Beyond Expected Information Gain

Stable Bayesian Optimal Experimental Design with Integral Probability Metrics

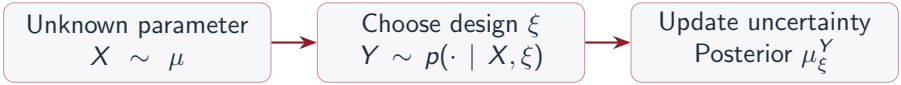
Di Wu

Joint work with Ling Liang and Haizhao Yang

University of Maryland, College Park

INFORMS Meeting

BOED: which experiment should we run?



Classical objective: maximize expected information gain (EIG),

$$\xi^* \in \arg\max_{\xi \in \Xi} \mathrm{EIG}(\xi), \qquad \mathrm{EIG}(\xi) = \mathbb{E}_{Y \sim m_\xi}\left[D_{\mathrm{KL}}(\mu_\xi^Y \| \mu)\right].$$

- **Utility choice:** which posterior changes matter for the task?
- **Numerical estimation:** how accurately can we evaluate that utility?

A better estimator improves evaluation; the choice of utility determines what we reward.

Wu, Liang & Yang: Introduction.

A finite utility can distinguish designs when EIG cannot

Independent coordinates $X_i \sim \mathrm{Unif}[-a_i, a_i]$; design j observes $Y = X_j$ exactly.

EIG: singular posterior

The posterior lies on $\{x : x_j = y\}$, a set of prior probability zero.

$$\mathrm{EIG}(j) = +\infty \quad \text{for every } j.$$

Wasserstein-1: finite transport cost

Under the Euclidean ground metric,

$$U_{W_1}(j) = \frac{2a_j}{3}, \qquad R(j) = \frac{1}{3}\sum_{i \neq j} a_i^2.$$

R: posterior-mean squared-error Bayes risk.

$(a_1, a_2, a_3) = (3, 2, 1)$	Observe X_1	Observe X_2	Observe X_3
W_1 utility $\uparrow$	**2**	$4/3$	$2/3$
Bayes risk $\downarrow$	**$5/3$**	$10/3$	$13/3$

Here, maximizing W_1 utility selects the Bayes-risk-optimal experiment.

Wu, Liang & Yang: Section 3.2, exact-observation example.

IPMs let us specify which distributional changes matter

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_P f - \mathbb{E}_Q f|, \qquad \boxed{U_{\mathcal{F}}(\xi) = \mathbb{E}_{Y \sim m_\xi}\left[\gamma_{\mathcal{F}}(\mu, \mu_\xi^Y)\right].}$$

Discrepancy	Test functions / construction	What the choice emphasizes
Wasserstein-1	1-Lipschitz functions	Displacement in the parameter metric
MMD	Unit ball of an RKHS	Features and scales selected by the kernel
Energy distance	Distance-induced RKHS†	Pairwise distances; no kernel bandwidth

- Evaluate discrepancies from prior and posterior samples.
- Choose geometry, coordinate scaling, and kernel to match the inference task.

†The theory uses $\mathrm{ED}/\sqrt{2} = \mathrm{MMD}_{k_{x_0}}$; this fixed factor preserves the maximizer.

Wu, Liang & Yang: Section 3.1.

Likelihood stability: geometry weights model error

Same prior μ, two likelihoods p and q, each with its own predictive and posterior.

$$d_{\mathcal{F}}(x,z) = \sup_{f\in\mathcal{F}} |f(x) - f(z)|, \qquad \Omega_{\mathcal{F},\mu}(x) = \mathbb{E}_{Z\sim\mu} d_{\mathcal{F}}(x,Z).$$

Likelihood stability theorem

If $d_{\mathcal{F}}$ is measurable and $\mathbb{E}_{X,Z\overset{\text{iid}}{\sim}\mu} d_{\mathcal{F}}(X,Z) < \infty$, then

$$|U_{\mathcal{F},p} - U_{\mathcal{F},q}| \le \mathbb{E}_{X\sim\mu}[\Omega_{\mathcal{F},\mu}(X)\, \|p(\cdot \mid X) - q(\cdot \mid X)\|_{L^1}].$$

Metric	Pairwise geometry $d_{\mathcal{F}}(x,z)$	Implication
Bounded-kernel MMD	$\le 2\kappa,\ k(x,x) \le \kappa^2$	Uniformly bounded weight
Normalized Euclidean ED	$\|x-z\|^{1/2}$	Moment-dependent weight
W_1	$\|x-z\|$	Stronger tail dependence

Wu, Liang & Yang: Section 4.2, likelihood stability theorem.

Prior stability requires likelihood–class compatibility

Fix the likelihood; replace the prior μ by $\widetilde{\mu}$.

Prior stability theorem

Under fixed-class evidence and multiplier conditions,

$$|U_{\mathcal{F}} - \widetilde{U}_{\mathcal{F}}| \le \overline{K}\,\gamma_{\mathcal{F}}(\mu, \widetilde{\mu}), \qquad \overline{K} = 1 + \int [C_L(y) + C_E(y) B_{\mathcal{F}}(\widetilde{\mu})]\, dy,$$

where $B_{\mathcal{F}}(\widetilde{\mu}) = \sup_{f \in \mathcal{F}} |\widetilde{\mu}(f)| < \infty$ and C_E, C_L are integrable.

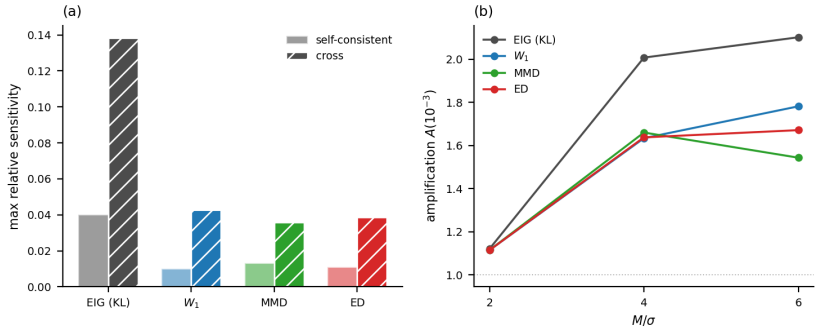
- C_E: controls changes in the evidence; C_L: controls likelihood weighting.
- Use an anchored class for W_1/ED and the native RKHS unit ball for MMD.
- Conditions are verified for additive Gaussian noise on $[0, R]$ with W_1, Matérn-MMD, and normalized ED.

IPM error can control utility error for empirical priors, when compatibility holds.

Wu, Liang & Yang: Section 4.3, Assumption 1, Proposition 4.1, prior stability theorem.

Likelihood misspecification: compare the utilities directly

$X \sim \mathcal{N}(0,1), \quad p(y \mid x, \xi) = (1-\epsilon)\mathcal{N}(\xi x, \sigma^2) + \epsilon \mathcal{N}(\xi x + M, \tau^2).$
Surrogate q replaces ϵ by $\delta \leq \epsilon$; utilities use deterministic quadrature.



Self-consistent: all quantities under each model. **Cross:** surrogate posterior, true observations.
$A(10^{-3})$: cross-error amplification relative to $\delta/\epsilon = 10^{-1}$.

Wu, Liang & Yang: Section 3.2, likelihood misspecification experiment ($\sigma = 1$, $\tau = 0.5$).

Preference learning: choose which item to query

Unknown preference threshold

$$X \sim \mathcal{N}(-20, 20^2).$$

Choose item characteristic ξ. Observe a noisy preference response Y.

Observation model

$$\eta \mid X, \xi \sim \mathcal{N}\left(\xi - X, (1 + |\xi|)^2\right),$$

$$Y = \mathrm{clip}_{[0.01, 0.99]}(\mathrm{sigmoid}(\eta)).$$

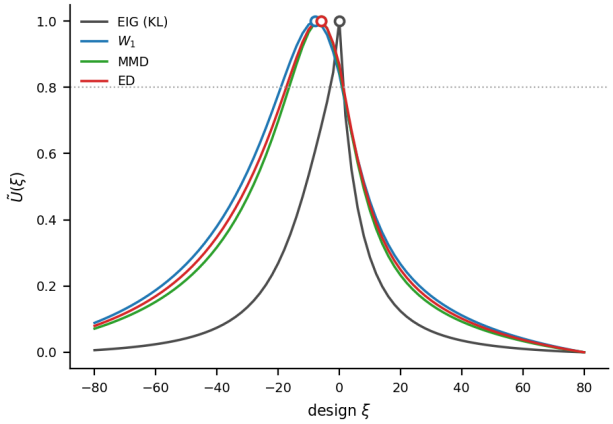
Noise increases with $|\xi|$; clipping creates two endpoint atoms.

Design tradeoff: probe the preference threshold while avoiding high observation noise and response saturation.

- 81 candidate designs: $\xi \in \{-80, -78, \ldots, 80\}$.
- Deterministic reference utilities; fixed MMD bandwidth $h_0 = 19.077$.
- Evaluate both **objective shape** and **independent posterior quality**.

Wu, Liang & Yang: Section 5.2, model formulation.

IPM objectives have broader near-optimal regions here



Compare shapes on a common scale:

$$\widetilde{U}(\xi) = \frac{U(\xi) - U_{\min}}{U_{\max} - U_{\min}}.$$

Utility	ξ^*	# designs with $\widetilde{U} \geq 0.8$
EIG	0	2
W_1	−8	10
MMD	−6	9
ED	−6	9

Broader regions give more tolerance to coarse candidate grids in this example.

Wu, Liang & Yang: Section 5.2, reference landscapes; normalization compares shapes, not absolute utility values.

Selected designs are competitive under external criteria

Evaluate posterior quality by deterministic quadrature. **Lower is better.**

Selector	ξ^*	Bayes RMSE	L^1 risk	90% interval width	IQR
EIG	0	13.5395	10.1944	**40.7203**	18.0399
W_1	-8	13.3338	10.3995	42.9604	17.5325
MMD / ED	-6	**13.0715**	**10.1093**	41.5229	**17.1762**
Prior only	—	20.0000	15.9577	65.7941	26.9796

- MMD/ED give the smallest RMSE, absolute-error risk, and IQR *among these selected designs*.
- EIG gives the narrowest expected central 90% interval.

Different utilities express different priorities; no criterion wins on every measure.

Wu, Liang & Yang: Section 5.2, posterior-quality table; interval width and IQR are predictive expectations.

64-dimensional benchmark: estimator assumptions matter

$X \sim \mathcal{N}(0, I_{64}), \quad Y \mid X, \xi \sim \mathcal{N}(\xi X, 0.1^2 I_{64}); \quad N = 64$ outer draws, $M = L = 512$.

Utility	Estimator	$\xi = 0.5$	$\xi = 2$	$\xi = 6$
EIG	PCE	94.01%	96.75%	97.62%
EIG	Fitted ACE	0.23%	0.91%	0.11%
EIG	Gaussian plug-in KL	0.70%	0.43%	0.01%
W_1	Empirical assignment	8.22%	1.27%	0.12%
MMD	V-statistic	0.97%	0.57%	0.50%
ED	V-statistic	0.80%	0.35%	0.30%
W_2^2	OT-ICNN (extension)	1.17%	1.12%	0.21%

Absolute relative errors against each method's **own utility reference**; no runtime comparison.

- PCE saturates at $\log(513) = 6.2403$ nats; exact EIG is 104.26–262.05.
- ACE and plug-in KL are accurate with effective proposals / a correct density family.
- W_2^2 demonstrates modularity; it lies outside the IPM stability theory.

Wu, Liang & Yang: Section 5.3; PCE/ACE: Foster et al. (2020); OT-ICNN: Makkuva et al. (2020).

Sample-based evaluation has explicit information costs

$$\widehat{U}_{\mathcal{F}}(\xi) = \frac{1}{N}\sum_{i=1}^{N} \widehat{\gamma}_{\mathcal{F}}\left(\{X_m\}_{m=1}^{M}, \{X_m^{(i)}\}_{m=1}^{M}\right), \quad Y_i \sim m_\xi, \quad X_m^{(i)} \sim \mu_\xi^{Y_i}.$$

Estimator	Original densities	Posterior samples	Density / proposal fit
PCE	Likelihood	No	No
Fitted ACE	Likelihood + prior	Yes	Gaussian proposal
Gaussian plug-in KL	None	Yes	Prior + posterior
Sample-based IPMs	None	Yes	None

- Density access above concerns **utility evaluation**; obtaining posterior samples can still be expensive.
- ACE proposal-scale errors yield 5–58% underestimation in sensitivity tests.
- A nonlinear reparameterization preserves EIG, but Gaussian plug-in errors reach 21–41% in the strongest tested case.

Wu, Liang & Yang: Section 5.3 and Appendix D, estimator requirements and sensitivity tests.

Takeaways: choose the utility to match the task

1. **A geometry-aware alternative.**
   IPMs can provide finite, useful design criteria when EIG is infinite.

2. **Explicit stability conditions.**
   Likelihood error is weighted by geometry; prior error requires compatibility.

3. **Separate three questions.**
   Assess model sensitivity, estimator accuracy, and downstream design quality.

Select the metric using the inference target, parameter geometry, and available computational access.

Thank you! dwuwd@umd.edu

Backup: the likelihood proof avoids centering the class

Write $\ell_y(x) = p(y \mid x)$ and $\mu(f) = \int f \, d\mu$.

$$U_{\mathcal{F}} = \int \sup_{f \in \mathcal{F}} |A_\mu(f, y)| \, dy, \qquad A_\mu(f, y) = \mu(f\ell_y) - \mu(f)\mu(\ell_y).$$

The predictive weight cancels the posterior normalization. For the *same* f,

$$A_\mu(f, y) = \iint [f(x) - f(z)] \ell_y(x) \, \mu(dx)\mu(dz).$$

Therefore,

$$|A_\mu(f, y) - A^q_\mu(f, y)| \le \int \Omega_{\mathcal{F},\mu}(x) |p(y \mid x) - q(y \mid x)| \, \mu(dx).$$

Take the supremum over f and integrate over y.

No assumption that $f - \mu(f)$ or $f - f(x_0)$ remains in $\mathcal{F}$.

Wu, Liang & Yang: Appendix A.

Backup: precise prior-compatibility conditions

Use a fixed class: anchored Lipschitz functions for W_1, a native RKHS ball for MMD, or the anchored distance-induced RKHS ball for normalized ED.

$$e_y = \begin{cases} \ell_y, & \text{MMD}, \\ \ell_y - \ell_y(x_0), & W_1 \text{ or normalized ED}. \end{cases}$$

For integrable $C_E, C_L > 0$, require

$$\frac{e_y}{C_E(y)} \in \mathcal{F}, \qquad \frac{f\ell_y}{C_L(y)} \in \mathcal{F} \quad \text{for every } f \in \mathcal{F}.$$

With $\Delta = \mu - \widetilde{\mu}$ and $m(y) = \mu(\ell_y)$,

$$A_\mu(f, y) - A_{\widetilde{\mu}}(f, y) = \Delta(f\ell_y) - m(y)\Delta(f) - \Delta(\ell_y)\widetilde{\mu}(f).$$

Bound the three terms by $[C_L(y) + m(y) + C_E(y)B_{\mathcal{F}}(\widetilde{\mu})]\gamma_{\mathcal{F}}(\mu, \widetilde{\mu})$.

Wu, Liang & Yang: Section 4.3 and Appendix B; finite $B_{\mathcal{F}}(\widetilde{\mu})$ and finite IPM are also required.

Backup: an interpretable forward-model error bound

Gaussian observation models with common covariance:

$$Y = G(X) + \eta, \qquad Y_* = G_*(X) + \eta, \qquad \eta \sim \mathcal{N}(0, \Sigma), \quad \Sigma \succ 0.$$

Assume $m_2 = \mathbb{E}_\mu \|X - x_0\|^2 < \infty$ and $G - G_* \in L^2_\mu$.

Wasserstein-1 consequence

$$|U_{W_1} - U^*_{W_1}| \le 2\|\Sigma^{-1/2}\|_{\mathrm{op}} \sqrt{m_2}\, \|G - G_*\|_{L^2_\mu}.$$

- Gaussian KL plus Pinsker bounds the likelihood L^1 error.
- $\|\Omega_{W_1,\mu}\|_{L^2_\mu} \le 2\sqrt{m_2}$; apply Cauchy–Schwarz.
- A common, nondegenerate noise covariance is essential to this statement.

Wu, Liang & Yang: Section 4.4 and Appendix C.

Backup: coarse grids and MMD bandwidth

Preference example: restrict to every kth grid point and vary the offset.

Own-utility regret: $\bar{r} = (U(\xi^*) - U(\widehat{\xi}))/(U_{\max} - U_{\min})$.

Mean regret (max)	$k = 2$	$k = 4$	$k = 6$	$k = 8$
EIG	.077 (.155)	.171 (.285)	.233 (.395)	.289 (.465)
ED	.001 (.002)	.012 (.025)	.030 (.066)	.052 (.131)
MMD	.003 (.006)	.015 (.033)	.035 (.077)	.060 (.145)
W_1	.005 (.010)	.014 (.037)	.029 (.075)	.048 (.121)

- PCE accurately recovers the EIG landscape: RMSE 0.0064 nats; reference maximizer recovered in all five runs ($L = 64$, $N = 2000$).
- MMD bandwidth sweep: 33 values, $h/h_0 \in [0.25, 4]$; selected design ranges from -2 to -8.
- The selected design stays -6 at tested bandwidths in $[0.545h_0, 1.297h_0]$.

Wu, Liang & Yang: Section 5.2; regret is normalized within each utility, not a common downstream loss.

Backup: complete utility references and estimates

Utility	Method	$\xi = 0.5$	$\xi = 2$	$\xi = 6$
EIG	Reference	104.2591	191.8068	262.0469
	PCE	6.2403	6.2403	6.2403
	Fitted ACE	104.4950	190.0654	261.7492
	Gaussian plug-in KL	104.9864	190.9730	262.0237
W_1	Reference	10.1043	10.9846	11.1753
	Empirical assignment	10.9346	11.1241	11.1882
MMD	Reference	0.6110	0.6255	0.6264
	V-statistic	0.6169	0.6219	0.6232
ED	Reference	3.0057	3.2686	3.3256
	V-statistic	3.0298	3.2573	3.3157
W_2^2	Reference	102.8971	121.6080	125.8670
	OT-ICNN	104.1002	120.2499	125.6078

EIG, W_1, W_2^2: analytic references. MMD, ED: deterministic quadrature.
All reference utilities prefer $6 > 2 > 0.5$; every reported estimator except PCE recovers this ordering.

Wu, Liang & Yang: Section 5.3.

Backup: estimator sensitivity at fixed sample budgets

ACE: change proposal scale

$$q_{y,c} = \mathcal{N}(\widehat{b}_y, c^2 \widehat{C}_y).$$

c	Relative underestimation
0.5	23.97–58.30%
2	5.20–12.49%

An effective posterior proposal matters at fixed $L = 512$.

Plug-in KL: change coordinates

$$Z_{2j-1} = X_{2j-1}, \qquad Z_{2j} = X_{2j} + a(X_{2j-1}^2 - 1).$$

Invertible transformation: exact EIG is unchanged.

Fit Gaussians in Z coordinates. At $a = 4$, relative overestimation is 21.14–40.63%.

Accurate EIG estimation is possible; its approximation assumptions must be checked.

Wu, Liang & Yang: Section 5.3 and Appendix D; ranges are across the three designs.