Kernel Debiased Plug-in Estimation

INFORMS DMDA Workshop, 2023

Brian Cho, Yaroslav Mukhin, Kyra Gan, Ivana Malenica

Cornell University

Outline for This Talk

1. Naive Plug-in Estimation: Why does this fail?
2. Existing Methods for Debiasing: TMLE
3. Our Method: KDPE!

Motivation

Example: ATE Estimation

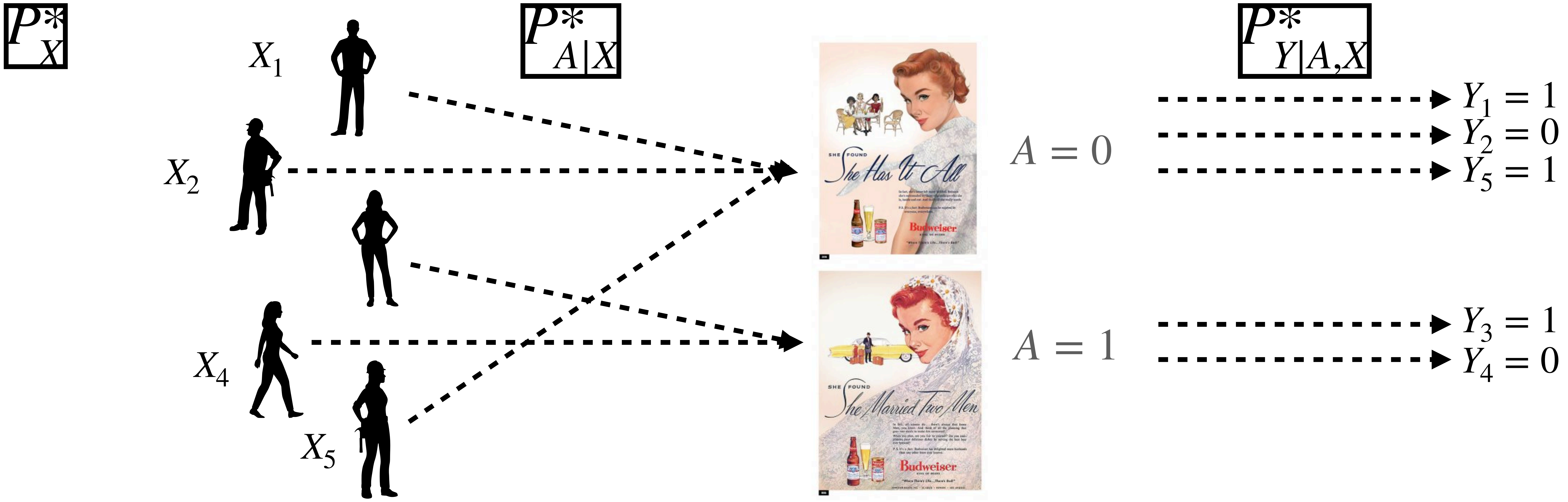
We have a fixed dataset $\{O_i\}_{i=1}^n$.

$$O_i = (X_i, A_i, Y_i) \sim_{i.i.d.} P^*$$

$X =$ (age, beer preference, etc.)

A : Ad Assignment

Y : Did they click?



Motivation

Example: ATE Estimation

We have a fixed dataset $\{O_i\}_{i=1}^n$.

$$O_i = (X_i, A_i, Y_i) \sim_{i.i.d.} P^*$$

Goal: From the data, we want a "**good**" estimator of useful quantities

$$\psi(P^*) = \mathbb{E}_{P^*}[\mathbb{E}_{P^*}[Y|A=1,X]] - \mathbb{E}_{P^*}[\mathbb{E}_{P^*}[Y|A=0,X]]$$

"Good Estimator"?

A. Enables Uncertainty Quantification: tractable limiting distribution via. CLT.

B. Data-efficient and consistent: converges to truth faster with less data

C. Retains simplicity of a plug-in approach

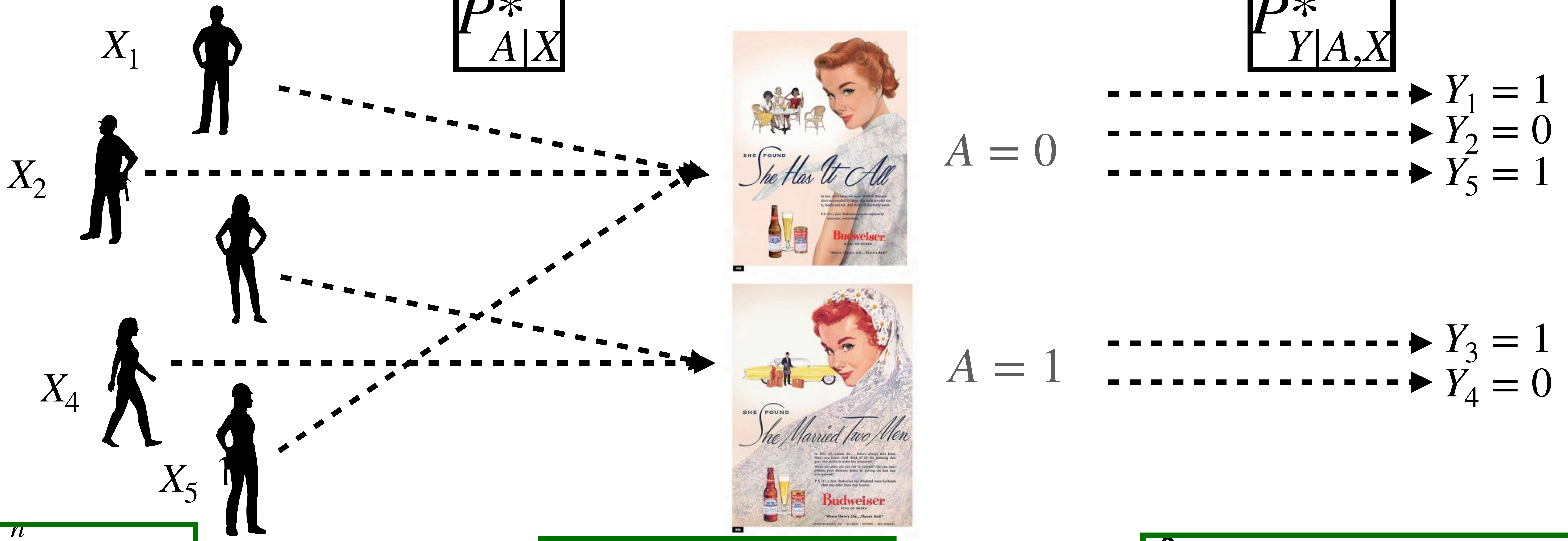
Problem : $P^* = (P^*_X, P^*_{A|X}, P^*_{Y|A,X})$ unknown!

- All we assume is that $P^* \in M$.
- M nonparametric - i.e. unwilling to make strong assumptions about the unknown P^*

What is naive plug-in estimation?

Estimate unknown components of distribution!

P^*_X $\quad$ $P^*_{A|X}$ $\quad$ $P^*_{Y|A,X}$



$$\hat{P}_X(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}[X_i = x]$$

$$\hat{P}_{A=1|X} = g(X)$$

$$\hat{P}_{Y=1|A,X} = f(A, X)$$

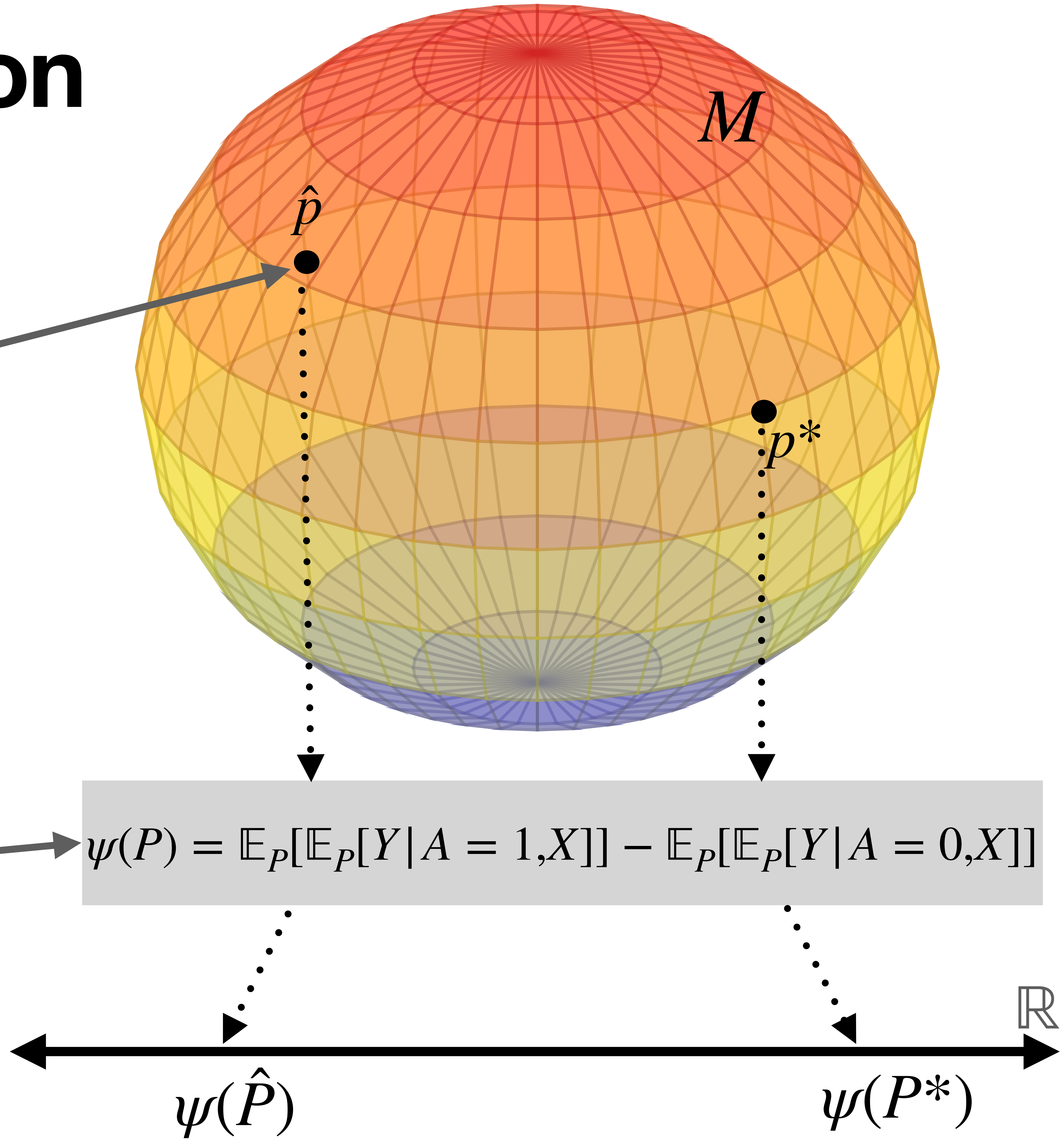
Plug-in Estimate $= \hat{\psi} = \psi(\hat{P}) = n^{-1} \sum_{i=1}^{n} f(1, X_i) - f(0, X_i)$

Naïve Plug-in Estimation

Estimated Distribution:
$\hat{P} = (\hat{P}_X, \hat{P}_{A|X}, \hat{P}_{Y|A,X})$

- $\hat{P}_x$: marginal empirical distribution
- $\hat{P}_{A=1|X}$: estimated propensities
- $\hat{P}_{Y=1|A,X}$: conditional regression func.

Target quantity of interest can be expressed as
$\psi : M \to \mathbb{R}$



What went wrong?

$$\psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y|A=1,X]] - \mathbb{E}_P[\mathbb{E}_P[Y|A=0,X]]$$

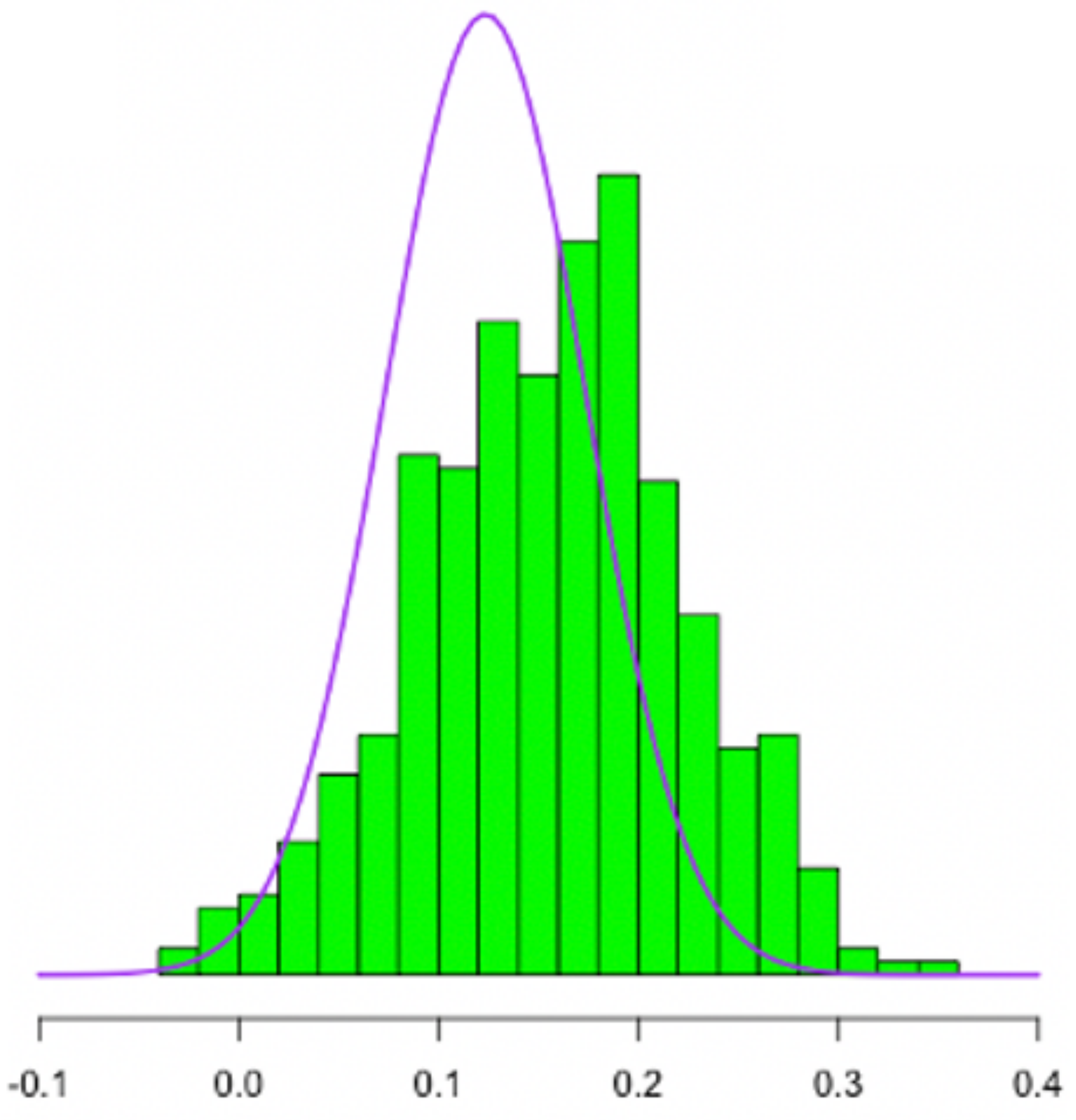


Fig 1: Distribution of Naïve Plug-in Estimates over 350 simulations, n=300.

Lemma 1:

$$\psi(\hat{P}) - \psi(P^*) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_{P^*}(O_i)}_{\to N(0,\mathbb{E}[\tilde{\psi}^2_{P^*}])} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_{\hat{P}}(O_i)}_{?} + o_{P^*}(1/\sqrt{n})$$

Purple Curve:
Distribution of an estimator $\hat{\psi}$ of $\psi(P^*)$ satisfying (A) and (B)

Plug-in Bias:
Generally, doesn't converge at the appropriate rate!

Outline

1. Naive Plug-in Estimation: Why does this fail?
   - **Plug-in bias fails criteria (A), (B)**
2. Existing Methods for Debiasing: TMLE
3. Our Method: KDPE!

"Good Estimator"?

A. Enables Uncertainty Quantification: tractable limiting distribution via. CLT.

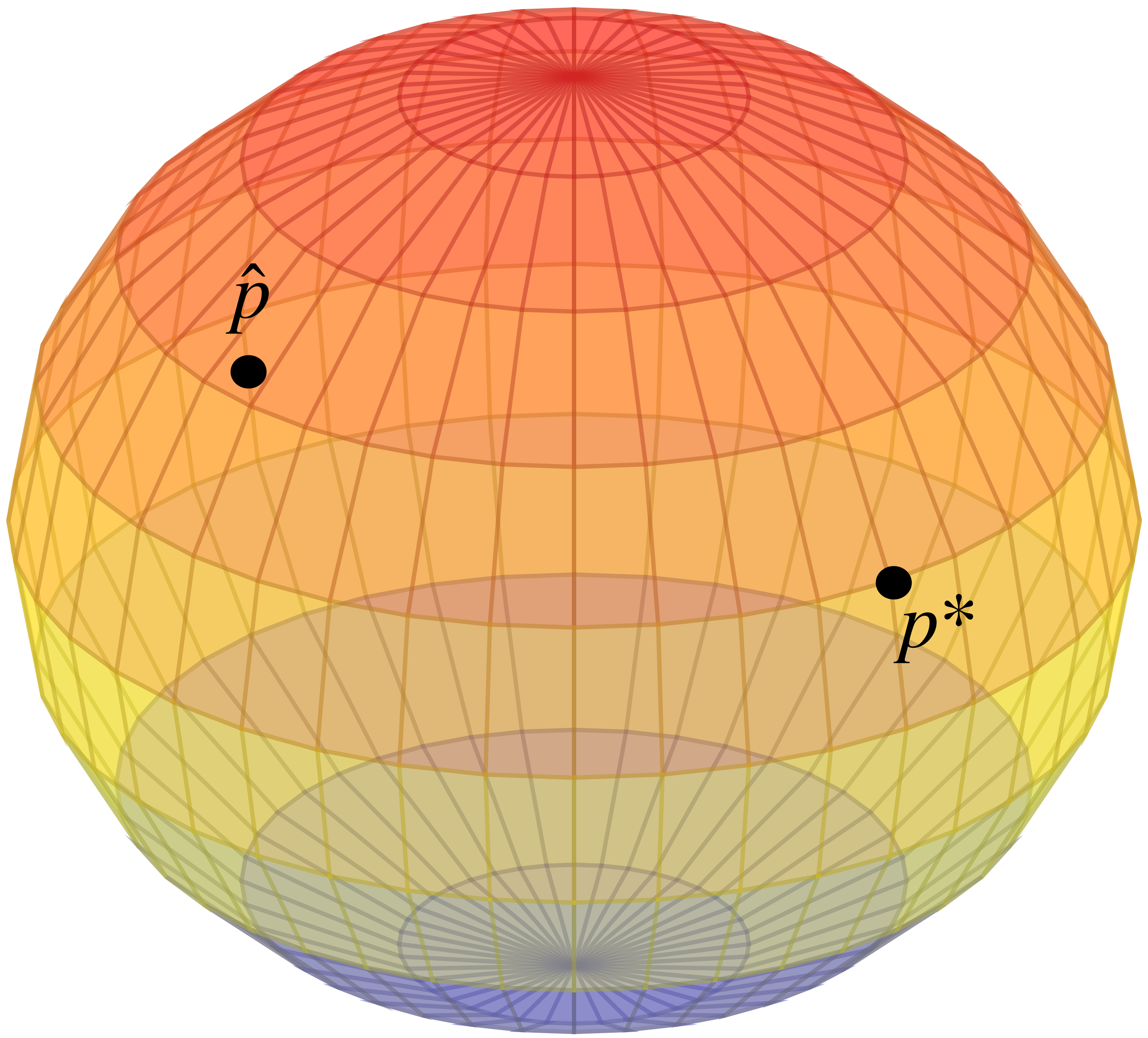
B. Data-efficient and consistent: converges to truth faster with less data

C. Retains simplicity of a plug-in approach

How do we find plug-in bias free $P \in M$ from $\hat{P}$?

<u>**How do we move in M?**</u>

1. **Scores:** “directions” we can move at $\hat{P}$
2. **MLE:** Along the direction we choose, how much do we move?



Set-up

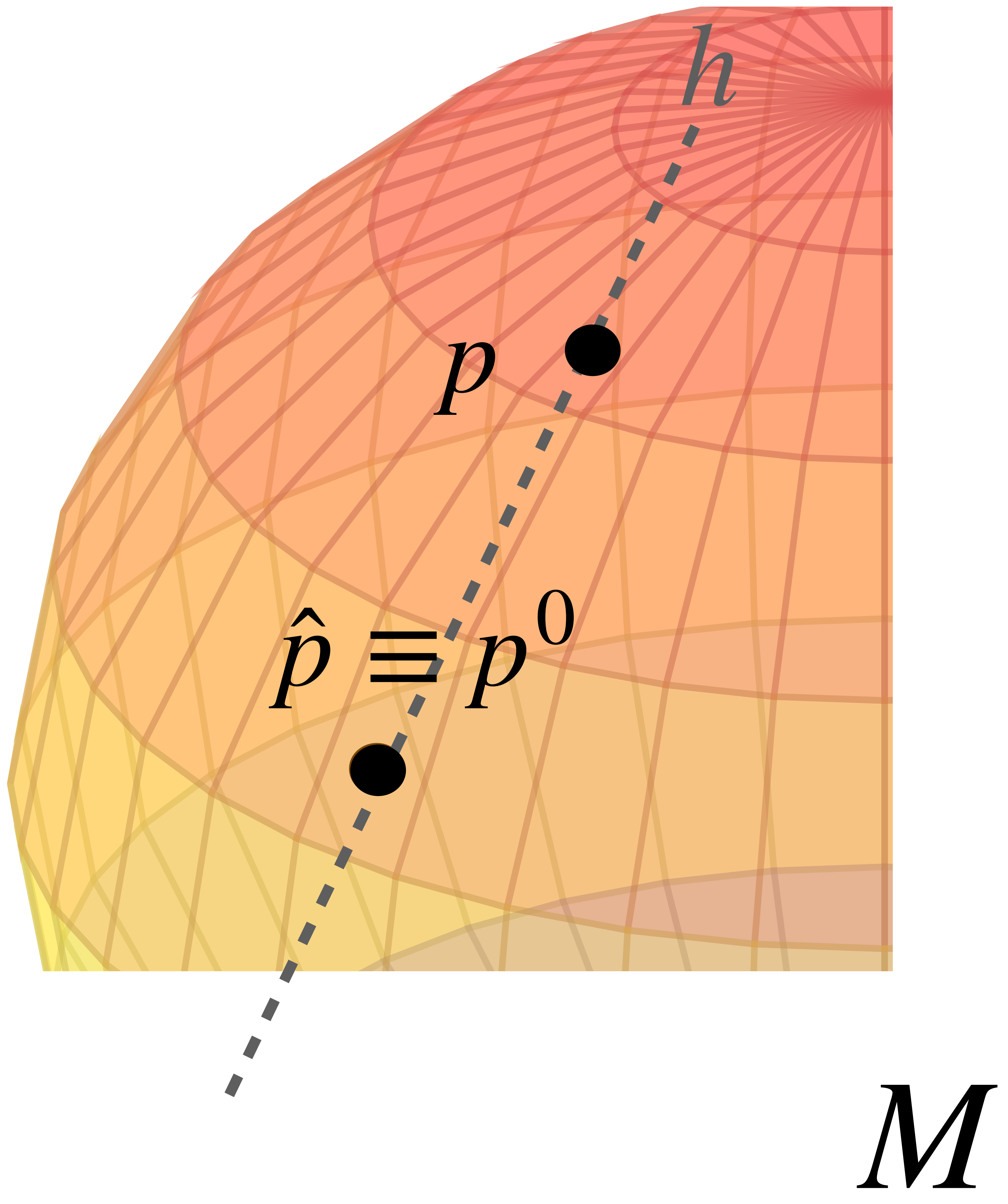
Tangent Spaces and Scores

Scores: One-Dimensional Sub-Models

$$p(\epsilon) = (1 - \epsilon)p^0 + \epsilon p = p^0(1 + \epsilon h)$$

where

$$h = \frac{p}{p^0} - 1 \text{ is the "direction"}$$



Set-up

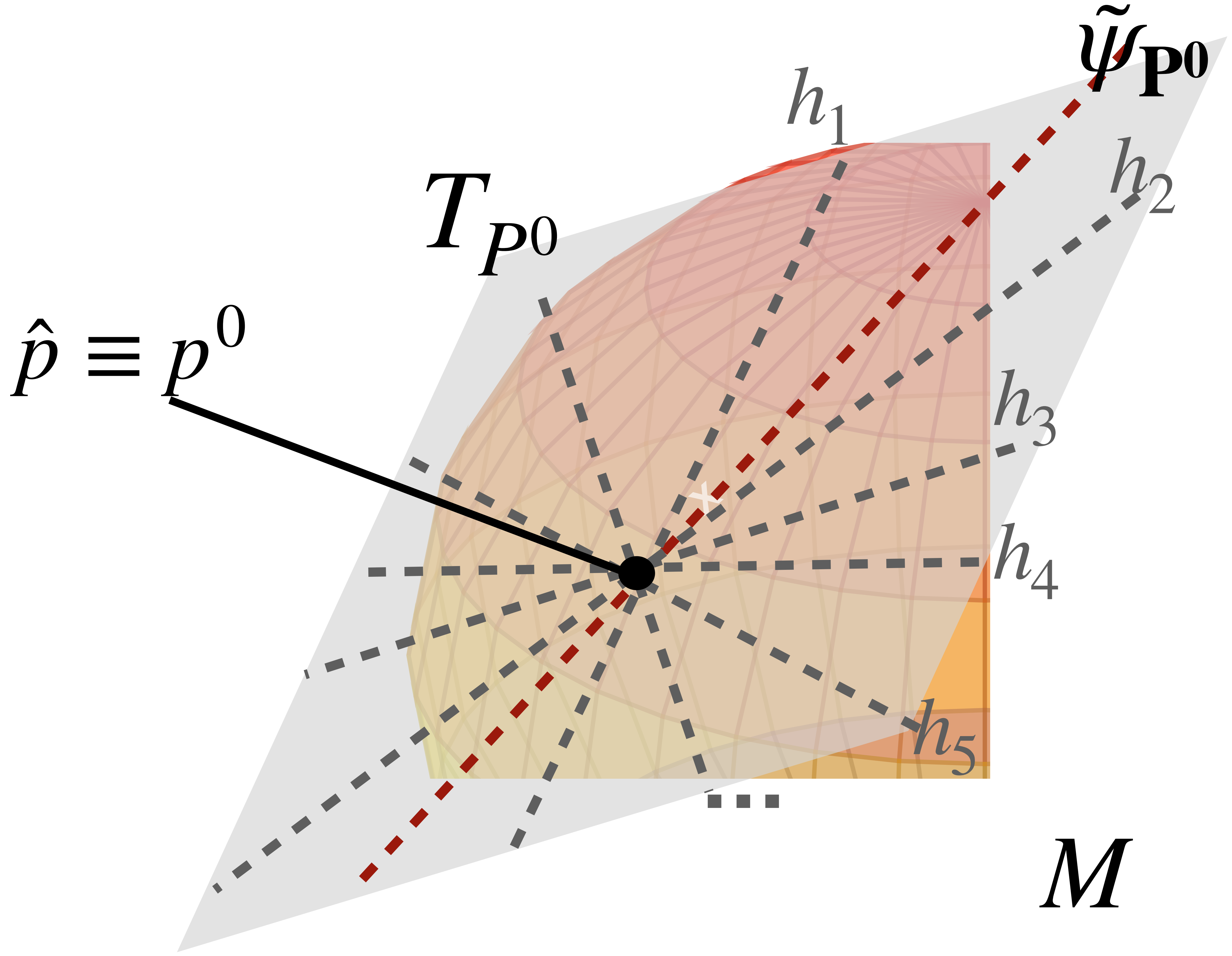
Tangent Spaces and Scores

Tangent Space:

$$T_P = \{h : \exists \epsilon_h \ s.t. (1 + \epsilon h)\ p \in M \ \forall \epsilon \leq \epsilon_h\}$$

$\tilde{\psi}_P \in T_P$ **is the direction of maximal change!**

$$\tilde{\psi}_P = \arg \max_{\|h\|=1, h \in T_P} \nabla_{\epsilon=0} \psi(\ [1 + \epsilon h]\ p\ )$$



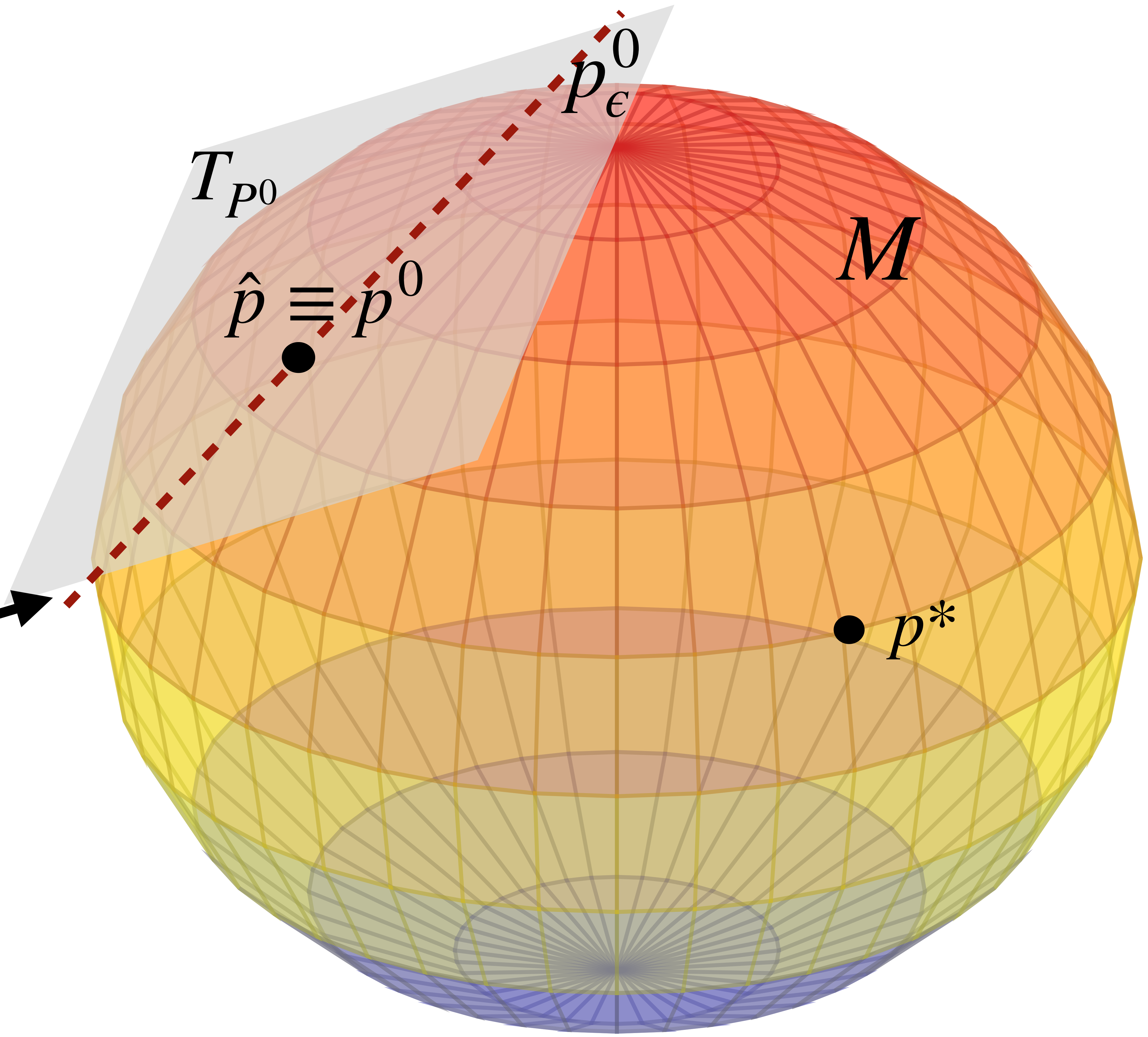
TMLE

van der Laan et. al. (2006)

(1): Construct the IF-based model:

$$p^0_\epsilon(O) = (1 + \epsilon\tilde{\psi}_{P^0}(O))\ p^0(O)$$

TMLE says to ONLY consider moving in the direction of $\tilde{\psi}_{P^0} \in T_P$!



TMLE

van der Laan et. al. (2006)

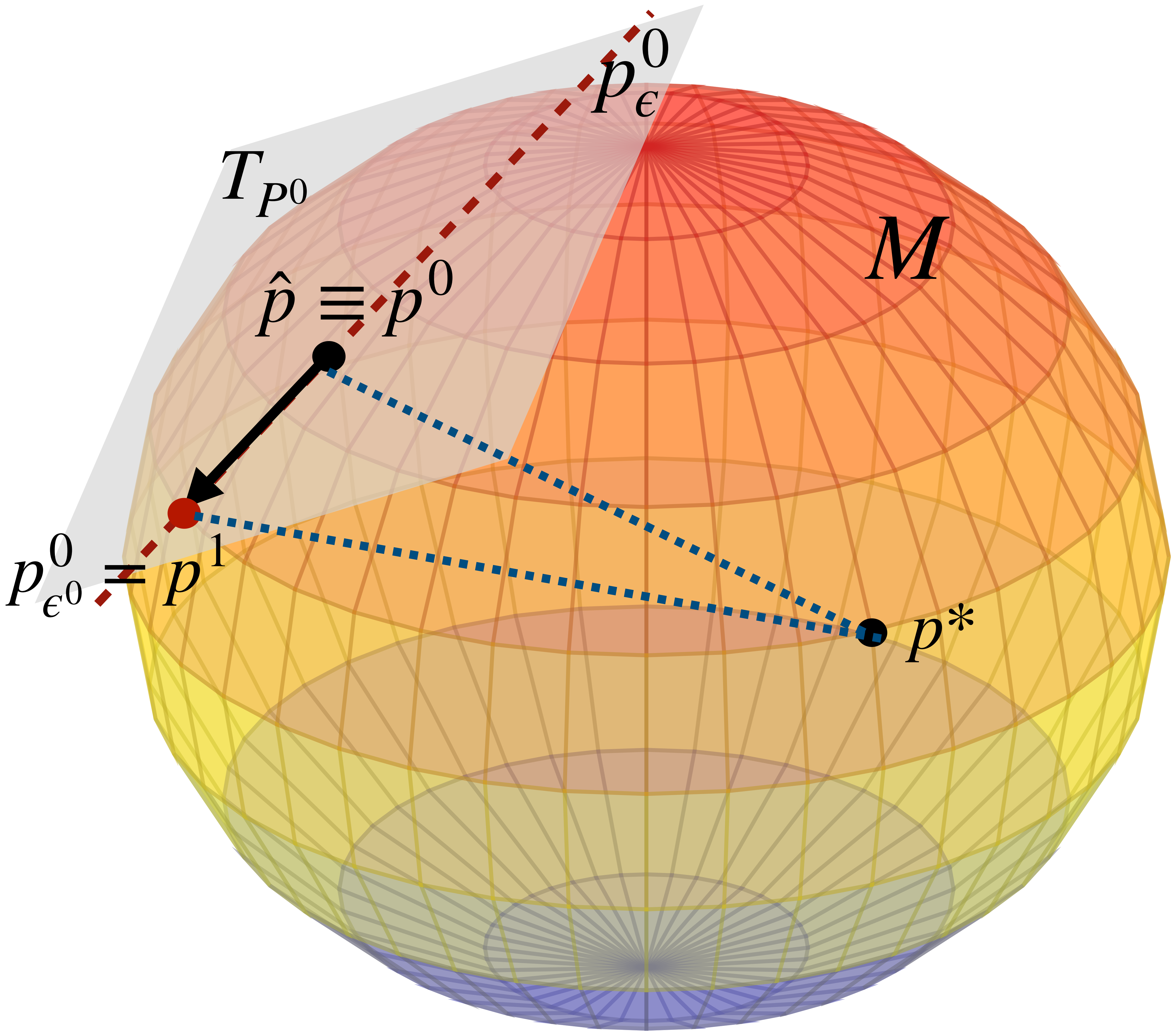
(1): Construct the IF-based sub-model:

$$p^0_\epsilon(O) = (1 + \epsilon\tilde{\psi}_{P^0}(O))\ p^0(O)$$

(2): Find ϵ^0 by MLE to get update!

$$\epsilon^0 = \arg\max_\epsilon \sum_{i=1}^{n} \log p^0_\epsilon(O_i)$$

$$p^1 = p^0_{\epsilon^0} = (1 + \epsilon^0\tilde{\psi}_{P^0})\ p^0$$



TMLE

van der Laan et. al. (2006)

Keep iterating this process until $\epsilon = 0$!

TMLE

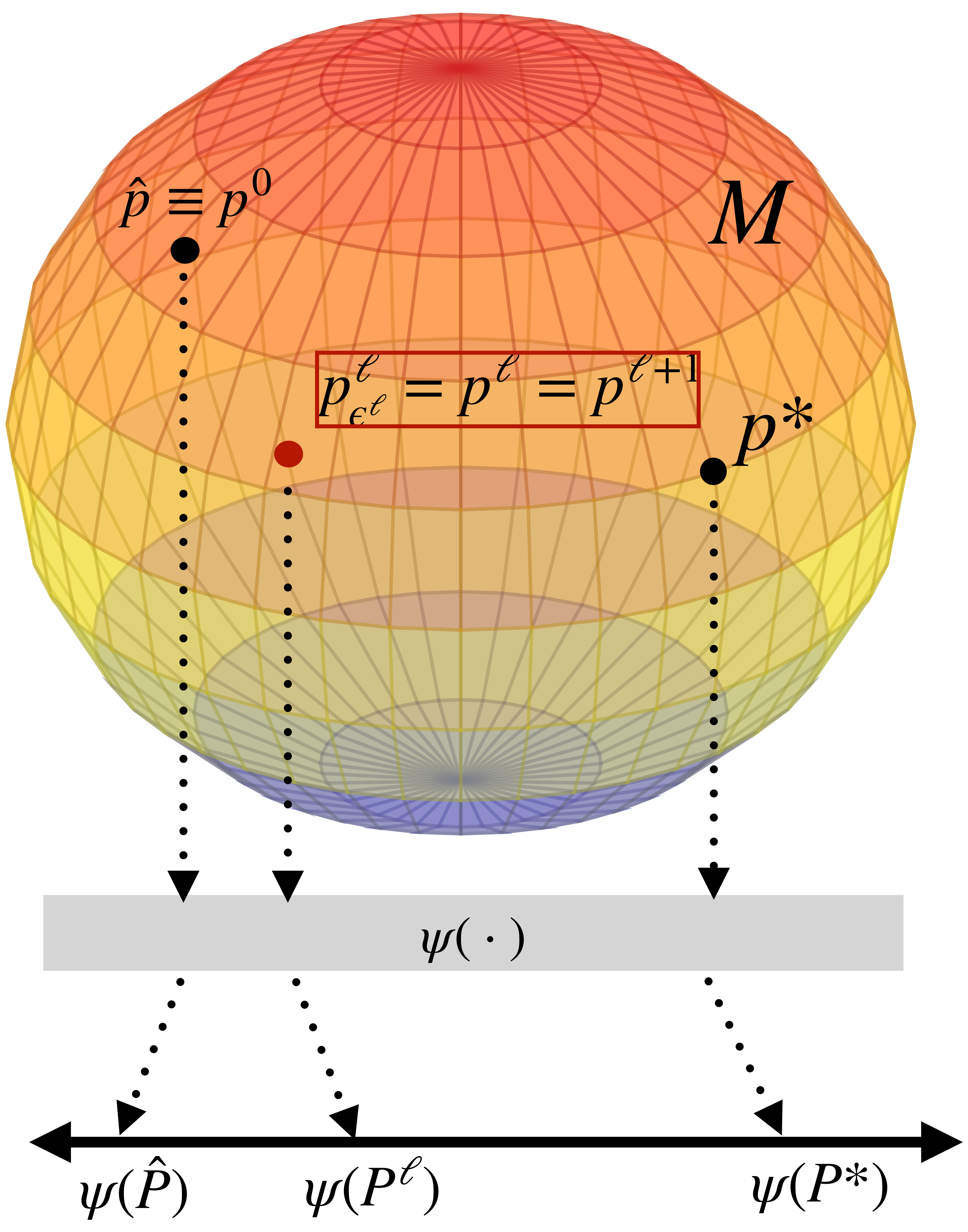
van der Laan et. al. (2006)

Lemma 1:

$$\psi(P^{\ell}) - \psi(P^{*}) \approx \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_{P^{*}}(O_i) - \frac{1}{n}\underbrace{\sum_{i=1}^{n} \tilde{\psi}_{P^{\ell}}(O_i)}$$

By FOC for MLE problem and $\epsilon = 0$,

$$\nabla_{\epsilon=0}\sum_{i=1}^{n}[\log(1+\epsilon\tilde{\psi}_{P^{\ell}})(O_i)\ p^{\ell}(O_i)] = \sum_{i=1}^{n}\tilde{\psi}_{P^{\ell}}(O_i) = 0!$$



Drawbacks of TMLE
(and other IF-based methods)

1. To run TMLE, we need the influence function $\tilde{\psi}_P$ for each ψ !
   - **Jordan et. al.** (2022): “deriving the actual (IF) that yields bias adjustment may require significant analytical effort.”
   - **Hines et. al.** (2021): “derivation of the IF often regarded as a ‘dark art’…not given much attention in traditional statistics education… some steps appearing as if from nowhere. ”
   - **Kennedy et. al.** (2019): “many researchers find IF-based estimators to be opaque or overly technical, which makes their use less prevalent and their benefits less available”

Drawbacks of TMLE
(and other IF-based methods)

1. **Need the influence function** $\tilde{\psi}_P$ for each quantity of interest ψ !
    - Jordan et. al. (2022), Hines et. al. (2021), Kennedy et. al. (2019)

2. Final plug-in P^ℓ built for ψ **doesn't work for a different quantity of interest** ψ'!

Outline

1. Naive Plug-in Estimation: Why does this fail?
   - **Plug-in bias fails criteria (A), (B)**
2. Existing Methods for Debiasing: TMLE
   - **Fails criteria (C)**
3. Our Method: KDPE!

“Good Estimator”?

A. Enables Uncertainty Quantification: tractable limiting distribution via. CLT.

B. Data-efficient: converges to truth faster with less data

C. Retains simplicity of a plug-in approach

Outline

1. Naive Plug-in Estimation: Why does this fail?
   - **Plug-in bias fails criteria (A), (B)**
2. Existing Methods for Debiasing: TMLE
   - **Fails criteria (C)**
3. Our Method: KDPE!

"Good Estimator"?

A. Enables Uncertainty Quantification: tractable limiting distribution via. CLT.

B. Data-efficient: converges to truth faster with less data

C. Retains simplicity of a plug-in approach

Can we find a general plug-in distribution P^{ℓ} that removes plug-in bias for many estimands ψ?

Kernel Debiased Plug-in Estimation

KDPE is a modified TMLE, with two major changes within in each iteration:

1. Only moving in the direction of $\tilde{\psi}_P \implies$ **moving in dense subset of T_P!**

2. Solving MLE for $\epsilon \in \mathbb{R} \implies$ **Solving MLE for $\alpha \in \mathbb{R}^n$**

Reproducing Kernel Hilbert Spaces

“Kernel” Debiased Plug-in Estimation

What is a Reproducing Kernel Hilbert Space?

Technical Definition:

Kernel Function: $K(O, O') = \exp(-\|O - O'\|_2^2)$

RKHS H**: set of functions that satisfy the following:**

1. $K(O, \cdot) \in H$ **for any** $O \in \mathcal{O}$
2. $\forall f \in H, O \in \mathcal{O}, \langle f, K(O, \cdot)\rangle = f(O)$

Why use RKHS?

1. **Certain RKHS (i.e. those associated with RBF Kernels) are sufficiently rich spaces.**
2. **Enables computationally tractable optimization for MLE**

KDPE

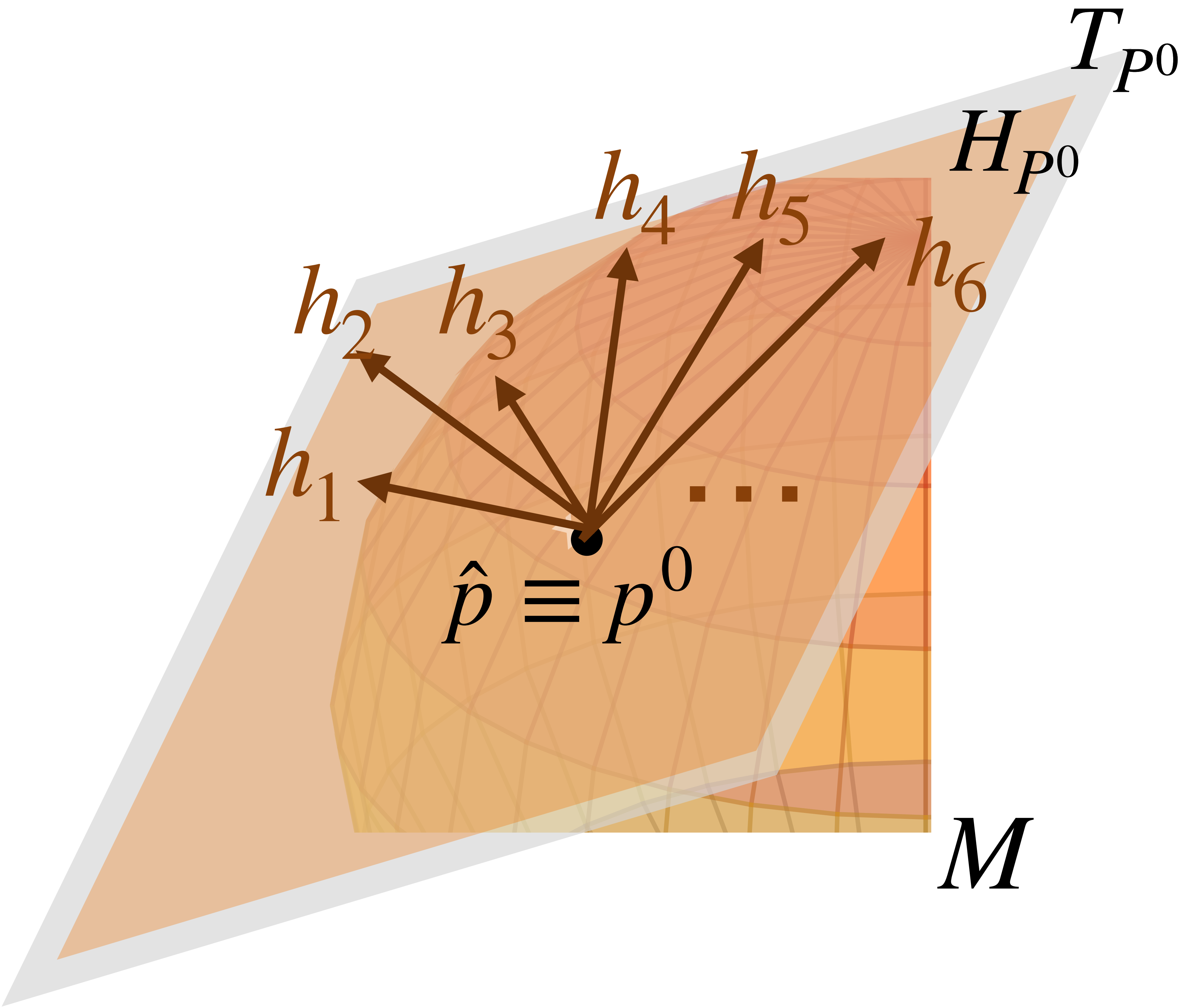
1. **Use an RKHS-Based Sub-model, not IF-based one:**

- Project universal RKHS H into tangent space T_P to get RKHS H_P

Lemma 1: By the fact that our chosen kernel K is *universal*, H_P, the projection of RKHS H, is **DENSE** in the set of feasible directions T_P and remains an RKHS.

- Construct RKHS-based model $\tilde{M}_{P0}$

$$\tilde{M}_{P0} = \{ \ (1+h)p^0 \ ; h \in H_{P0} \ \} \cap M$$



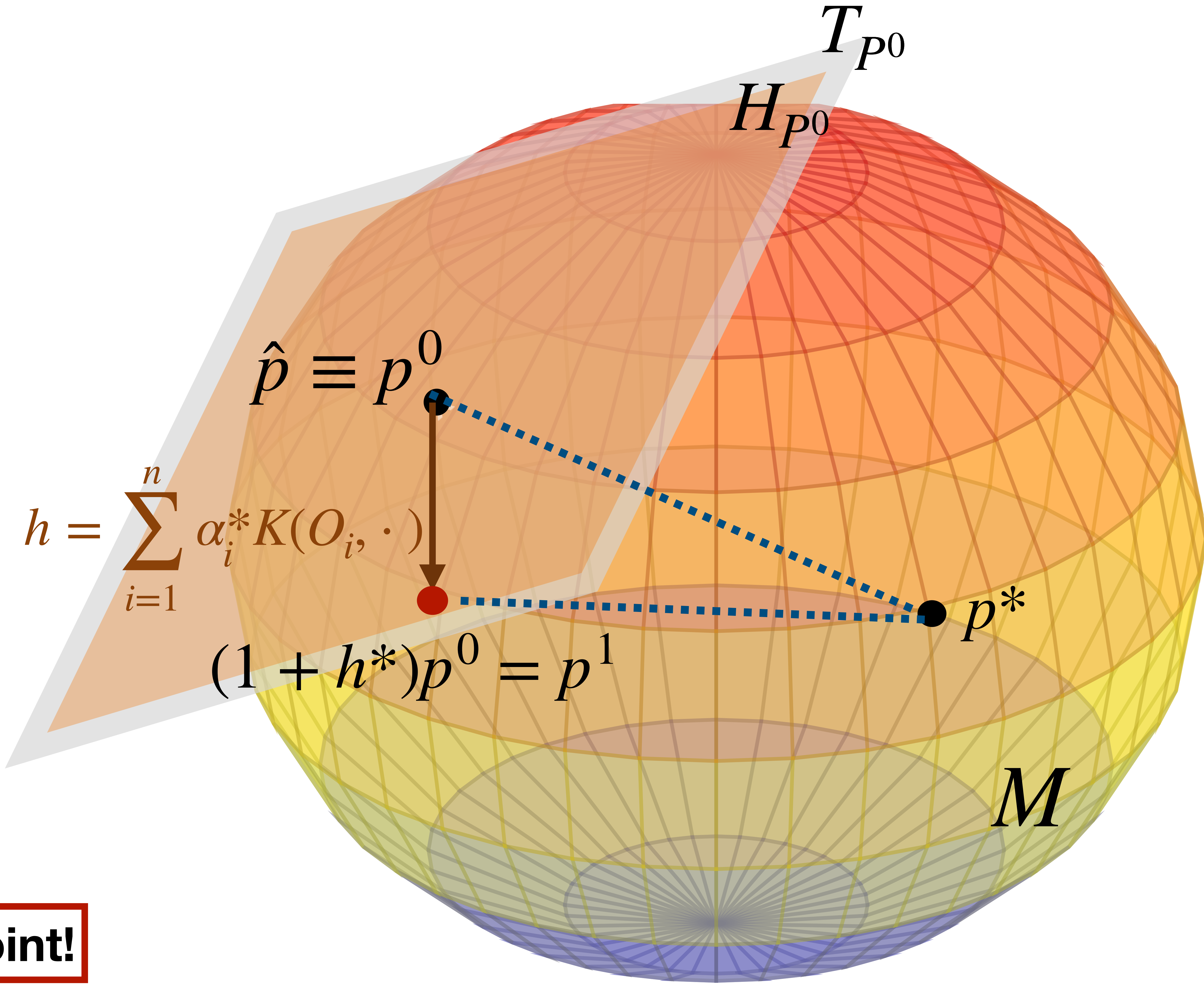
KDPE

2. MLE determines both step size and direction within $\tilde{M}_{P0}$.

Theorem (RKHS Representer):
For the MLE problem, defined as

$$h^* = \arg \max_{h \in H_{P\ell}, (1+h)p^0 \in \tilde{M}_P^0} \sum_{i=1}^{n} \log(\, [1 + h(O_i)]p^0(O_i) \,) - \lambda \|h\|_{H_{P0}}$$

The solution is given by $h^* = \sum_{i=1}^{n} \alpha_i^* K_{P0}(O_i, \cdot \,)$



Like TMLE, iterate these updates until we reach a fixed point!

Theoretical Guarantees for KDPE

Prop. 2:

$$\sum_{i=1}^{n} h(O_i) = 0 \text{ for ALL } h \in H_{P^\ell}$$

Key Takeaway

We are leveraging the first-order conditions, and NOT to obtain some optimal solution / approximate the influence function in any way.

Theoretical Guarantees for KDPE

For many quantities I want to estimate…

Theorem 2 (asymptotic linearity of KDPE). *Let $\psi : \mathcal{M} \to \mathbb{R}$ be a pathwise-differentiable functional of the distribution P with influence function $\tilde{\psi}_P \in L_0^2(P)$ and von Mises expansion:*

$$\psi(\bar{P}) - \psi(P) = \int \tilde{\psi}_{\bar{P}} d(\bar{P} - P) + R_2(\bar{P}, P) \quad \textit{for any } \bar{P}, P \in \mathcal{M},$$

which defines the second-order reminder term $R_2(\bar{P}, P)$. Then, under necessary regularity conditions, the plug-in bias satisfies $\mathbb{P}_n \tilde{\psi}_{\widehat{P}} = o_{P^*}(1/\sqrt{n})$ and the KDPE estimator satisfies

$$\psi(\widehat{P}) - \psi(P^*) = \mathbb{P}_n \tilde{\psi}_{P^*} + o_{P^*}(1/\sqrt{n}) \approx N(0, [\tilde{\psi}_{P^*}]^2/n).$$

$$n^{-1} \sum_{i=1}^{n} \tilde{\psi}_{P*}(O_i)$$

Plug-in Bias Disappears appropriately!

Outline

1. Naive Plug-in Estimation: Why does this fail?
   - **Plug-in bias fails criteria (A), (B)**
2. Existing Methods for Debiasing: TMLE
   - **Fails criteria (C)**
3. Our Method: KDPE!
   - **Satisfies (A), (B), (C)**

"Good Estimator"?

A. Enables Uncertainty Quantification: tractable limiting distribution via. CLT.

B. Data-efficient: converges to truth faster with less data

C. Retains simplicity of a plug-in approach

Can we find a general plug-in distribution P^ℓ that removes plug-in bias for many estimands ψ? KDPE!

Simulation Studies

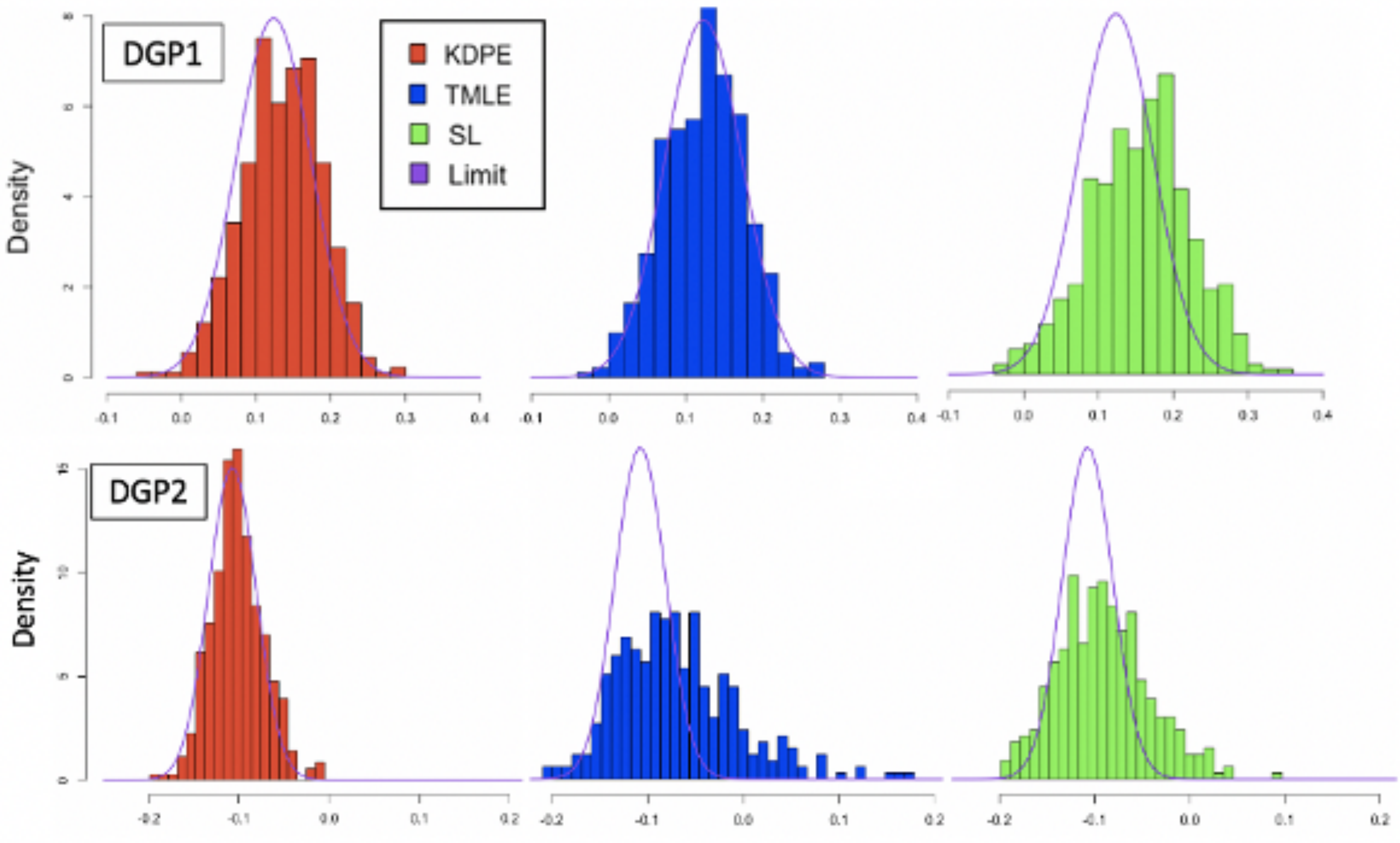


Figure 1: Simulated distributions of $\widehat{\psi}_{\mathrm{ATE}}$ compared to their asymptotic distributions. TMLE distribution in the second row corresponds to LTMLE for DGP2.

Simulation Studies

	Method	ψ_{ATE}	ψ_{RR}	ψ_{OR}
	SL	0.0803	0.2623	0.6796
DGP1	TMLE	0.0574	0.1723	0.4059
	KDPE	0.0592	0.1752	0.4303
	SL	0.0508	0.0925	0.1555
DGP2	LTMLE	0.0731	0.1481	0.2648
	KDPE	0.0295	0.0778	0.0827

Table 1: Root Mean Squared Error (RMSE) of KDPE, (L)TMLE, and SL for DGP1, DGP2

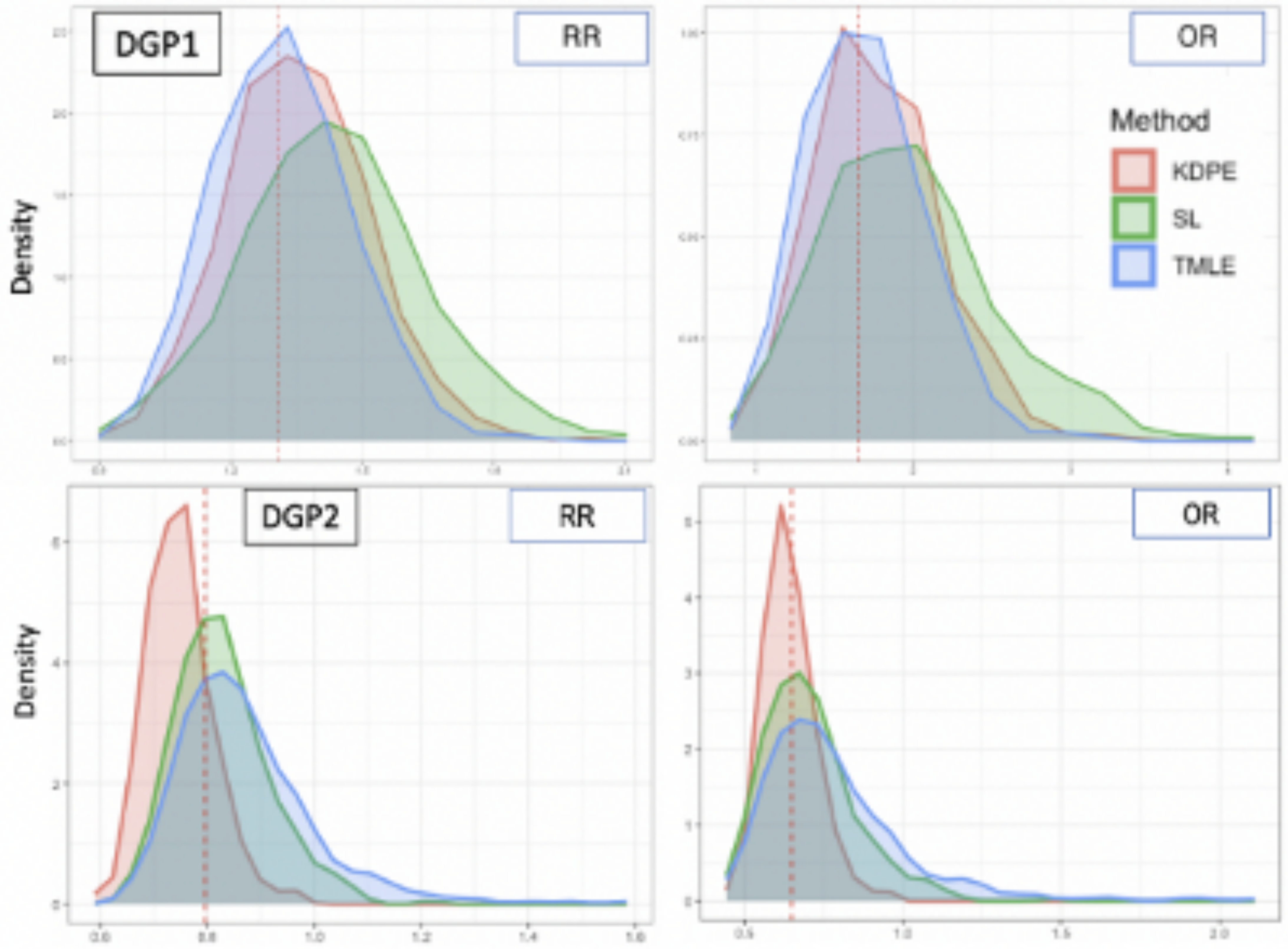


Figure 2: Simulated distributions of $\widehat{\psi}_{\text{RR}}, \widehat{\psi}_{\text{OR}}$. First row corresponds to DGP1, and second row to DGP2. Red line denotes true value of the target parameter.

References

- Targeted Maximum Likelihood Learning.
  Mark J. van der Laan and Daniel Rubin (2006).
- Empirical gateaux derivatives for causal inference.
  Michael I. Jordan, Yixin Wang, and Angela Zhou. (2022)
- Demystifying statistical learning based on efficient influence functions.
  Hines, O., Dukes, O., Diaz-Ordaz, K., and Vanstee- landt, S. (2022)