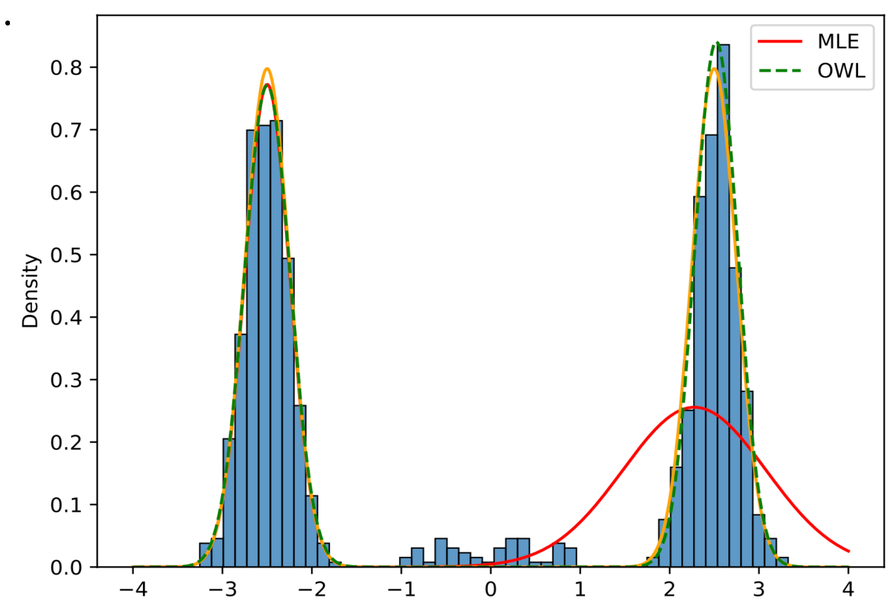
Robustly fit models by using the most model friendly re-weighting within a neighborhood of the observed data.

PRESENTER:
Miheer Dewaskar
Duke UNIVERSITY

Summary: Maximum Likelihood Estimation (MLE) can be brittle to even minor model misspecification like data contamination and outliers. Under contamination, we propose to optimistically reweight the data points to match the closest model in the total-variation distance. This results in down-weighting of the corrupted data points, resulting in a more robust parameter estimate.

Gaussian mixture model example: Only 5% of the samples are corrupted U[-1,1], but MLE sacrifices model fit on majority of the data.



Brittleness of MLE is not limited to mixture models.

Optimistically Weighted Likelihoods (OWL)

Given observations $x_1, \dots, x_n$, a family of model densities $\{p_\theta : \theta \in \Theta\}$, and corruption fraction $0 \leq \epsilon \leq 1$, we find the robust parameter estimate that solves

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(x_i)^{n w_i(\hat{\theta})}$$

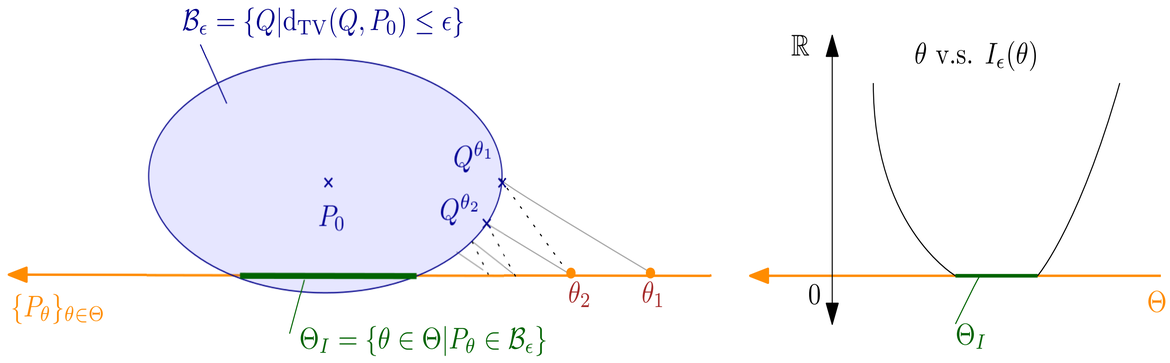
where data weights $\{w_i(\theta)\}_{i=1}^n$ solve the following optimization on the n-dimensional probability simplex Δ_n with a total-variation constraint:

$$\hat{I}_\epsilon(\theta) = \min_{\substack{w \in \Delta_n \\ \frac{1}{2}\|w-o\|_1 \leq \epsilon}} \sum_{i=1}^{n} w_i \log \frac{n w_i \hat{p}(x_i)}{p_\theta(x_i)}$$

where $o = \left(\frac{1}{n}, \frac{1}{n} \dots, \frac{1}{n}\right) \in \Delta_n$ and $\hat{p}$ is a density estimator of the data.

- We implement OWL for various models with product likelihoods: Linear/Logistic Regression and Bernoulli/Gaussian Mixture models.
- Theory support via connection to OKL (below) and Coarsened Inference (Miller & Dunson, 2018).
- Automatic procedure to tune the corruption fraction ϵ.

Theory: Robustly identified parameters Θ_I minimize the OKL function



Optimistic-Kullback-Leibler (OKL) function

$$I_\epsilon(\theta) = \inf_{Q : \mathrm{d_{TV}}(Q, P_0) \leq \epsilon} \mathrm{KL}(Q | P_\theta)$$

OKL Estimation Consistency
Let data $x_1, \dots, x_n \in X$ be i.i.d. with unknown distribution P_0, and fix $\theta \in \Theta$ such that $I_\epsilon(\theta) < \infty$.

Theorem (Finite Spaces): Suppose X is finite and $\mathrm{supp}(P_\theta) \subseteq \mathrm{supp}(P_0)$. Then the bound:

$$\left| I_\epsilon(\theta) - \hat{I}_\epsilon(\theta) \right| = \tilde{O}_p(n^{-1/2})$$

holds with high probability for large values of n.

Theorem (Euclidean Spaces): Suppose $X \subseteq R^d$ is bounded and has finite surface area, and densities p_0 and p_θ are suitably smooth and supported on X. Then for a kernel-based variant $\hat{I}_{h,\epsilon}(\theta)$ with bandwidth h > 0, the bound

$$\left| I_\epsilon(\theta) - \hat{I}_{h,\epsilon}(\theta) \right| = \tilde{O}_p(n^{-1/2} h^{-d} + \sqrt{h})$$

holds with high probability for large values of n.

OKL and Coarsened Inference
To formally allow misspecification in inference, Miller & Dunson (2018) introduce the coarsened likelihood:

$$L_\epsilon(\theta | x_{1:n}) \doteq \mathbb{P}_\theta \left(\boldsymbol{d}(\hat{P}_{Z_{1:n}}, \hat{P}_{x_{1:n}}) \leq \epsilon \right)$$

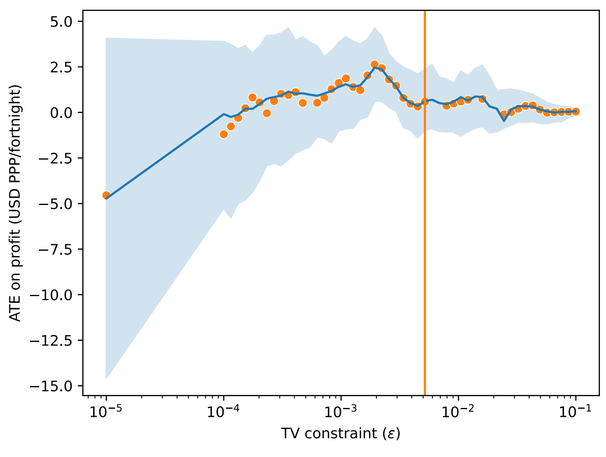
defined as the probability that the empirical measure $\hat{P}_{Z_{1:n}}$ based on hypothetical i.i.d. samples $Z_1, \dots, Z_n$ from P_θ lie in the $(\boldsymbol{d}, \epsilon)$ neighborhood of the empirical measure $\hat{P}_{x_{1:n}}$ from the observed data. Here $\boldsymbol{d}$ is a discrepancy on probability measures $\mathcal{P}(X)$.

Theorem: If data $x_1, \dots, x_n$ are sampled i.i.d. from P_0 and $\boldsymbol{d}$ is a convex distance on $\mathcal{P}(X)$ that is continuous w.r.t. weak-convergence, then as $n \to \infty$:

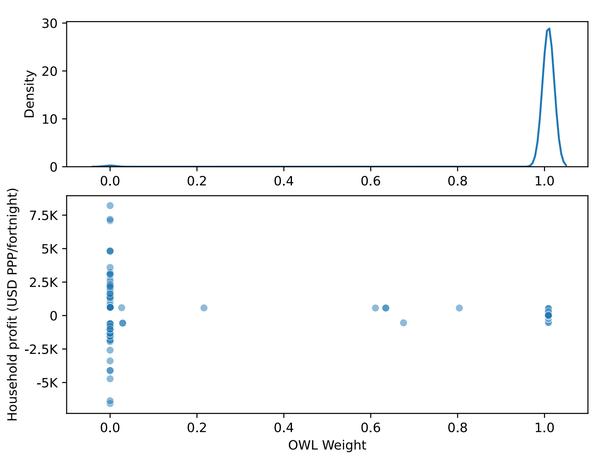
$$-\frac{1}{n} \log L_\epsilon(\theta | x_{1:n}) \xrightarrow{P} \inf_{Q : \boldsymbol{d}(Q, P_0) \leq \epsilon} \mathrm{KL}(Q | P_\theta).$$

Examples of $\boldsymbol{d}$: Wasserstein, MMD & smoothed TV.

Estimated ATE from OWL



Estimated weights from OWL (ϵ_0=0.005)



Application: Micro-credit study

To study the impact of micro-credit, Angelucci et al. (2015) worked with one of the largest micro-lenders in Mexico to randomize their credit rollout across 238 geographical regions in the Sonora state. Within 18-34 months after this rollout, the authors surveyed *n=16,560* households for various outcome measures.

Model for Average Treatment Effect (ATE) on household profits

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \qquad i = 1, \dots, n$$

Here Y_i is the profit for household i in the last two weeks, and $T_i \in \{0,1\}$ is the treatment status of household i. We are interested in inferring the ATE: i.e., the coefficient β_1.

Broderick, Giordano and Meager (2023) showed that the OLS estimate of the average treatment effect (ATE) β_1 is *brittle* to the removal of a handful of households. On the right, we robustly infer ATE by fitting the model using OWL for 50 values of ϵ placed uniformly $\log_{10}$ from -4 to -1.

OWL tuning procedure selected $\epsilon_0 = 0.005$, and down-weighted 1% of the households with extreme profit values. Estimated ATE at ϵ_0 of $\beta_1 = 0.6$ USD PPP/2 weeks is stable with respect to ϵ and has relatively narrow OS-bootstrap confidence bands (left).

Download full paper here:

Robustifying Likelihoods by Optimistically Re-weighting data

Miheer Dewaskar, Christopher Tosh,
Jeremias Knoblauch, David Dunson