n^{-1/d}$$

Wine quality data set with varying ρ



Numerical simulations

- Conditional distribution

$$Y \mid X \sim \mathcal{N}\left(X^\top \beta + \sin(X_1) + 0.2 X_3^2, 1\right)$$

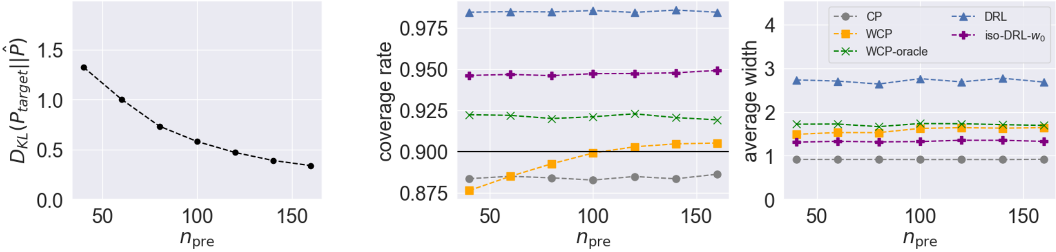
- Marginal distributions

$$\begin{cases} \text{training distribution} & P: \quad X \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), \\ \text{test distribution} & P^{\text{test}}: \quad X \sim \mathcal{N}(\mu, \mathbf{I}_d + \zeta \cdot \mathbf{\Omega}), \end{cases}$$

- $d = 5$, $\mathbf{\Omega} = \mathbf{1}\mathbf{1}^\top$, and $\mu = (2/\sqrt{d}) \cdot (1, \cdots, 1)^\top$
- $\zeta = 0$: well-specified $\widehat{w}$ via logistic regression; $\zeta > 0$: misspecified $\widehat{w}$

Varying splitting ratio η: well-specified density ratio

Estimated density ratio $\widehat{w}$ via logistic regression using $\eta \times 100\%$ data



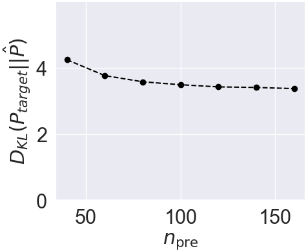
(a) $D_{\mathrm{KL}}(P^{\mathrm{test}}||\widehat{P})$ versus n_{pre}

(b) Comparison with varying n_{pre}

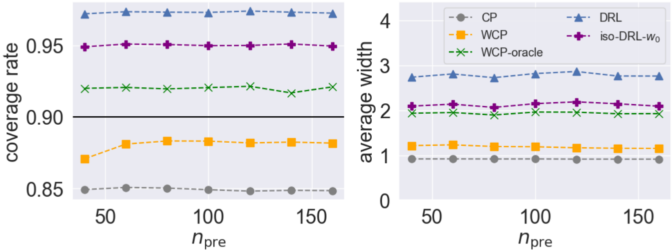
Results with well-specified density ratio $(\zeta = 0)$¶

¶ $\rho = \rho^* = D_{\mathrm{KL}}(P^{\mathrm{test}}||P)$

Varying splitting ratio η: misspecified density ratio



(a) $D_{\mathrm{KL}}(P^{\mathrm{test}}||\widehat{P})$ versus n_{pre}



(b) Comparison with varying n_{pre}

Results with misspecified density ratio ($\zeta = 1$)

Summary

- Distribution shift can harm the validity of statistical inference
- By incorporating shape constraints, (iso-DRL) offers one way to balance the misspecification of reweighting methods and the pessimism of DRL

Thank you!

Condition on $\mathcal{Q}$

- Change of "variable"

$$Q \in \mathcal{Q} \quad \text{if and only if} \quad w_{\#}P \in \mathcal{B}$$

- Convex ordering ($\overset{cvx}{\preccurlyeq}$): for two distributions Q and P,

$$Q \overset{cvx}{\preccurlyeq} P \quad \text{if and only if} \quad \mathbb{E}_Q[\psi(X)] \leq \mathbb{E}_P[\psi(X)] \quad \text{for any convex function } \psi$$

Condition (Closedness under convex ordering)

The set $\mathcal{B}$ is closed under convex ordering such that

$$\text{if } Q' \in \mathcal{B}, \text{ then } Q \in \mathcal{B} \text{ for any } Q \overset{cvx}{\preccurlyeq} Q' \qquad \text{(conditions)}$$

A DETOUR: CONFORMAL PREDICTION

- Any distribution $P_{X,Y}$ (completely unknown)
- $\{(X_i, Y_i)\}_{i \leq n+1} \sim P_{X,Y}$ are exchangeable with unobserved Y_{n+1}

Finite-sample validity

Construct marginal confidence intervals any $\alpha \in (0, 1)$

$$\mathbb{P}\left(Y_{n+1} \in C_{1-\alpha}(X_{n+1})\right) \geq 1 - \alpha$$

Split conformal prediction

- Split dataset into a training set and a calibration set $\mathcal{D}_{\text{calib}} = \{(X_i, Y_i)\}_{i \leq n}$
- Prefit $\widehat{\mu} : \mathcal{X} \to \mathcal{Y}$ on the training set $\Longrightarrow$ nonconformity score $R(x, y)$

Split conformal prediction

- Split dataset into a training set and a calibration set $\mathcal{D}_{\text{calib}} = \{(X_i, Y_i)\}_{i \leq n}$
- Prefit $\widehat{\mu} : \mathcal{X} \to \mathcal{Y}$ on the training set $\implies$ nonconformity score $R(x, y)$
- Exchangeability of $\mathcal{D}_{\text{calib}} \cup \{(X_{n+1}, Y_{n+1})\}$

$$\left(R(X_{n+1}, Y_{n+1}) \middle| \{R(x_i, y_i)\}_{i \leq n+1} \right) \sim \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R(x_i, y_i)}$$

Calculate the quantile

$$q_{1-\alpha} = \text{Quantile}_{1-\alpha} \left(\frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R(x_i, y_i)} + \frac{1}{n+1} \delta_{\infty} \right)$$

Construct the prediction interval

$$C_{1-\alpha}(X_{n+1}) = \left\{ y : \; R(X_{n+1}, y) \leq q_{1-\alpha} \right\}$$

Two ingredients of conformal prediction

- Exchangeable data $\{(X_i, Y_i)\}_{i \leq n+1}$
- Symmetric algorithm $\mathcal{A}$ (not required in split conformal prediction)

Two ingredients of conformal prediction

- Exchangeable data $\{(X_i, Y_i)\}_{i \le n+1}$
- Symmetric algorithm $\mathcal{A}$ (not required in split conformal prediction)

Question: What if $\{(X_i, Y_i)\}_{i \le n+1}$ are not exchangeable? How can we fix this?

- Weighted exchangeability

Definition (Tibshirani et al., 2019)

Random variables $\{V_i\}_{i \leq n+1}$ are said to be weighted exchangeable with weight functions $\{w_i\}_{i \leq n+1}$ if the joint density can be factorized by

$$f(v_1, \cdots, v_{n+1}) = \left\{ \prod_{i \leq n+1} w_i(v_i) \right\} \cdot g(v_1, \cdots, v_{n+1})$$

where g is any function that does not depend on the ordering of its inputs.

CP UNDER WEIGHTED EXCHANGEABILITY

- If $\{Z_i = (X_i, Y_i)\}_{i \leq n+1}$ are weighted exchangeable with weight functions w_i

$$\left\{ R(Z_{n+1}) \middle| \{R(z_i)\}_{i \leq n+1} \right\} \sim \sum_{i \leq n+1} p_i(Z_1, \cdots, Z_{n+1}) \delta_{R(Z_i)}$$

where p_i's are standardized weights

$$p_i^w(z_1, \cdots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j \leq n+1} w_j\left(z_{\sigma(j)}\right)}{\sum_{\sigma} \prod_{j \leq n+1} w_j\left(z_{\sigma(j)}\right)}, \qquad i = 1, \cdots, n+1$$

- If $\{Z_i = (X_i, Y_i)\}_{i \le n+1}$ are weighted exchangeable with weight functions w_i

$$\left\{ R(Z_{n+1}) \middle| \{R(z_i)\}_{i \le n+1} \right\} \sim \sum_{i \le n+1} p_i(Z_1, \cdots, Z_{n+1}) \delta_{R(Z_i)}$$

where p_i's are standardized weights

$$p_i^w(z_1, \cdots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j \le n+1} w_j(z_{\sigma(j)})}{\sum_{\sigma} \prod_{j \le n+1} w_j(z_{\sigma(j)})}, \qquad i = 1, \cdots, n+1$$

- Construct the prediction interval

$$\widehat{C}_{1-\alpha}(X_{n+1}) = \{y \in \mathcal{Y} : \ R(X_{n+1}, y) \le q_{1-\alpha}^w\}$$

with the threshold

$$q_{1-\alpha}^w = \text{Quantile}_{1-\alpha}\left(\sum_{i \le n} p_i^w \delta_{R(Z_i)} + p_{n+1}^w \delta_{\infty} \right)$$

Barbu, A., D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz (2019). Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems 32*.

Ben-Tal, A. and A. Nemirovski (1998). Robust convex optimization. *Mathematics of operations research 23*(4), 769–805.

Blanchet, J., Y. Kang, and K. Murthy (2019). Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability 56*(3), 830–857.

Chatterjee, S. and J. Lafferty (2019). Adaptive risk bounds in unimodal regression. *Bernoulli*.

Cortez, P., A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems 47*(4), 547–553.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee.

Duchi, J. and H. Namkoong (2019). Variance-based regularization with convex objectives. *Journal of Machine Learning Research 20*(68), 1–55.

El Ghaoui, L. and H. Lebret (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications 18*(4), 1035–1064.

Han, Q., T. Wang, S. Chatterjee, and R. J. Samworth (2019). Isotonic regression in general dimensions. *The Annals of Statistics*.

Lam, H. (2016). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research 41*(4), 1248–1275.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika 74*(1), 13–26.

Sugiyama, M. (2011). Learning under non-stationarity: Covariate shift adaptation by importance weighting. In *Handbook of Computational Statistics: Concepts and Methods*, pp. 927–952. Springer.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*(476), 1619–1637.

Thams, N., S. Saengkyongam, N. Pfister, and J. Peters (2023). Statistical testing under distributional shifts. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(3), 597–663.

Tibshirani, R. J., R. F. Barber, E. J. Candès, and A. Ramdas (2019). Conformal prediction under covariate shift. In *NeurIPS*.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world*, Volume 29. Springer.