Optimizer's Information Criterion: Dissecting and Removing Bias in Data-Driven Optimization

Garud Iyengar, Henry Lam, Tianyu Wang

Contact: tw2837@columbia.edu | Columbia University

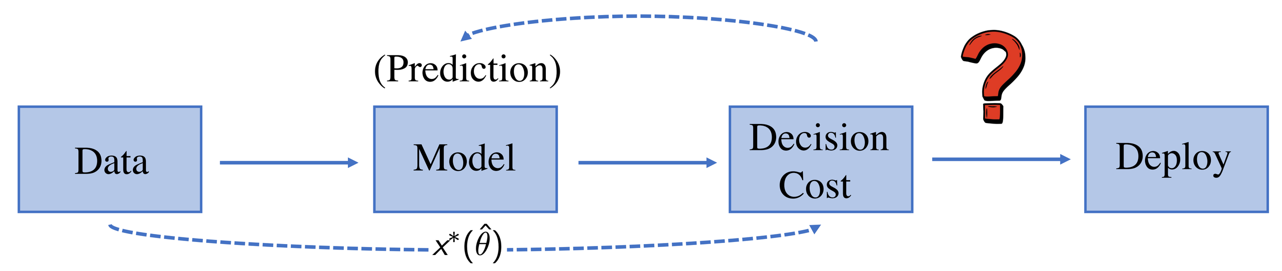
Data-Driven Optimization

$$\min_{x\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}^*}[h(x;\xi)]\right\}\quad \text{(Basic Framework)}$$

- Unknown: distribution $\xi\sim\mathbb{P}^*$;
- We only observe samples $\mathcal{D}_n := \{\xi_i\}_{i=1}^n \sim (\mathbb{P}^*)^n$;
- Generalization: Contextual Optimization, Risk-aversed Optimization.

General optimization procedures:

- Output $x^*(\hat\theta)\in\{x^*(\theta)\in\mathcal{X}|\theta\in\Theta\}$, where $\hat\theta$ is optimized from $\mathcal{D}_n$.



- Variants: Model class + Opt design (prediction-based, robust, ...)

Which optimization procedure should we choose?

Criterion to evaluate and compare decisions:

$$A := \mathbb{E}_{\mathcal{D}_n}\left(\mathbb{E}_{\mathbb{P}^*}[h(x^*(\hat\theta);\xi)]\right)\quad \text{(Example: Expected Cost)}$$

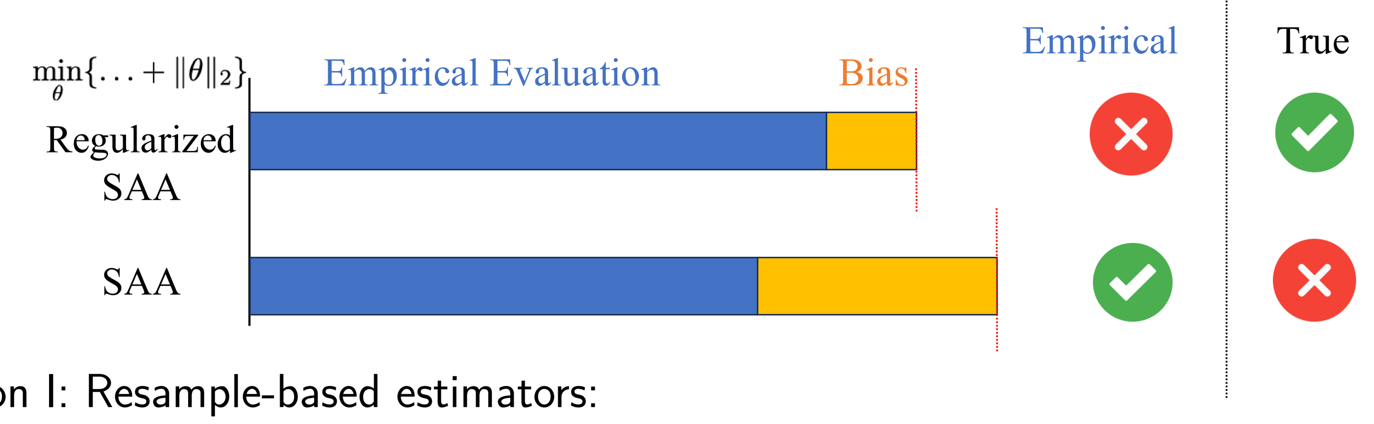
Goal: Find an estimator $\hat A(x^*(\hat\theta))$ for every $x^*(\hat\theta)$.

- **Accurate Evaluation**: $\mathbb{E}_{\mathcal{D}_n}[\hat A]\approx A$;
- **Accurate Comparison**: A smaller $\hat A$ implies smaller A.

Evaluation Bias and Existing Solutions

Empirical Evaluation: $\hat A_o := \frac{1}{n}\sum_{i=1}^n h(x^*(\hat\theta);\xi_i)$ with the smallest $\hat A_o$

- **Evaluation bias**: $A-\mathbb{E}[\hat A_o]>0$.
- **Selection bias**: Always select SAA $\hat\theta\in\arg\min_\theta\{\sum_{i=1}^n h(x^*(\theta);\xi_i)\}$;



Solution I: Resample-based estimators:

- K-Fold / Leave-one-out (LOO) Cross-validation, bootstrap;
- Refitting many optimization procedures decisions.

Solution II: Problem-specific estimators:

- Approximate Leave-one-out (ALO), Akaike Information Criterion (AIC);
- Provide bias formulas / computationally efficient ways for specific cost functions.

Our Work: A general and efficient approach that evaluates data-driven decisions by removing the bias.

Optimizer's Information Criterion (OIC)

OIC corrects the bias through direct estimation:

$$\hat A := \frac{1}{n}\sum_{i=1}^n h(x^*(\hat\theta);\xi_i) - \frac{1}{n^2}\sum_{i=1}^n \underbrace{\nabla_\theta h(x^*(\hat\theta);\xi_i)^\top}_{Decision}\cdot\underbrace{\widehat{IF}(\xi_i)}_{Estimation}$$

- OIC yields nearly unbiased performance evaluation: $\mathbb{E}[\hat A]=A+o(1/n)$.
- OIC is close to LOOCV: $n(\hat A_{\text{OIC}}-\hat A_{\text{LOOCV}})\xrightarrow{p}0$. (LOOCV has superior evaluation and selection performance.)

Tools and Applications

Optimization Procedure: $T(\cdot):\mathcal{D}_n\mapsto$ Decision Parameter,
$\hat\theta=T(\hat{\mathbb{P}}_n)$ and $\theta^*=T(\mathbb{P}^*)$.

Influence Function:

$$IF(\xi;T,\mathbb{P}^*)=\lim_{\epsilon\to0^+}\frac{T(\epsilon\delta_\xi+(1-\epsilon)\mathbb{P}^*)-T(\mathbb{P}^*)}{\epsilon}.$$

$IF(\xi)$ captures the impact of a point ξ on the optimization procedure T.
Estimated Influence Function: $\widehat{IF}(\xi)=IF(\xi;\hat T,\hat{\mathbb{P}}_n)$ for some approximated optimization procedure $\hat T(\cdot)$.

Proof Sketch:

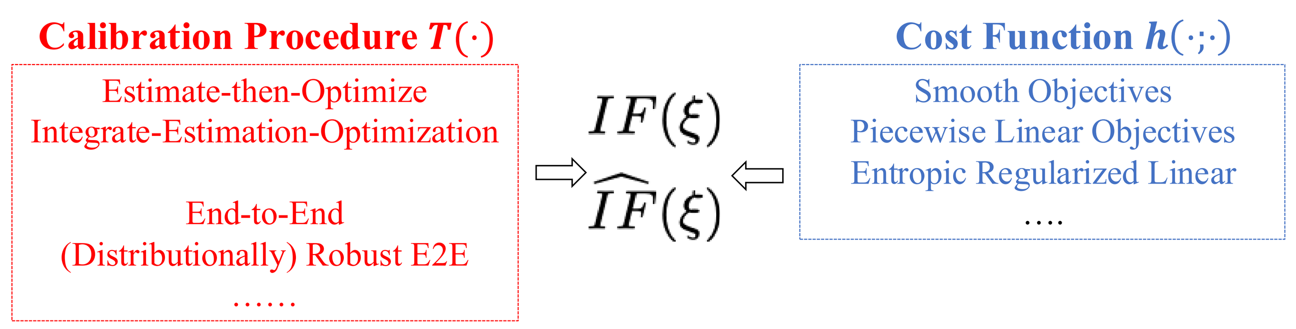
- Deriving the expected bias based on properties of the influence function:

$$A=\mathbb{E}_{\mathcal{D}^n}\mathbb{E}_{\mathbb{P}^*}\left[\hat A_o\right]-\underbrace{\frac{\mathbb{E}_{\mathbb{P}^*}[\nabla_\theta h(x^*(\theta^*);\xi)^\top IF(\xi)]}{n}}_{\text{Bias}}+o\left(\frac{1}{n}\right).$$

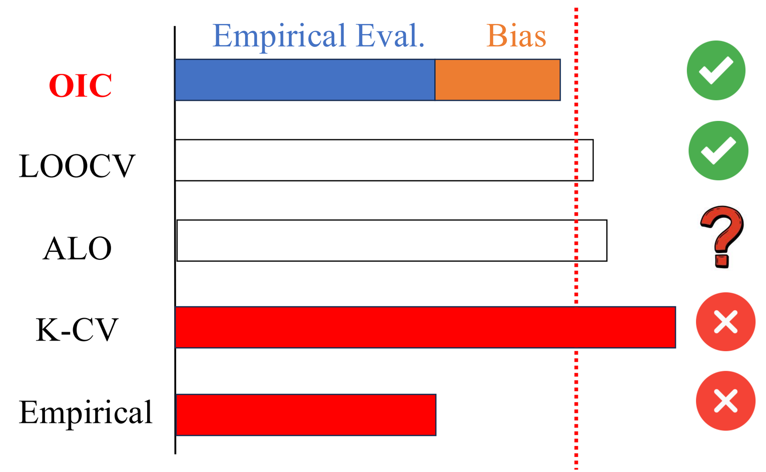
- Approximate unknown $\nabla_\theta h(x^*(\theta^*);\xi)$ and $IF(\xi)$ with empirical counterparts:

$$\mathbb{E}_{\mathcal{D}_n}[\|\widehat{IF}(\xi)\|_2]-\|IF(\xi)\|_2=o(1).$$

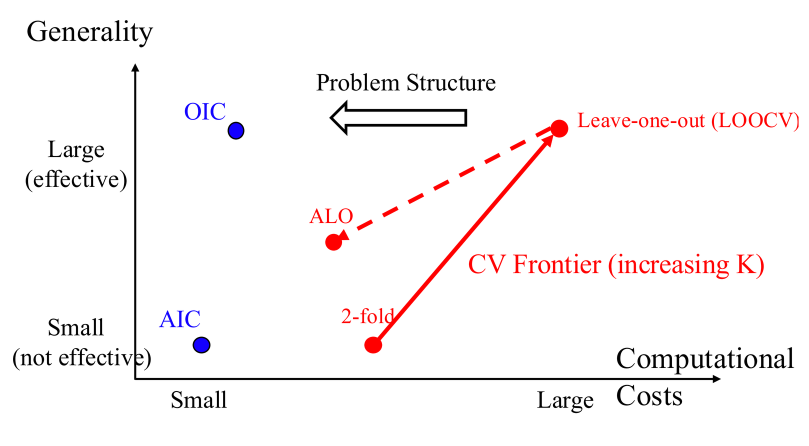
Generality: OIC can be computed for general (constrained) (non)smooth objectives.



Computation: $\widehat{IF}(\xi)$ only requires gradient and hessian of the cost function, which is more efficient than refitting optimization procedures.



(a) Statistical Performance



(b) Statistics Computation Frontier

Optimization Procedure Example [Estimate-then-Optimize]

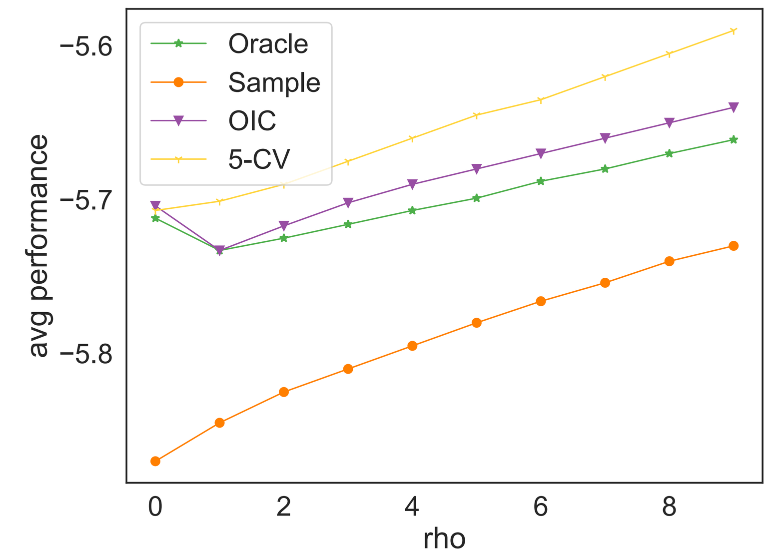
$T(\mathbb{P})\in\arg\min_\theta\mathbb{E}_{\mathbb{P}}[\phi(\theta;\xi)],\quad x^*(\theta)\in\arg\min_{x\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_\theta}[h(x;\xi)].$

$$IF(\xi)=-(\mathbb{E}_{\mathbb{P}^*}[\nabla_{\theta\theta}\phi(\theta^*;\xi)])^{-1}\nabla_\theta\phi(\theta^*;\xi)$$
$$\widehat{IF}(\xi)=-(\mathbb{E}_{\hat{\mathbb{P}}_n}[\nabla_{\theta\theta}\phi(\hat\theta;\xi)])^{-1}\nabla_\theta\phi(\hat\theta;\xi)$$

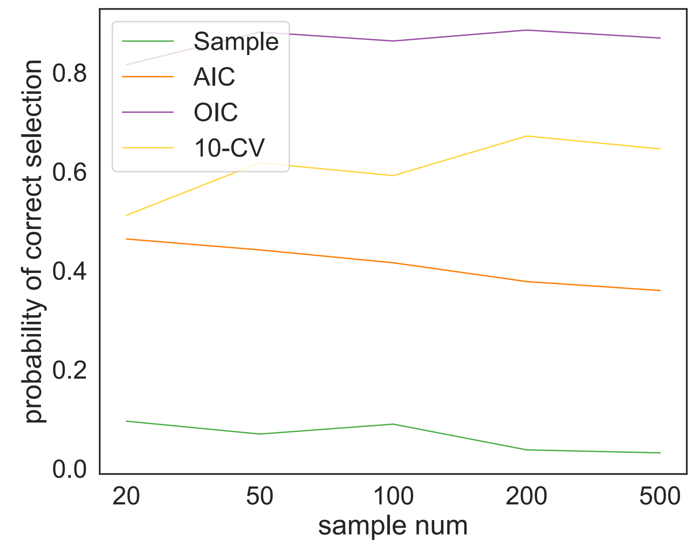
For example, $\phi(\theta;\xi)=-\log p_\theta(\xi)$ in maximum likelihood estimation.

Benefit I: Statistically Improved Decision Selection

OIC identifies optimal hyperparameters and models through improved evaluation.



(a) Parameter Selection: Mean-Variance Portfolio Optimization



(b) Model Selection: Newsvendor Problem (Exponential versus Normal)

- Performance evaluation integrated in `https://python-dro.org`.

Benefit II: Reduction of Computational Costs

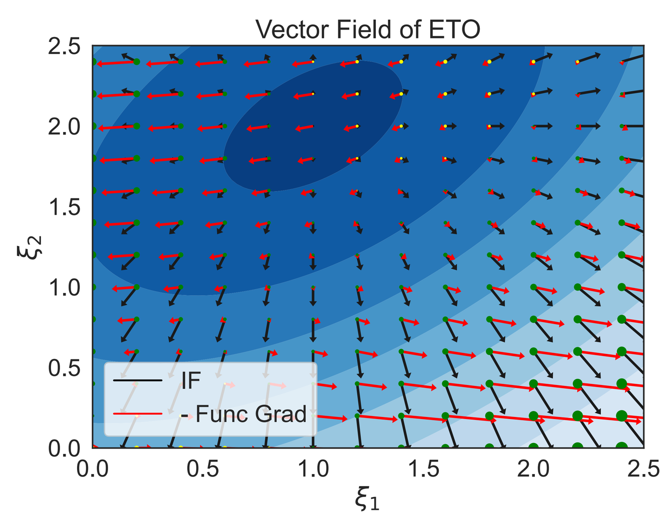
OIC does not need to solve additional optimization problems.

Task	Method	OIC	5-CV	ALO
Portfolio	ERM	$\mathbf{1.64\times10^{-3}}$	3.84×10^{-3}	4.10×10^{-3}
	Param	5.50×10^{-1}	$\mathbf{4.18\times10^{-1}}$	3.23×10^{0}
	DRO	$\mathbf{1.64\times10^{-3}}$	6.59×10^{0}	4.29×10^{-3}
Regression	Ridge	$\mathbf{1.1\times10^{-2}}$	3.4×10^{-2}	8.4×10^{-2}
	Neural Network	$\mathbf{1.8\times10^{2}}$	8.3×10^{2}	1.9×10^{2}-

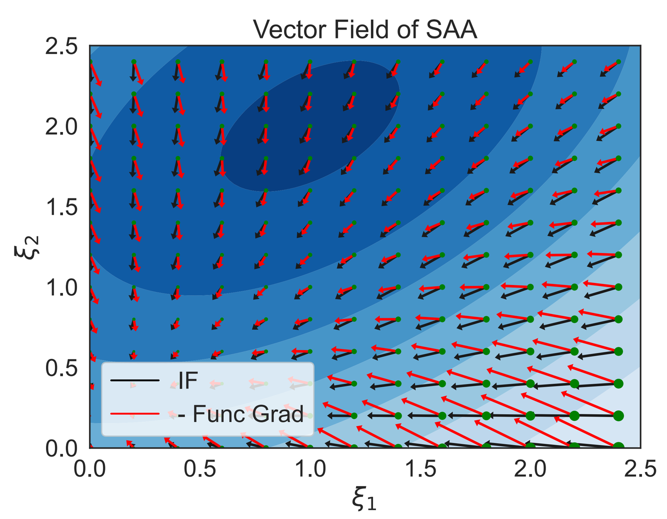
Note. running time for each evaluation procedure (unit: seconds).

Benefit III: Transparent Design Principle

Understanding the evaluation bias leads to a better design of decision-focused learning.



(a) ETO Bias ($A_c=0.82$)



(b) SAA Bias ($A_c=3.79$)

→ Incorporate OIC in the evaluation, selection and training procedure