Optimistic Distributionally Robust Optimization for Nonparametric Likelihood Approximation

Viet Anh Nguyen[†], Soroosh Shafieezadeh-Abadeh[†], Man-Chung Yue[‡], Daniel Kuhn[†] and Wolfram Wiesemann[§]

[†]EPFL, [‡]The Hong Kong Polytechnic University, [§]Imperial College London

Contributions

Evaluating the likelihood of an observation is a fundamental question in various fields

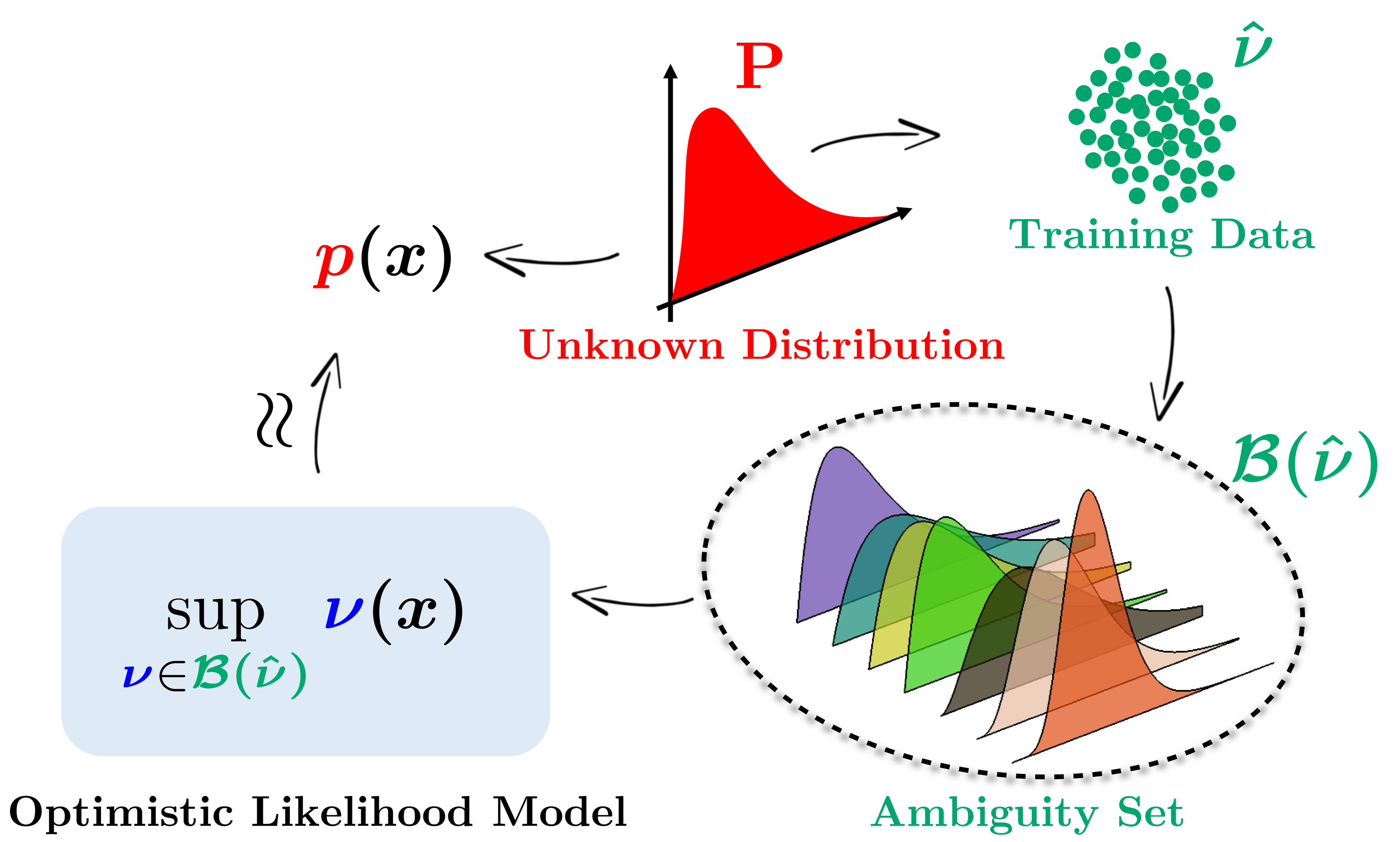
Motivation:

- Data-generating distribution $\mathbb{P}$ is unknown
- Estimation of parameters from training samples leads to overfitting
- Ignoring this uncertainty leads to poor out-of-sample performance

Contributions:

- Nonparametric estimation using **Optimistic Likelihood** (OL) approach
- KL ambiguity sets $\Rightarrow$ a finite convex problem
- Mean-variance ambiguity sets $\Rightarrow$ a semidefinite program with analytical solution
- Wasserstein ambiguity sets $\Rightarrow$ a linear program solved with sorting

Generic Optimistic Likelihood Model



Kullback-Leibler Divergence

For distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ with density functions $p_0(x)$ and $p_1(x)$, we have

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) = \int_{-\infty}^{\infty} p_0(x) \log\left(\frac{p_0(x)}{p_1(x)}\right) \mathrm{d}x$$

The KL ambiguity set with radius $\varepsilon > 0$:

$$\mathcal{B}(\hat{\nu}) = \{\nu \in \mathcal{M}(\mathcal{X}) : \mathrm{KL}(\hat{\nu} \parallel \nu) \le \varepsilon\}$$

OL solution:

- $x \in \{\hat{x}_1, \ldots, \hat{x}_N\}$: OL reduces to the finite convex program

$$\sup_{\nu \in \mathcal{B}_{\mathrm{KL}}(\hat{\nu})} \nu(x) = \max\left\{\sum y_i \mathbb{1}_x(\hat{x}_i) : y \in \Delta, \sum \hat{\nu}_i \log(\hat{\nu}_i / y_i) \le \varepsilon\right\},$$

- $x \notin \{\hat{x}_1, \ldots, \hat{x}_N\}$: OL is independent of x

$$\sup_{\nu \in \mathcal{B}_{\mathrm{KL}}(\hat{\nu})} \nu(x) = 1 - \exp(-\varepsilon)$$

Mean-Variance Ambiguity Set

The mean-variance ambiguity set defined as

$$\mathcal{B}_{\mathrm{MV}}(\hat{\nu}) = \left\{\nu \in \mathcal{M}(\mathcal{X}) : \mathbb{E}_\nu[\tilde{x}] = \hat{\mu},\ \mathbb{E}_\nu[\tilde{x}\tilde{x}^\top] = \hat{\Sigma} + \hat{\mu}\hat{\mu}^\top\right\},$$

where $\hat{\mu} = \frac{1}{N}\sum \hat{x}_i$ and $\hat{\Sigma} = \frac{1}{N}\sum (\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})^\top$

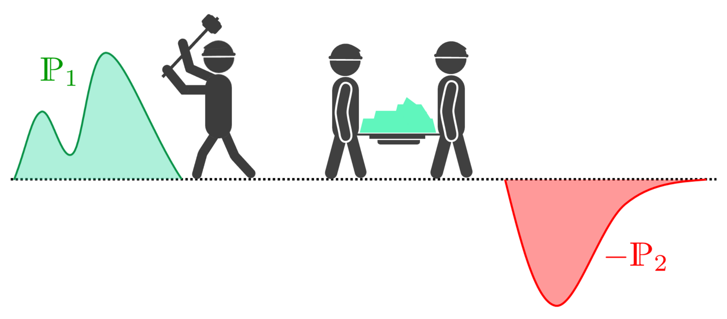
OL solution:

$$\sup_{\nu \in \mathcal{B}_{\mathrm{MV}}(\hat{\nu})} \nu(x) = \frac{1}{1 + (x - \hat{\mu})^\top \hat{\Sigma}^{-1} (x - \hat{\mu})} \in (0, 1].$$

Wasserstein Ambiguity Set

Wasserstein distance:

$$\mathrm{W}(\mathbb{P}_1, \mathbb{P}_2) \triangleq \inf_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}_\pi\left[d(\xi_1, \xi_2)\right]$$



Wasserstein ambiguity set:

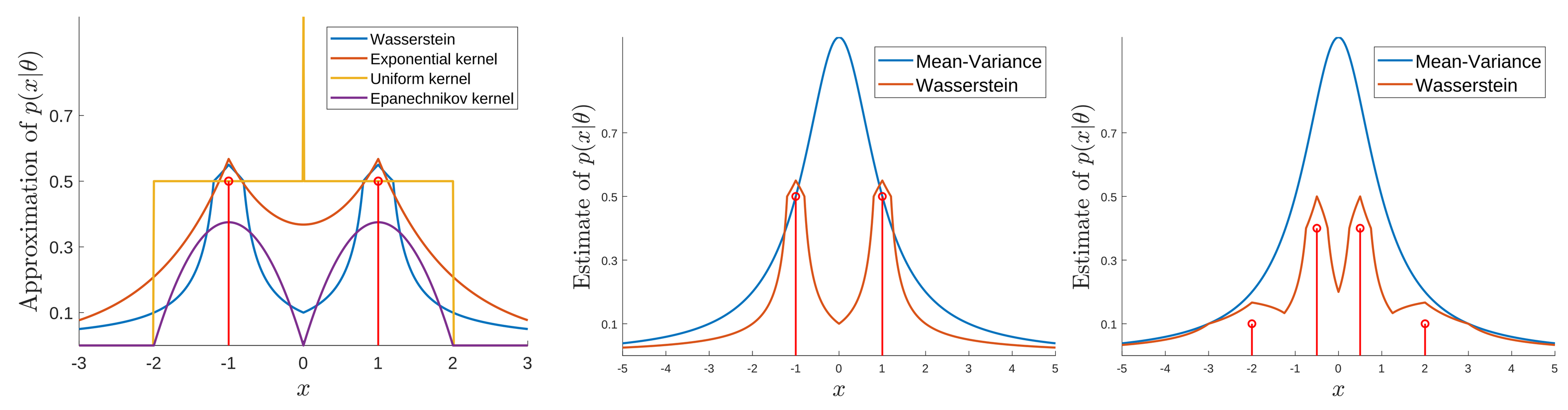
$$\mathcal{B}_{\mathrm{W}}(\hat{\nu}) = \{\nu \in \mathcal{M}(\mathcal{X}) : \mathrm{W}(\nu, \hat{\nu}) \le \varepsilon\}$$

OL Solution:

- Reduce to the linear program

$$\sup_{\nu \in \mathcal{B}_{\mathrm{W}}(\hat{\nu})} \nu(x) = \max\left\{\sum y_i : y_i \in [0, 1/N], \sum d(x, \hat{x}_i)\, y_i \le \varepsilon, \right\}$$

- Can be solved in $\mathcal{O}(N \log N)$ using greedy heuristics.



Wasserstein OL vs Kernel

Wasserstein OL vs Mean-Variance OL

ELBO Problem

Assumption: Finite parameter space $\Theta = \{\theta_1, \ldots, \theta_C\}$

- **Evidence Lower BOund**:

$$\mathcal{J}^{\mathrm{true}} \triangleq \min_{\mathbb{Q} \in \mathcal{Q}} \mathrm{KL}(\mathbb{Q} \parallel \pi) - \mathbb{E}_{\mathbb{Q}}[\log p(x|\theta)]$$
$$= \min_{q \in \mathcal{Q}} \sum q_i (\log q_i - \log \pi_i) - \sum q_i \log p(x|\theta_i)$$

- **Unknown** $p(x|\theta_i)$
- **ELBO with OL**: Given observations $\hat{x}_1, \ldots, \hat{x}_{N_i} \sim p(\cdot|\theta_i)$ for all $i \in [C]$

$$\hat{\mathcal{J}}_{\mathcal{B}} = \min_{q \in \mathcal{Q}} \sum q_i (\log q_i - \log \pi_i) - \sum q_i \log\left(\sup_{\nu_i \in \mathcal{B}(\hat{\nu}_i)} \nu_i(x)\right)$$

Question: is $\hat{\mathcal{J}}_{\mathcal{B}}$ **close to** $\mathcal{J}^{\mathrm{true}}$?

Performance Guarantees

KL Ambiguity Set – Asymptotic Guarantee: set $n = \min\{N_1, \ldots, N_C\}$

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}^\infty(\mathcal{J}^{\mathrm{true}} < \hat{\mathcal{J}}_{\mathcal{B}}) \le -\min \varepsilon_i < 0.$$

Wasserstein Ambiguity Set – Finite sample guarantee: If $\varepsilon_i = \varepsilon_i(\beta, C, N_i)$ defined as

$$\varepsilon_i(\beta, C, N_i) \triangleq \begin{cases} \left(\frac{\log(k_{i1} C \beta^{-1})}{k_{i2} N_i}\right)^{1/\max\{m,2\}} & \text{if } N_i \ge \frac{\log(k_{i1}) C \beta^{-1}}{k_{i2}}, \\ \left(\frac{\log(k_{i1} C \beta^{-1})}{k_{i2} N_i}\right)^{1/a_i} & \text{if } N_i < \frac{\log(k_{i1}) C \beta^{-1}}{k_{i2}}, \end{cases}$$

then $\mathbb{P}^N\left(\mathcal{J}^{\mathrm{true}} < \hat{\mathcal{J}}_{\mathcal{B}}\right) \le \beta$.

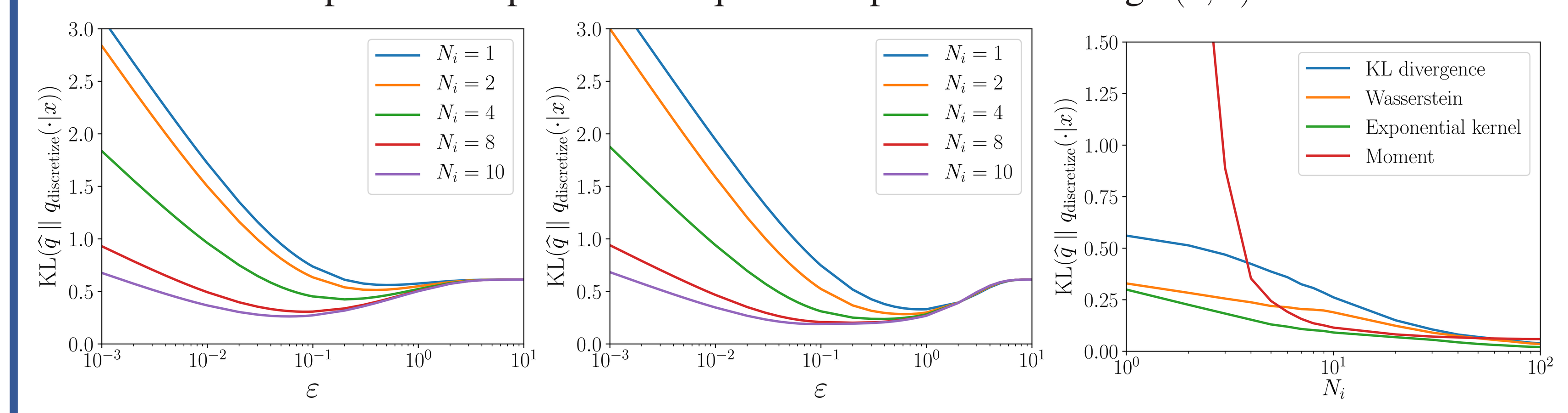
Wasserstein Ambiguity Set – Asymptotic Guarantee:

If $\varepsilon_i = \varepsilon_i(\beta, C, N_i)$, then $\hat{\mathcal{J}}_{\mathcal{B}} \to \mathcal{J}^{\mathrm{true}}$ as $N_1, \ldots, N_C \to \infty$ almost surely.

Simulation Results

Beta-Binomial Inference:

- True (unknown) likelihood: $p(x|\theta) = \mathrm{Bin}(x|M, \theta)$
- Prior: $\pi(\theta) = \mathrm{Beta}(\theta|\alpha, \beta)$
- True posterior: $q(\theta|x) = \mathrm{Beta}(\theta|x + \alpha, M - x + \beta)$
- Discretized parameter space Θ: 20 equidistant points in the range $(0, 1)$

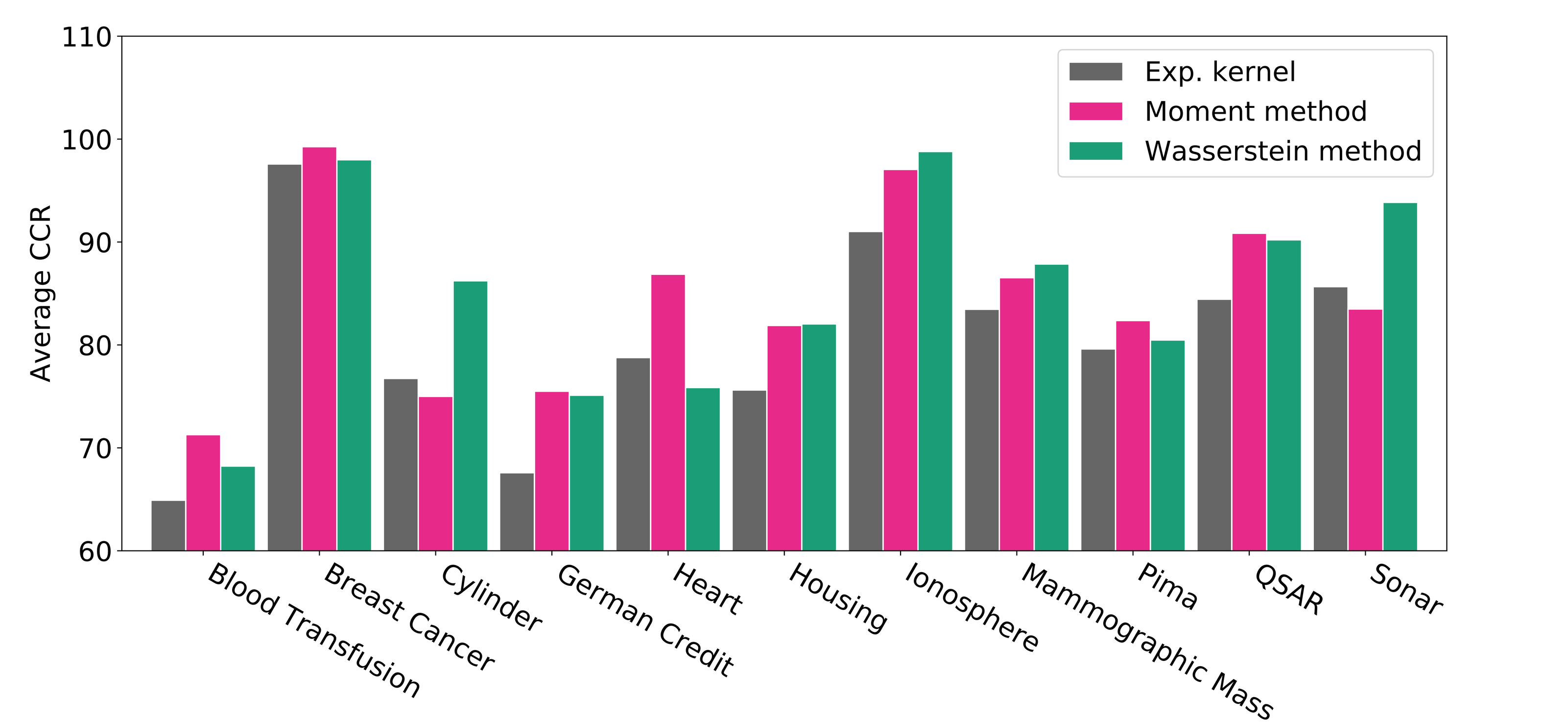


KL ambiguity set

Wasserstein ambiguity set

Comparison among different methods

Empirical Experiments (UCI dataset): average area under the precision-recall curve



References

- Nguyen, V.A., Shafieezadeh-Abadeh, S., Yue, M.C., Kuhn, D. and Wiesemann, W. (2019). Calculating optimistic likelihoods using (geodesically) convex optimization. *Advances in Neural Information Processing Systems.*