Beyond Expected Information Gain: Stable Bayesian Optimal Experimental Design with Integral Probability Metrics and Plug-and-Play Extensions

Di Wu
Joint work with Ling Liang and Haizhao Yang

University of Maryland, College Park

May 9, 2026

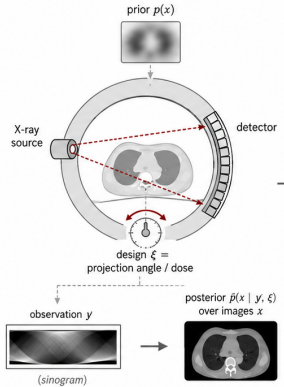
Outline

1. Motivation
2. IPM-based BOED
3. Theoretical Stability
4. Experiments
5. Conclusion

The Whole Story

Stable & Scalable BOED via Integral Probability Metrics

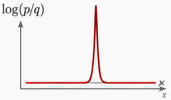
1 BOED in CT Imaging



Few projections, max information — choose ξ that shrinks posterior uncertainty most.

2 Measuring Information Gain: KL vs IPM

KL / EIG ✗



non-Lipschitz · needs absolute continuity · nested MC

IPM ✓

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P f - \mathbb{E}_Q f \right|$$

W_1 | MMD | ED

sample-based · geometry-aware · no density ratios

3 Guarantees & Scalability

Likelihood stability

Prior stability

$$|U_{\mathcal{F}} - U^*_{\mathcal{F}}| \leq \bar{K} \cdot \gamma_{\mathcal{F}}(\mu, \tilde{\mu})$$

Plug-and-play OT-ICNN — scales to high-D imaging where nested MC fails.

Why Bayesian Optimal Experimental Design (BOED)?

Data acquisition is the bottleneck in many scientific and ML tasks:

- Calibration of expensive simulators (PDEs, climate, materials).
- Inverse problems with limited measurements.
- Active data selection / preference learning in modern ML.

The BOED question

Given a prior $p(x)$ and a forward model, *which experiment ξ should we run to learn x most efficiently?*

Mathematical Framework

- Parameter $x \in \mathcal{X}$ with prior $p(x)$.
- Design $\xi \in \Xi$ controls the experiment.
- Observation $y = G_\xi(x) + \eta$ from a forward model with noise η.
- Bayes' rule yields the posterior $p(x \mid y, \xi)$.

Goal

Choose ξ^* that **maximizes the expected discrepancy** between prior and posterior — a more informative experiment moves the posterior further from the prior.

Classical choice: KL divergence $\Longrightarrow$ **Expected Information Gain (EIG)**.

Classical EIG: The Default and Its Drawbacks

$$\xi^* \;=\; \arg\max_{\xi\in\Xi} \; \mathbb{E}_{p(y|\xi)}\Big[\mathrm{KL}\big(p(x \mid y,\xi) \,\|\, p(x)\big)\Big].$$

Three structural problems with KL-based EIG

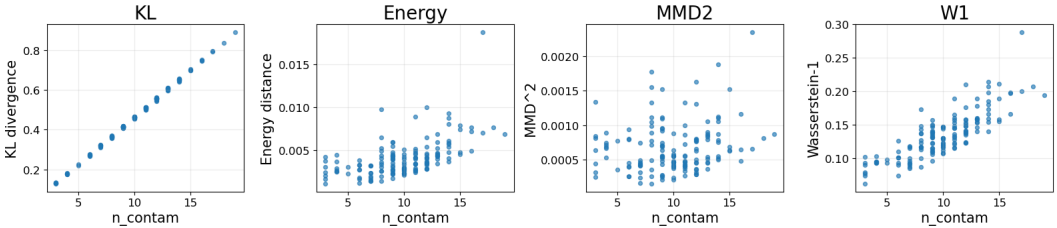
1. **Tail / rare-event fragility:** log density-ratios explode in low-probability regions.
2. **Strict support condition:** requires absolute continuity — breaks under empirical priors.
3. **Computational bottleneck:** nested expectations $\Rightarrow$ high-variance nested Monte Carlo.

Estimator-level fixes (variational, NMC) leave the underlying log-density-ratio objective unchanged.

Motivating Example: KL is Fragile to Rare Events

Two distributions differing only by a low-probability tail event ($\epsilon = 10^{-2}$):

$$Q \sim \mathcal{N}(0,1), \qquad P = (1-\epsilon)\mathcal{N}(0,1) + \epsilon\,\mathcal{N}(10,\, 0.1^2).$$



KL response is nearly **linear in the contamination count** $n_{\text{contam}} \Rightarrow$ extreme variance.
IPMs (W_1, MMD, ED) respond **smoothly and robustly**.

A Paradigm Shift: Integral Probability Metrics

Replace density-ratio metrics with a **geometry-aware** discrepancy:

$$\gamma_{\mathcal{F}}(P, Q) \;=\; \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \right|.$$

Why IPMs?

- Operate on **measures**, not densities $\Rightarrow$ no support overlap required.
- **Sample-based** estimation $\Rightarrow$ no log-density evaluation, no bandwidth tuning.
- Choosing the test class $\mathcal{F}$ encodes geometric structure (Lipschitz / RKHS / kernel).

Three Representative IPMs

1-Wasserstein (W_1)

$\mathcal{F} = \{1\text{-Lipschitz functions}\}$. Captures **transport geometry** of the parameter space.

Maximum Mean Discrepancy (MMD)

$\mathcal{F}$ = unit ball of an RKHS $\mathcal{H}_k$. With **bounded** radial kernels (Gaussian, Matérn).

Energy Distance (ED)

$\mathrm{ED}_\rho(P, Q)^2 = 2\,\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y')$.
Equivalent to MMD under a specific (unbounded) kernel; sublinear growth.

Plug-and-play: same BOED template, different geometry.

The IPM-based BOED Framework

We define the **IPM design utility**:

$$U_{\mathcal{F}}(\xi) := \int_{\mathcal{Y}} \gamma_{\mathcal{F}}\big(p(x),\, p(x \mid y, \xi)\big)\, p(y \mid \xi)\, dy.$$

- Same problem, different "ruler".
- No log-density-ratio anywhere $\Rightarrow$ stable to support mismatch and rare events.
- Estimable from **samples** only — no density estimation step.

Two questions to address

Theory: How does utility behave under model misspecification?
Practice: Does it give better designs and scale to high dimensions?

Stability of Design Utilities

If we slightly perturb the likelihood (surrogate error) or the prior, how much does the design utility change?

Stability is governed by the **growth profile**:

$$\omega_{\mathcal{F}}(x) := \sup_{f \in \mathcal{F}} |f(x)|.$$

IPM	Growth profile $\omega_{\mathcal{F}}(x)$
MMD (bounded kernel)	$\mathcal{O}(1)$
Energy Distance	$\mathcal{O}(\|x\|^{1/2})$
Wasserstein-1	$\mathcal{O}(\|x\|)$

The smaller the growth, the more universal the stability bound.

Theorem 1: Likelihood Stability

Likelihood Stability Bound

Let $p(y|x), p_*(y|x)$ be the true and surrogate likelihoods, with utilities $U_{\mathcal{F}}, U^*_{\mathcal{F}}$. Then

$$\left|U_{\mathcal{F}} - U^*_{\mathcal{F}}\right| \;\leq\; \int_{\mathcal{Y}}\int_{\mathcal{X}} \left(\omega_{\mathcal{F}}(x) + \bar{\omega}\right) \left|p(y|x) - p_*(y|x)\right| \mu(dx)\, dy.$$

- Utility error $\leq L^1$ **likelihood error**, weighted by the growth profile.
- **Bounded MMD:** unconditional, uniform L^1 control.
- **Unbounded IPMs:** bound holds under prior moment + Lipschitz forward model.

Contrast with KL. A near-zero surrogate likelihood causes **logarithmic blow-up** — no analogous bound exists for KL.

Theorem 2: Prior Stability

Modern pipelines replace continuous μ with empirical $\tilde{\mu}$ — where KL is undefined.

Prior Stability Bound

Under a mild compatibility assumption,

$$\left|U_{\mathcal{F}} - \tilde{U}_{\mathcal{F}}\right| \;\leq\; \bar{K} \cdot \gamma_{\mathcal{F}}(\mu, \tilde{\mu}), \qquad \bar{K} \;=\; 1 + \int_{\mathcal{Y}} \left(C_L(y) + 2\, C_E(y)\, D(y)\right) dy.$$

- Utility error **linear** in IPM distance between priors.
- **Bounded MMD:** $\bar{K}$ finite without any tail assumption.
- **Unbounded IPMs:** $\bar{K}$ finite under sub-Gaussian prior + Lipschitz G.

Contrast with KL. The KL divergence between continuous and empirical μ is **undefined**.

Bounded vs. Unbounded IPMs: A Principled Trade-off

Metric	Growth $\omega_{\mathcal{F}}(x)$	Universal stability	Geometric richness
MMD (bounded kernel)	$\mathcal{O}(1)$	**Yes**	Kernel-dependent
Energy Distance	$\mathcal{O}(\|x\|^{1/2})$	Sub-Gaussian + Lipschitz G	Sublinear transport
Wasserstein-1	$\mathcal{O}(\|x\|)$	Sub-Gaussian + Lipschitz G	**Full transport**
KL (EIG)	— (log-ratio)	None	—

- No free lunch: stronger geometry $\Rightarrow$ stronger tail conditions on the prior.
- All IPM choices avoid the **logarithmic small-denominator instability** of KL.
- User picks based on application: tail-robustness (MMD) vs. transport sensitivity (W_1).

Experiment 1: Robustness Under Coarse Discretization

Preference learning (1D, nonlinear sigmoid likelihood, 81 candidate designs). High-utility region $R_t = \{d : U(d) \geq t \cdot \max U\}$; restrict to every k-th grid point and report mean (max) normalized regret.

		Mean normalized regret $\bar{r}$ (max in parentheses)			
Metric	$\lvert R_{0.80}\rvert$	$k = 2$	$k = 4$	$k = 6$	$k = 8$
KL Divergence	2	0.018 (0.035)	0.121 (0.231)	0.180 (0.337)	0.234 (0.419)
MMD^2	4	0.080 (0.160)	0.118 (0.161)	0.150 (0.225)	0.195 (0.383)
Energy Distance	7	0.068 (0.135)	0.093 (0.151)	0.117 (0.168)	0.148 (0.299)
Wasserstein-1	**11**	**0.001** (0.002)	**0.023** (0.056)	**0.038** (0.081)	**0.054** (0.134)

- Wider near-optimal region ($\lvert R_{0.80}\rvert$) $\Leftrightarrow$ smaller regret under coarsening.
- KL: regret grows **13**$\times$ as the grid coarsens (k: 2 $\rightarrow$ 8); W_1: only **54**$\times$ in absolute terms but stays **below** 6%.
- Geometric stability $\Rightarrow$ more reliable design under MC noise and approximate optimization.

Experiment 2: High-Dim Linear-Gaussian ($p = 64$)

Conjugate model with closed-form W_2^2 and EIG ground truth:

$$x \sim \mathcal{N}(0, I_p), \qquad y \mid x, \xi \sim \mathcal{N}(\xi x,\, \sigma^2 I_p), \qquad \xi \in \{0.5, 2, 6\},\ \sigma = 0.1.$$

Design	W_2^2 utility (OT-ICNN) Estimate	W_2^2 utility (OT-ICNN) Rel. err.	EIG (Nested MC) Estimate	EIG (Nested MC) Rel. err.
$\xi = 0.5$	1.01×10^2	$\mathbf{2.18 \times 10^{-2}}$	9.49×10^2	8.10×10^0
$\xi = 2$	1.21×10^2	$\mathbf{2.70 \times 10^{-3}}$	1.54×10^4	7.91×10^1
$\xi = 6$	1.25×10^2	$\mathbf{3.10 \times 10^{-3}}$	1.34×10^5	5.12×10^2

Plug-and-play with neural OT (ICNN) keeps relative error below 2%; nested-MC **fails by 2–3 orders of magnitude**.

Experiment 3: Sign-Ambiguous Bimodal Posterior ($d = 32$)

$y_i = x_i^2 + \epsilon_i$ with bimodal prior $\Rightarrow$ **Gaussian VB is fundamentally misspecified**.

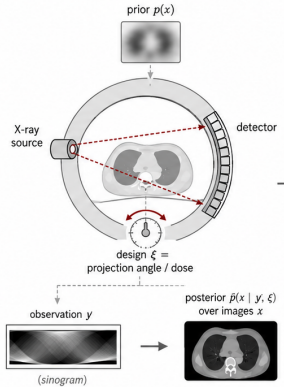
Design	Composition	EIG		W_2^2	
		Exact	Gaussian VB	Exact	OT-ICNN
A	8 strong	21.09 (1)	−13.91 (5)	1.943 (1)	**1.608** (1)
B	8 medium	15.53 (2)	−8.27 (4)	1.596 (2)	**1.585** (2)
C	8 weak	10.00 (4)	−2.53 (2)	1.337 (3)	**1.249** (3)
D	4 strong + 4 null	11.61 (3)	−6.34 (3)	1.097 (4)	**0.793** (4)
E	8 null	2.12 (5)	+1.23 (1)	0.251 (5)	**0.161** (5)
Ranking		ABDCE	ECDBA	ABCDE	**ABCDE**

Gaussian VB completely flips the ranking (worst design becomes the "best").
OT-ICNN within our framework recovers the **exact correct order**.

Closing the Loop

Stable & Scalable BOED via Integral Probability Metrics

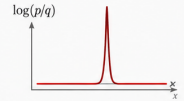
1 BOED in CT Imaging



Few projections, max information — choose ξ that shrinks posterior uncertainty most.

2 Measuring Information Gain: KL vs IPM

KL / EIG ✗



non-Lipschitz · needs absolute continuity · nested MC

IPM ✓

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P f - \mathbb{E}_Q f \right|$$

W_1 | MMD | ED

sample-based · geometry-aware · no density ratios

3 Guarantees & Scalability

Likelihood stability

Prior stability

$$|U_{\mathcal{F}} - U^*_{\mathcal{F}}| \leq \bar{K} \cdot \gamma_{\mathcal{F}}(\mu, \tilde{\mu})$$

Plug-and-play OT-ICNN — scales to high-D imaging where nested MC fails.

Summary

Beyond EIG: stable, sample-based BOED with IPMs

1. **Theory.** Likelihood + prior stability bounds, controlled by the growth profile $\omega_{\mathcal{F}}(x)$ — no log-ratio fragility.
2. **Practice.** Broader near-optimal regions, smoother landscapes, robust to coarse-grid selection.
3. **Scalability.** Plug-and-play with neural OT estimators handles 64-D conjugate *and* 32-D bimodal cases where EIG estimators collapse.

arXiv:2604.21849

Thank you!

Any questions or feedback are very welcome.