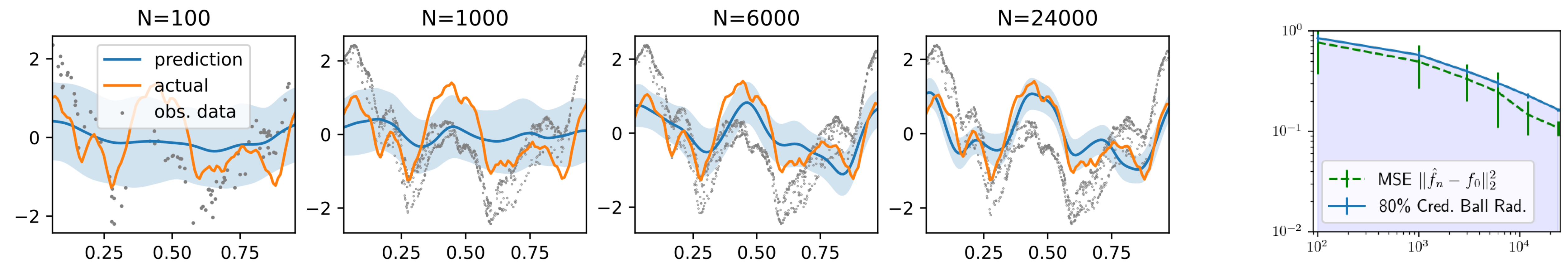
# Quasi-Bayesian Dual Instrumental Variable Regression

Ziyu Wang* (Tsinghua), Yuhao Zhou* (Tsinghua), Tongzheng Ren (UT Austin), Jun Zhu (Tsinghua)

wzy196@gmail.com
https://arxiv.org/abs/2106.08750v2
https://github.com/meta-inf/qbdiv

ML ECON 2021

- Quasi-Bayesian uncertainty quantification for kernelized IV models, without requiring the knowledge of the full data generating process
- Frequentist guarantees: minimax rates of posterior contraction, analysis of credible balls
- Scalable approximate inference with a modified "randomized prior trick"
- Heuristic application to wide neural network models




Left: visualizations of the mean estimator $\hat{f}_n$ and quasi-posterior. Right: $MSE(\hat{f}_n, f_0)$ vs $L_2$ credible ball radius.

## Background

**IV regression**: estimate causal effect $f : X \to Y$ on confounded data, through the use of instruments $z$. Conditional moment restriction formulation:

$$\mathbf{E}(f(\mathbf{x}) - \mathbf{y} \mid \mathbf{z}) = 0 \text{ a.s.} \quad \text{(CMR)}$$

Estimation requires

- An estimator $\hat{E}$ for (the restriction on $H$ of) $E : f \mapsto E(f(\mathbf{x}) \mid \mathbf{z} = \cdot)$
- A choice of $\|\cdot\|$ to weight violation of $(CMR)$

**Dual/minimax formulation**: two-stage estimation ⇒ minimax optimization

$$\hat{f}_n = min_{f\in H} \underbrace{max_{g\in I} \sum_{i=1}^n \left(2(f(x_i) - y_i - g(z_i))g(z_i) - g^2(z_i)\right) - \nu\|g\|_I^2}_{=n\|\hat{E}f - \hat{E}(y|z)\|^2 \text{ for some choices of } \hat{E} \text{ and } \|\cdot\|} + \lambda\|f\|_H^2 \quad (1)$$

**Kernelized IV**:[1] use RKHS for $H$. Compared with classical models, improves adaptivity to the smoothness of the data distribution.

## Quasi-Bayesian IV

Bayesian IV requires knowledge of the full data generating process. Not in (CMR) Assume a (conditional?) generative model for $p(x \mid z)$:

- Additional risk of misspecification
- Need *Bayesian inference* over the (parameter of) generative model, computation extremely expensive

⇒ Quasi-Bayes (Chernozhukov and Hong, 2003): use the Gibbs distribution

$$p_\lambda(df) \propto \pi(df) exp\left(-\lambda^{-1} n\|\hat{E}f - \hat{E}(y|z)\|^2\right)$$

to quantify uncertainty. Trades off between evidence and prior belief:

$$p_\lambda = argmin_\rho \mathbf{E}_{\rho(df)} n\|\hat{E}f - \hat{E}(y|z)\|^2 + \lambda KL(\rho\|\pi).$$

Estimation error in $\hat{E}$ complicates computation and frequentist justification

## Prior Work

- A vast literature on truncated series / smoothing-based approaches
- Kato (2013) established optimal contraction rates for a series-based quasi-posterior, under the requirement that both stages use the same number of bases

Past works on various kernelized IV/CMR estimators mostly provide rates for $\|E(\hat{f}_n - f_0)\|_2$. This can be compared with the OLS lower rate on $\{(y_i, z_i)\}$.[2]

- Singh et al (2019) matches the optimal OLS rate (w.r.t. $N_{\text{stage 2}}$) under source conditions, but requires $N_{\text{stage 1}} \gg N_{\text{stage 2}}$
- When $max\{logN(H_1, \|\cdot\|_\infty, \epsilon), logN(I_1, \|\cdot\|_\infty, \epsilon)\} \lesssim \epsilon^{-\frac{2}{b}}$, Dikkala et al (2020) establishes $\|E(\hat{f}_n - f_0)\|_2 = O(n^{-\frac{b/2}{b+1}})$ for the estimator (1)
- Mastouri et al (2021) provides a rate for $\|\hat{f}_n - f_0\|_H^2$ under source conditions, which is at best $O(n^{-1/4})$

## Kernelized Quasi-Bayesian Dual IV

Assume $GP(0, k_x)$ prior for $f$. Plug in the choice of $\|\hat{E}f - \hat{E}(y \mid z)\|$ from (1). Closed-form quasi-posterior:

$$\Pi(f(x_*) \mid D^{(n)}) = N(K_{*x}(\lambda + LK_{xx})^{-1}LY,\ K_{**} - K_{*x}L(\lambda I + K_{xx}L)^{-1}K_{x*})$$
$$\text{where } L = K_{zz}(K_{zz} + \nu I)^{-1}$$

**Proposition.** For functionals $L \in H^*$, $\Pi(Lf \mid D^{(n)}) = \mathbf{E}_{Y|X}(L(\hat{f}_n) - L(f_0))^2$ on unconfounded data, for (i) the worst-case $f_0 \in H$, (ii) the average-case $f \sim GP(0, k_x)$.

- Still explains a non-trivial proportion of error

## Frequentist Theory

Assuming:

- Mildly ill-posed problem: $\lambda_i(E^\top E) \asymp i^{-2p}$;
- Link condition between $E$ and Mercer bases of $k_x$;
- $f_0$ in a certain $L_2$-Sobolev space $\bar{H}$, s.t. $logN(\bar{H}_1, \|\cdot\|_2, \epsilon) \lesssim \epsilon^{-2/b}$;[3]
- $H, I$ are correctly specified Matérn RKHSs.
  - Paper states more general conditions allowing for "almost all" bounded $\bar{H}$'s satisfying $L_2$ entropy bounds, and suitable $I$ incl. Nyström approximated

Then we have, in $L_2$ and Sobolev norms,

❶ Posterior **contracts** at asymptotically **minimax optimal rates**:

$$\mathbf{P}_{D^{(n)}} \Pi\left(\|f - f_0\|^2_{[L_2(P(dx)), H]_{\alpha,2}} > Mn^{-\frac{(1-\alpha)b}{b+2p+1}} \,\middle|\, D^{(n)}\right) \to 0, \quad \forall \alpha \in \left[0, \frac{b}{b+1}\right)$$

$$\Rightarrow \|\hat{f}_n - f_0\|_2 \lesssim n^{-\frac{b/2}{b+2p+1}}, \|\hat{f}_n - f_0\|_\infty \lesssim n^{-\frac{(b-1)/2}{b+2p+1}}, \|E(\hat{f}_n - f_0)\|_2 \lesssim n^{-\frac{(b+2p)/2}{b+2p+1}} \ll n^{-\frac{b/2}{b+1}}.$$

❷ Radii of credible balls have the **correct order of magnitude**.

## Inference & Heuristic Application to NNs

Consider wide NNs in the kernel regime. Then network weights are (over-parameterized) random features.

"Randomized prior" trick: *Sampling* from the quasi-posterior ⇔ *Optimization* of the perturbed MAP objective

$$min_f max_g \sum_{i=1}^n \left(2(f(x_i) - y_i - \tilde{e}_i - g(z_i))g(z_i) - g^2(z_i)\right) - \nu\|g - \tilde{g}_0\|_I^2 + \lambda\|f - \tilde{f}_0\|_H^2,$$

where $\tilde{e}_i \sim N(0, \lambda), \tilde{f}_0 \sim GP(0, k_x), \tilde{g}_0 \sim GP(0, \lambda\nu^{-1}k_z)$.

For random feature, RKHS norm = $L^2$ feature norm.

Correction for NTK/NNGP discrepancy follows the development in OLS/GPR (He et al, 2020)

## Numerical Experiments

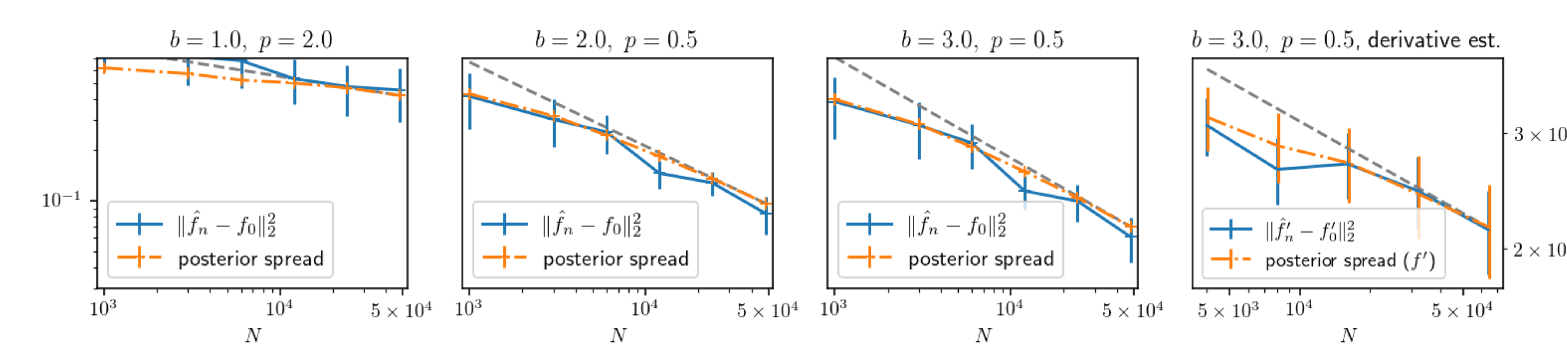
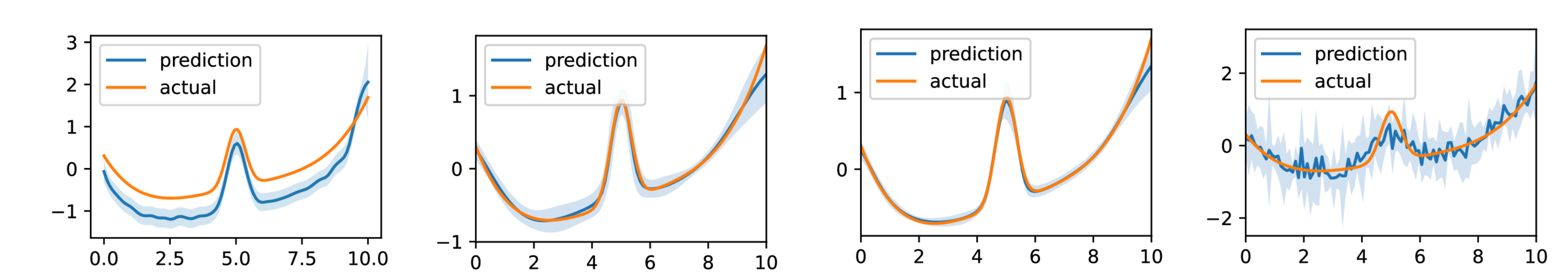



**Figure:** Validation of asymptotic results, using Matérn/Sobolev kernels on $\mathbb{T}^1$. All assumptions hold with known constants, and $f_0 \sim GP$




(a) *BayesIV*, $N = 1k$ (b) QB, $N = 1k$ (c) QB, $N = 10k$ (d) QB, high-d

**Figure:** Results on the *airline demand* dataset. (a-c) corresponds to the low-dim variant of the dataset. QB uses NN models.

(See paper for additional experiments and full results.)

[1]: Zhang et al (2020); Mastouri et al (2021) studies alternative use of RKHS. [2]: See paper for a complete review. [3]: $\bar{H} \supsetneq H$ as is standard practice in GPR. This is due to the mismatch between prior and RKHS regularity; see van der Vaart and van Zanten (2011)