On the Misinformation in a Statistical Experiment

Jake Callahan
The University of Arizona

Tommie Catanach
Sandia National Laboratories

Abstract

The principle that more informative experiments are always better is a cornerstone of Bayesian experimental design. This principle assumes the practitioner's model and inference are correct. In practice, both the data-generating model and the inferential approximation are inevitably misspecified, and we show that under these conditions the classical framework for comparing experiments breaks down. Designs ranked as most informative can become actively harmful, amplifying bias to produce confident but incorrect inferences. We demonstrate that the commonly-accepted axioms of experimental utility, such as Blackwell monotonicity, fail under misspecification, and that information measures proposed to handle it, like the Expected Generalized Information Gain (EGIG), do not obey these axioms. To resolve this, we propose a generalized axiomatic framework for robust Bayesian experimental design. We prove that EGIG satisfies our axioms as a criterion that penalizes inferential error, providing a principled foundation for its use in Bayesian experimental design. As a complementary approach, we derive a new measure that instead penalizes model error. Finally, we demonstrate our framework's utility across common modes of misspecification, showing it provides a reliable guide for experimental design where classical methods fail.

1 INTRODUCTION

While George Box noted that "all models are wrong," optimal experimental design (OED) theory typically assumes the practitioner's model and inference are correct. Within the classical framework, more data are always beneficial. This principle is formalized through Blackwell's theory of experiments (Blackwell, 1951, 1953), establishing that "more informative" experiments always improve decision-making. Building on this, Ginebra (2007) established criteria any valid information measure must satisfy, including Blackwell monotonicity: finer experiments must yield higher utility than coarser ones.

Recent work has begun addressing these limitations. Duersch and Catanach (2020) established theoretical foundations for robust information measures accounting for evolving belief, while Catanach and Das (2023) demonstrated the practical utility of Expected Generalized Information Gain (EGIG) in Bayesian optimal experimental design (BOED). However, these advances lack a principled axiomatic framework, mirroring the criteria of Ginebra (2007), justifying why robust measures like EGIG are preferable under misspecification and whether more general classes of robust information measures can be formed. Our work addresses this gap.

In practice, the practitioner's inferential procedure is often misspecified, and the classical framework for comparing experiments can break down. When the model or inference algorithm is flawed, precise data can amplify bias, producing posteriors that are confident but wrong. Experiments that classical theory ranks as most informative can become actively misleading. This arises in many applications, such as experiments modeled with simplified physics that are inaccurate under extreme conditions, where BOED often samples near boundaries without knowing the model fails there. This raises our central question: *what theoretical framework can motivate robust experimental design criteria?*

We diagnose the theoretical failure of the classical framework and build from it a new axiomatic foundation for robust design. We show the breakdown results from a quantifiable penalty (the *misspecification gap*) between the true information an experiment provides and the value perceived by the practi-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

tioner's flawed procedure. This gap invalidates Blackwell monotonicity, explaining why a practitioner might prefer a strictly worse experiment. This failure occurs when the practitioner's full procedure (statistical model and inference algorithm) diverges from the true data-generating process and exact Bayesian update.

After developing this axiomatic foundation, we show EGIG satisfies these axioms, providing principled justification for its use in Bayesian OED as the robust analogue to EIG. We provide examples encompassing structural, model, and inferential misspecification to demonstrate EGIG's efficacy across these failure modes, going beyond simple likelihood misspecification for linear dynamical system models considered in Catanach and Das (2023). Finally, we derive a general class of robust information measures distinct from EGIG that also satisfy the axioms, opening new avenues for research into other robust BOED metrics.

2 THE FAILURE OF CLASSICAL BOED UNDER MISSPECIFICATION

Classical experimental design assumes a well-specified model, which enables a rigorous framework for valuing information, but it fails under misspecification. In this section, we review the classical framework, demonstrate its failure with an example, and diagnose the underlying theoretical breakdown. Finally, we review Expected Generalized Information Gain and demonstrate why it does not fit into the classical framework.

2.1 The BOED Framework

In Bayesian optimal experimental design (BOED) (Lindley, 1956), initial uncertainty about unknown parameters θ is encoded in a prior $p(\theta)$. We configure experiments through a design choice $d \in \mathcal{D}$, which influences the data-generating process via the likelihood $p(y \mid \theta, d)$. After observing data y, Bayes' rule updates our beliefs to the posterior $p(\theta \mid y, d)$. Standard BOED maximizes the Expected Information Gain (EIG), defined as the expectation over potential data of the KL divergence between posterior and prior:

$$\begin{aligned} \mathrm{EIG}(d) &= \mathbb{E}_{p(y|d)}\left[\mathrm{D}_{\mathrm{KL}}(p(\theta \mid y, d) \,||\, p(\theta))\right] \\ &= \mathbb{E}_{p(y,\theta|d)}\left[\log \frac{p(\theta \mid y, d)}{p(\theta)}\right]. \end{aligned} \tag{1}$$

BOED seeks the design that maximizes this utility:

$$d^* = \arg\max_{d\in\mathcal{D}} \mathrm{EIG}(d). \tag{2}$$

2.2 The Information in an Experiment

The choice of information-based utilities like EIG is motivated by Blackwell's formal framework for comparing experiments (Blackwell, 1951, 1953). An experiment d_1 is *sufficient for* (or "more informative than") d_2 if data from d_2 can be simulated by applying a stochastic transformation to data from d_1; that is, if $y_2 \sim K(y_2 \mid y_1)$ for some transition kernel K, where y_1 is from experiment d_1. This defines the Blackwell ordering of experiments by informativeness.

Building on this, Ginebra (2007) proposed that any valid information measure $U(d)$ must satisfy: **(G1)** $U(d)$ is a finite real number; **(G2)** If $Y \perp \theta$, then $U(d) = 0$; and **(G3)** If d_1 is sufficient for d_2, then $U(d_1) \geq U(d_2)$ (Blackwell monotonicity).

The Blackwell-Sherman-Stein theorem (Ginebra, 2007) connects this abstract ordering to practical utilities: d_1 is sufficient for d_2 if and only if, for any prior $p(\theta)$ and convex function φ on the space of posteriors,

$$\mathbb{E}_{p(y_1|d_1)}[\varphi(p(\theta \mid y_1, d_1))] \geq \mathbb{E}_{p(y_2|d_2)}[\varphi(p(\theta \mid y_2, d_2))]. \tag{3}$$

Thus, more informative experiments generate posteriors that are more separated in the sense of convex ordering (proof in Appendix C.1).

Classical Information Measures. These axioms motivate a general class of information measures

$$I_\varphi(d) = \mathbb{E}_{p(y|d)}[\varphi(p(\theta \mid y, d))], \tag{4}$$

where φ is convex with $\varphi(p(\theta)) = 0$. EIG corresponds to φ as the KL divergence. This framework's success assumes the practitioner's procedure accurately reflects reality. When this fails, information measures following these axioms may not provide useful guidance.

2.3 A Practical Example of Failure

Consider inferring the parameters of a forced oscillator. The true system follows nonlinear dynamics:

$$\ddot{x} + \mu|\dot{x}|\dot{x} + \alpha x + \beta x^3 = \gamma \sin(\omega t). \tag{5}$$

However, the practitioner employs a simpler working model that omits the cubic nonlinear term:

$$\ddot{x} + \mu|\dot{x}|\dot{x} + \alpha x = \gamma \sin(\omega t). \tag{6}$$

Here x denotes the displacement of the mass as a function of time, the coefficient β controls the strength of the cubic restoring force, while γ controls the amplitude of the external forcing. The experimental design

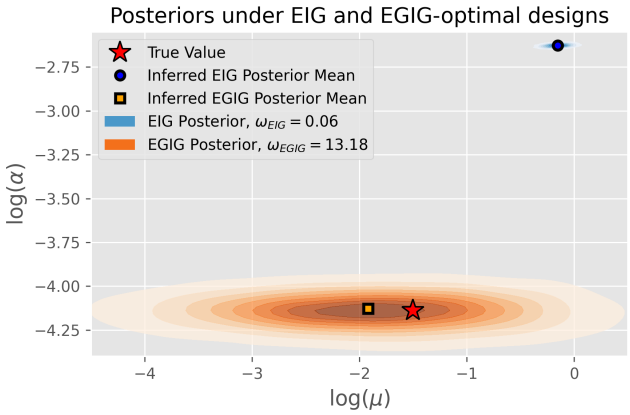

Figure 1: Failure of classical EIG under model misspecification. EIG optimization under the working model selects $\omega = 0.06$. When data from the true nonlinear system is analyzed using the working model at this design point, the resulting posterior (blue contours) is precise but catastrophically biased, entirely missing the true parameter value (red star). EGIG optimization selects $\omega = 13.18$. When data from the true nonlinear system is analyzed using the working model at this design point, the resulting posterior (orange contours) is less tightly concentrated than that generated using ω_{EIG}, but it exhibits minimal bias.

problem is to choose the driving frequency $d = \omega$ to best infer the parameters $\theta = (\mu, \alpha)$.

Following classical BOED, the practitioner maximizes EIG under their working model. Since the working model is most sensitive to its parameters at lower frequencies, EIG optimization selects $\omega_{\text{EIG}} = 0.06$. Figure 1 shows the EIG surface and the resulting posterior when data generated from the true nonlinear model is analyzed using the working model with $\omega = \omega_{\text{EIG}}$. The resulting posterior is unimodal and concentrated, suggesting the experiment was highly informative, yet it completely misses the true parameter value. The classical framework has guided the experimenter to a design that makes them confidently wrong.

2.4 Expected Generalized Information Gain

To address such failures, Catanach and Das (2023) proposed Expected Generalized Information Gain (EGIG) based on the ideas of information as rational belief presented in Duersch and Catanach (2020). Rather than using the working model's expected data distribution, EGIG takes expectation with respect to $\mathcal{M}^\star$:

$$\text{EGIG}(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(\theta, y \mid d)} \left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p(\theta)} \right] \tag{7}$$

where $p^\star_{\mathcal{M}}(\theta, y \mid d)$ represents the true joint distribution and $p_{\mathcal{M}}(\theta \mid y, d)$ is the working posterior.

EGIG evaluates the working posterior against data that will actually be generated, revealing when designs produce confident but incorrect inferences. Catanach and Das (2023) demonstrated that EGIG-based designs avoid the pathological behavior seen in our spring example, maintaining posteriors closer to truth under misspecification.

The computation of EGIG requires sampling from the true data-generating process $\mathcal{M}^\star$. For the spring-mass problem, we assume the true nonlinear dynamics are known for design selection. This situation often occurs when surrogate or reduced-order models are used for inference but a high-fidelity model is available. When sampling from $\mathcal{M}^\star$ is not feasible, a natural extension is to marginalize over plausible candidate models.

Using EGIG instead of EIG to select the optimal design yields $\omega_{EGIG} = 13.18$. Figure 1 displays the posterior from analyzing data sampled from the true model (with forcing amplitude $\gamma = 0.1$ and cubic coefficient β drawn from its prior) using the working model with $\omega = \omega_{\text{EGIG}}$. This posterior is less concentrated than the EIG-based design but centered near the true θ value. Thus, EGIG finds designs yielding reasonably accurate posteriors despite misspecification.

Despite its empirical success, EGIG is incompatible with the Ginebra framework. First, classical measures (Equation 4) are guaranteed non-negative, whereas EGIG, as an expectation over the true data-generating process, can be negative. This is a key feature: a negative value signals an experiment is actively misleading, where the prior is more accurate than the inferred posterior.

Moreover, EGIG can violate axiom (G3), *Blackwell monotonicity*, which requires that a more informative experiment always have higher utility. Under misspecification, EGIG may prefer a less informative experiment because more precise data from a sufficient one can cause a flawed procedure to become overconfident in the wrong parameter region. EGIG thus correctly penalizes designs that amplify bias, reducing it at the cost of a less concentrated posterior (Figure 1). Therefore, while empirically successful, EGIG's theoretical foundations lie outside classical experimental design theory.

3 A MORE GENERAL FRAMEWORK FOR ROBUST BOED

The classical framework is unsuitable for misspecified settings because it cannot describe the necessary trade-off between information gain and the risk of amplifying bias. A robust theory for BOED thus requires a new axiomatic foundation to manage this trade-off, which we propose here. We begin by diagnosing the precise mechanism of the classical framework's failure, which will then motivate a new set of axioms and a general class of robust utilities.

3.1 Theoretical Diagnosis: The Misspecification Gap

EGIG's success stems from using the true data-generating process in its expectation, a natural principle since experimental outcomes depend on this true process, not the working model. We investigate whether importing this principle into the classical measures from Equation (4) can remedy the pathological behavior.

We distinguish the true data-generating process $\mathcal{M}^\star$, which yields the true posterior $p_{\mathcal{M}^\star}(\theta \mid y, d)$ via exact Bayesian inference, from the practitioner's full procedure: a working model $\mathcal{M}$ (which may have incorrect likelihood or prior) combined with an inference algorithm (which may be approximate, e.g., variational inference). This produces the working posterior $p_{\mathcal{M}}(\theta \mid y, d)$, potentially misspecified due to model error, inferential error, or both.

We define a *misspecified* version of the general utility from Equation 4, termed the misspecified φ-information:

$$I_\varphi^{mis}(d) := \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}[\varphi(p_{\mathcal{M}}(\theta \mid y, d))]\,. \tag{8}$$

While seemingly sensible, this breaks the classical framework. The following proposition reveals it decomposes into true information value and a penalty term.

Proposition 3.1 (Misspecification Gap Decomposition)**.** *Let φ be convex on the space of probability densities over Θ with $\varphi(p(\theta)) = 0$. For an experiment where data is generated by $\mathcal{M}^*$ but inference assumes $\mathcal{M}$, the misspecified φ-information decomposes as*

$$I_\varphi^{mis}(d) = I_\varphi^{true}(d) - \Delta_\varphi(d), \tag{9}$$

where

$$I_\varphi^{true}(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}[\varphi(p_{\mathcal{M}^\star}(\theta \mid y, d))]\,, \tag{10}$$

and

$$\Delta_\varphi(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}[\varphi(p_{\mathcal{M}^\star}(\theta \mid y,d)) - \varphi(p_{\mathcal{M}}(\theta \mid y,d))] \tag{11}$$

is the misspecification gap.

Proof. The result follows directly from the definition of $I_\varphi^{mis}(d)$ by adding and subtracting $\mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}[\varphi(p_{\mathcal{M}^\star}(\theta \mid y, d))]$. □

This decomposition shows the perceived utility splits into true information $I_\varphi^{true}(d)$ (depending only on the true process) and a penalty $\Delta_\varphi(d)$ (quantifying distortion from the working model). When $\mathcal{M} = \mathcal{M}^*$, the gap vanishes and Equation (4) is recovered.

3.1.1 Consequences for the Ginebra Axioms

This decomposition reveals why classical axioms fail. **(G2)** is violated: for a truly non-informative experiment where $Y \perp \theta$ under $\mathcal{M}^\star$, true information is zero but a misspecified procedure may still update on noise, yielding $I_\varphi^{mis}(d_0) = -\Delta_\varphi(d_0) < 0$.

This breaks **(G3)**: a practitioner may prefer the non-informative d_0 over a more informative d_1. This occurs when the weak signal in d_1 misleads the procedure into a confident but wrong posterior, creating a large penalty $\Delta_\varphi(d_1)$ that exceeds its true information gain, making the useless experiment appear superior. Appendix B.1 provides a concrete example.

3.2 Desiderata for Robust Information Measures

We propose that any robust information measure $U(d)$ for experimental design under misspecification should satisfy the following criteria.

(R1) Real-Valued: $U(d) \in \mathbb{R}$ is finite $\forall d \in \mathcal{D}$.

(R2) Non-Informative: If $Y \perp \theta$ under the true model $\mathcal{M}^*$, then $U(d) \leq 0$, with equality if and only if the working posterior equals the prior almost everywhere.

(R3) Information-Penalty Structure: The utility admits a decomposition

$$U(d) = U^{\text{true}}(d) - \Delta_U(d)$$

where $U^{\text{true}}(d) \geq 0$ measures true classical information gain satisfying the axioms (G1)–(G3), and $\Delta_U(d) \geq 0$ quantifies a misspecification penalty. This decomposition implies the utility is bounded by its components:

$$-\Delta_U(d) \leq U(d) \leq U^{\text{true}}(d).$$

(R4) Bounded Contraction Under Sufficiency: Suppose design d_1 is sufficient for design d_2 (i.e., the outcome under d_2 comes from applying a transition kernel to the outcome under d_1). Then:

- True information contracts under sufficiency:

$$U^{\text{true}}(d_2) \leq U^{\text{true}}(d_1).$$

- If d_2 is sufficient for d_1 under both the working and true models via a *common* transition kernel $K(y_2 \mid y_1)$, then the penalty also contracts:

$$\Delta_U(d_2) \leq \Delta_U(d_1).$$

These criteria adapt the classical Ginebra axioms to address the failures demonstrated in Section 2.

Axiom (R2) directly addresses the failure of non-informativeness. Since a misspecified model may still update its beliefs on uninformative data, the allowance for a negative utility reflects that this apparent information is an artifact of misspecification.

The decomposition (R3) requires that any robust utility be the difference between a classical information measure (U^{true}) and a misspecification penalty (Δ_U). This structure allows the cost of misspecification to be explicitly quantified and mandates that any compliant utility must implicitly balance the pursuit of classical information against the risk of amplifying error.

Axiom (R4) replaces Blackwell monotonicity with a contraction principle. Under misspecification, finer experiments can amplify bias in the working inferential procedure, as in Section 2.3, so monotonicity cannot hold. Contraction instead means that moving to a coarser (less sufficient) experiment reduces both the true information and, under mild assumptions, the misspecification penalty. The admissible utility range $[-\Delta_U, U^{\text{true}}]$ therefore shrinks toward zero. This ensures that while coarser experiments are less informative, they are also less risky, as they cannot amplify bias to inflict a large penalty.

4 MEASURES OF ROBUST INFORMATION

We now consider several measures of information that satisfy (R1)–(R4). We begin with our primary focus, Expected Generalized Information Gain, and show that it is a natural consequence of our framework.

4.1 Expected Generalized Information Gain

We first formally state that EGIG is a valid measure within our framework.

Proposition 4.1. *Expected Generalized Information Gain (EGIG) satisfies the robust axioms (R1–R4), and admits a decomposition*

$$\text{EGIG}(d) = \text{EIG}^{true}(d) - \Delta_{\text{EGIG}}(d),$$

where $EIG^{true}(d)$ is the EIG computed with the true model and

$$\Delta_{\text{EGIG}}(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}\left[D_{KL}(p_{\mathcal{M}^\star}(\theta \mid y, d) \;||\; p_{\mathcal{M}}(\theta \mid y, d))\right]. \tag{12}$$

Proof. See Appendix C.2.

The justification for EGIG's form comes from correcting the failure of simpler approaches identified in Section 2. The naive misspecified φ-information (I^{mis}_{φ}) fails axiom (R2) because it scores the working posterior in isolation, rewarding spurious learning. EGIG corrects this by instead evaluating the practitioner's belief update from the perspective of the true posterior, which motivates its form as defined in Equation (7). The non-negative penalty $\Delta_{\text{EGIG}}(d)$ is the expected KL divergence between the true and working posteriors, so EGIG is therefore fundamentally concerned with the quality of the final inference, rewarding designs that produce an accurate final answer.

4.2 Other Robust Information Measures

While EGIG motivates our framework, it is interesting to consider other valid robust information measures. We construct a general class of utilities using the family of *f-divergences* (Ali and Silvey, 1966), defined as

$$D_f(p \,||\, q) = \int q(x) f(p(x)/q(x)) dx \tag{13}$$

for a convex function f with $f(1) = 0$. Different f recover familiar divergences, such as the KL divergence ($f(t) = t \log t$) or the χ^2-divergence ($f(t) = (t-1)^2$).

This leads to the following definition, which constructs a robust utility by starting with a measure of true information (the f-mutual information) and subtracting a penalty for misspecification.

Definition 4.2 (Robust f-Information Measure)**.** *Let f be a convex function with $f(1) = 0$. The* Robust f-Information Measure (RFIM) *is*

$$\begin{aligned} I^\star_f(d) := D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \;||\; p(\theta) p_{\mathcal{M}^\star}(y \mid d)) \\ - D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \;||\; p_{\mathcal{M}}(\theta, y \mid d)). \end{aligned} \tag{14}$$

By construction, this satisfies (R1)–(R4):

Proposition 4.3 (Properties of Robust f-Information)**.** *Let $f : [0, \infty) \to \mathbb{R}$ be convex*

and non-affine, with $f(1) = 0$. *Then the Robust* f*-Information Measure satisfies (R1)–(R4) with*

$$I_f^\star(d) = I_f^{\text{true}}(d) - \Delta_f(d),$$

where

$$I_f^{\text{true}}(d) = D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \;||\; p(\theta)p_{\mathcal{M}^\star}(y \mid d)), \quad (15)$$

$$\Delta_f(d) = D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \;||\; p_{\mathcal{M}}(\theta, y \mid d)). \quad (16)$$

Proof Sketch. Properties (R1)–(R3) follow directly from the definition and properties of f-divergences. For (R4), the bounds contract via the data processing inequality. See Appendix C.3 for a full proof.

If we instantiate the RFIM with the Kullback-Leibler divergence (i.e., $f = x \log x$), the penalty term becomes the KL divergence between the joint distributions and it can be interpreted by decomposing the penalty with respect to θ:

$$\begin{aligned}\Delta_{KL}(d) &= D_{KL}(p_{\mathcal{M}^\star}(\theta, y \mid d) \;||\; p_{\mathcal{M}}(\theta, y \mid d)) \quad (17)\\ &= \mathbb{E}_{p(\theta)}\left[D_{KL}(p_{\mathcal{M}^\star}(y \mid \theta, d) \;||\; p_{\mathcal{M}}(y \mid \theta, d))\right].\end{aligned}$$

This views the penalty as the expected divergence between the true and working *likelihoods* (assuming $p_{\mathcal{M}^\star}(\theta) = p_{\mathcal{M}}(\theta)$; otherwise, a prior-mismatch term $D_{KL}(p_{\mathcal{M}^\star}(\theta)\,||\,p_{\mathcal{M}}(\theta))$ is added). It focuses on the error in the model's generative components, averaged over the prior. Therefore, $I_{\text{KL}}^\star$ becomes

$$\begin{aligned}I_{\text{KL}}^\star(d) &= EIG^{true}(d) \quad (18)\\ &\quad - \mathbb{E}_{p(\theta)}\left[D_{KL}(p_{\mathcal{M}^\star}(y \mid \theta, d) \;||\; p_{\mathcal{M}}(y \mid \theta, d))\right].\end{aligned}$$

$I_{\text{KL}}^\star(d)$ highlights penalizing the model's components (the likelihoods, as in $I_{KL}^\star$) versus penalizing the final inferential product (i.e., the posteriors, as in the EGIG penalty Δ_{EGIG}). For our experimental evaluation, we focus on EGIG, as its penalty on the final posterior aligns most directly with the practitioner's ultimate goal of achieving an accurate final answer. The rigorous comparison of these different philosophies is a rich area for future work.

We note that there is a strict ordering between EGIG and $I_{\text{KL}}^\star$.

Lemma 4.4 (Ordering of Robust Information Measures)**.** *For any design* $d \in \mathcal{D}$*, let* EGIG(d) *be the Expected Generalized Information Gain and* $I_{\text{KL}}^\star(d)$ *be the Robust* f*-Information Measure instantiated with the Kullback-Leibler divergence. These measures satisfy the following relationship:*

$$\begin{aligned}\text{EGIG}(d) =& I_{\text{KL}}^\star(d)\\ &+ \mathbb{E}_{p_{\mathcal{M}^*}(y|d)}\left[D_{KL}(p_{\mathcal{M}^*}(y|d) \parallel p_{\mathcal{M}}(y|d))\right].\end{aligned} \quad (19)$$

Consequently, EGIG$(d) \geq I_{\text{KL}}^\star(d)$*, with equality if and only if the working marginal data distribution matches the true marginal distribution almost everywhere.*

Proof. See Appendix C.4. This relationship implies that $I_{\text{KL}}^\star$ is a strictly more risk-averse criterion than EGIG, as it requires the model to be accurate not just in its marginals, but in its full joint distribution.

5 RELATED WORK

Modern scalable BOED relies on approximations that introduce misspecification, including variational methods (Foster et al., 2020; Dong et al., 2025), amortized policies (Foster et al., 2021), neural estimators (Kleinegesse and Gutmann, 2020), and surrogate models (Lei et al., 2021). Such approximate inference can cause BOED to fail (Rainforth et al., 2023; Kennedy and O'Hagan, 2001; Brynjarsdóttir and O'Hagan, 2014), a problem amplified by active learning's sampling bias (Sloman et al., 2022) that has motivated stability analysis of error propagation (Duong et al., 2023). A related decision-theoretic approach (Overstall and McGree, 2022) distinguishes a "Fitted Model" ($\mathcal{M}$) from a "Designer Model" ($\mathcal{M}^\star$) to evaluate an "External Expected Loss":

$$L(d) = \mathbb{E}_{p_{\mathcal{M}^\star(\theta,y|d)}}[\lambda(p_{\mathcal{M}}(\theta \mid y, d))], \quad (20)$$

for a loss function λ chosen to encode the experimental aim. While similar to our framework in using separate true and working models, this loss is not measured against a baseline like the prior's utility. Thus, it can compare experiments, but its value does not admit a meaningful interpretation on its own. Hence, $L(d)$ quantifies expected loss but cannot detect actively harmful experiments where posteriors are worse than priors, which yield negative information in our framework. Although their formulation can be adapted to recover EGIG When $\lambda = \log$, it cannot express all robust measures such as $I_{\text{KL}}^\star$.

Alternative approaches augment models with nonparametric terms (e.g., Gaussian Processes) to capture structural discrepancy (Feng, 2015; Tsirpitzi et al., 2023; Petsagkourakis and Galvanin, 2021); frame the issue as covariate shiftand apply importance-weighted estimators (Sugiyama, 2005; Tang et al., 2025; Ali et al., 2014); or use BOED to actively detect misspecification (Catanach and Das, 2023; Forster et al., 2025). Minimax frameworks address model uncertainty by optimizing designs against a worst-case "ambiguity set" of plausible model components, stabilizing criteria or posterior covariance (Go and Isaac, 2022; Attia et al., 2025).

6 EXPERIMENTAL EVALUATION

The following examples demonstrate how EGIG handles various forms of misspecification. Prior work

(Catanach and Das, 2023) focused on perturbed linear dynamical systems; here we extend validation to inference errors, surrogate-model approximations, and incorrect non-linear dynamics. Our examples assume the ability to sample from the true data-generating model $\mathcal{M}^\star$, which is practical in settings where a high-fidelity simulator provides $\mathcal{M}^\star$ or when exact inference is intractable but sampling remains feasible. In EGIG, the high-fidelity model simulates data, while the tractable working model is used for analysis. While we use a single known $\mathcal{M}^\star$, the framework naturally extends to model uncertainty by averaging utility over candidate models or seeking minimax designs.

6.1 Spring–Mass Dynamics

Figure 2 shows utility curves for EIG, EGIG, and $I^\star_{KL}$ for the misspecified spring-mass model (Section 2.3), explaining why the EIG-optimal design yields a biased posterior while the EGIG-optimal design is accurate (Figure 1). Curves were estimated via Nested Monte Carlo (Ryan, 2003) (Appendix A).

EIG peaks at low frequency ($\omega_{\text{EIG}} = 0.06$), a region the robust criteria penalize due to severe model mismatch. Both EGIG and $I^\star_{KL}$ find optimal designs at higher frequency ($\omega_{\text{EGIG}} = \omega_{I^\star_{KL}} = 13.18$), though the shared optimum may reflect design space discretization. As expected, $I^\star_{KL}$ is consistently lower than EGIG because it penalizes likelihood rather than posterior discrepancies. This demonstrates how robust criteria trade perceived information for robustness, avoiding designs that amplify model bias.

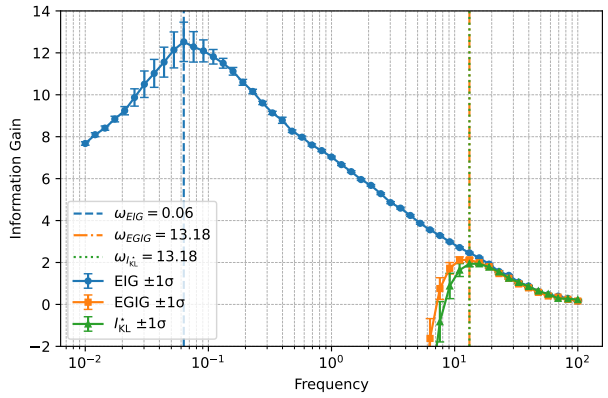

Figure 2: Utility curves for EIG, EGIG, and $I^\star_{KL}$ as a function of design frequency ω. Classical EIG (blue), computed under the misspecified model, peaks at low frequency. Both robust criteria, evaluated against the true model, penalize this region and find a more robust, higher-frequency optimum.

6.2 Procedural Misspecification with Variational Inference

We next handle procedural misspecification, where the model is correct but inference is approximate. The goal is to infer the $D = 6$ weights θ of a Bayesian polynomial regression model by choosing $n = 4$ data locations $d = \{d_1, \ldots, d_4\} \in [-1, 1]^4$. The true forward model is $y = \Phi(d)\theta + \epsilon$, where $\Phi(d)$ is the $n \times D$ Vandermonde matrix, $\theta \sim \mathcal{N}(0, \sigma_p^2 I)$ with $\sigma_p = 1$, and $\epsilon \sim \mathcal{N}(0, \sigma_y^2 I)$ with $\sigma_y = 0.1$. The practitioner knows this true model $\mathcal{M}^\star$, which yields a Gaussian posterior with dense covariance, but is constrained to use amortized mean-field variational inference (VI) for analysis. This involves offline training of a neural network mapping any dataset to a factorized Gaussian posterior $q_\phi(\theta \mid y, d) = \prod_{i=1}^{D} \mathcal{N}(\theta_i \mid \mu_i(y, d), \sigma_i^2(y, d))$, used for rapid online inference. The factorized q_ϕ cannot capture the true posterior's correlations, providing the misspecification.

We compare two strategies: the classical approach optimizes true EIG to find d_{EIG}, ignoring the known inferential limitations; the robust approach optimizes EGIG using q_ϕ as the working posterior to find d_{EGIG}.

6.2.1 Design Optimization for Amortized VI

To find the robust design, we maximize the EGIG objective depending on the amortized VI output:

$$\mathcal{U}(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(\theta, y \mid d)} \left[\log q(\theta \mid y, d) - \log p(\theta) \right].$$

Generally, this requires jointly optimizing design d and network parameters ϕ using the reparameterization trick (Kingma and Welling, 2013) with stochastic gradient ascent. For our Bayesian linear regression, however, the optimal mean-field posterior can be computed analytically via moment matching. We use this analytical solution to represent perfectly trained amortized VI, isolating the design choice's impact on mean-field approximation quality, separate from numerical optimization artifacts.

This objective equals the variational lower bound on EIG, $\mathcal{L}_{post}$, from Foster et al. (2020). While they frame this as a tractable approximation to classical EIG, our framework provides a new interpretation: this is the exact utility for robust design under procedural misspecification. One is not approximating an ideal problem but solving the practical problem where amortized VI is the known misspecification source.

6.2.2 Results

The classical d_{EIG} design creates a worst-case scenario for mean-field VI, as shown in Figure 3. While the approximate posterior mean is unbiased, its credible

interval is severely overestimated, failing to capture the true posterior's precision. The robust d_{EGIG} design yields a VI posterior that nearly perfectly matches the true posterior with well-calibrated uncertainty.

Figure 4, which shows marginal distributions for two of the six regression parameters, reveals the underlying cause. For the EIG design (left), the true posterior is highly correlated, which the axis-aligned MF-VI approximation fundamentally cannot capture. The EGIG design (right) yields a less dependent posterior that MF-VI captures accurately. The complete set of all fifteen 2D marginal posteriors is provided in Appendix B.2.

The misspecification gap confirms this: $\Delta_{\text{KL}}(d_{\text{EIG}}) = 1779.71$ nats versus $\Delta_{\text{KL}}(d_{\text{EGIG}}) = 3.05$ nats, quantifying the substantial improvement.

We additionally optimized the design vector using the $I^{\star}_{\text{KL}}$ criterion. The resulting optimal designs are nearly identical: $d_{\text{EGIG}} = [-0.4170, -0.1569, 0.1572, 0.4171]$ and $d_{I^{\star}_{\text{KL}}} = [-0.4174, -0.1557, 0.1557, 0.4174]$. To understand why these criteria converge here, consider the decomposition from Lemma 4.4. In this experiment, the working model uses the true generative parameters but employs an approximate mean-field inference method. Because the generative process is perfectly specified, the marginal data distributions match exactly ($p_{\mathcal{M}}(y|d) = p_{\mathcal{M}^*}(y|d)$), causing the marginal KL divergence penalty to vanish.

Thus, in settings with purely inferential error, EGIG and $I^{\star}_{\text{KL}}$ are mathematically equivalent, and both simplify to the true EIG minus the expected posterior KL divergence. Consequently, both metrics penalize designs that create strong posterior correlations, as these correlations represent the exact information lost by the mean-field approximation.

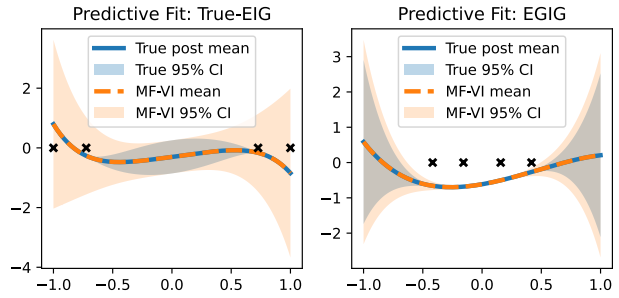

Figure 3: Posterior predictive distributions. (Left) For the EIG design, MF-VI (orange) correctly captures the mean but severely overestimates uncertainty versus the true posterior (blue). (Right) For the EGIG design, the approximation is nearly perfect. Design points are designated by black crosses.

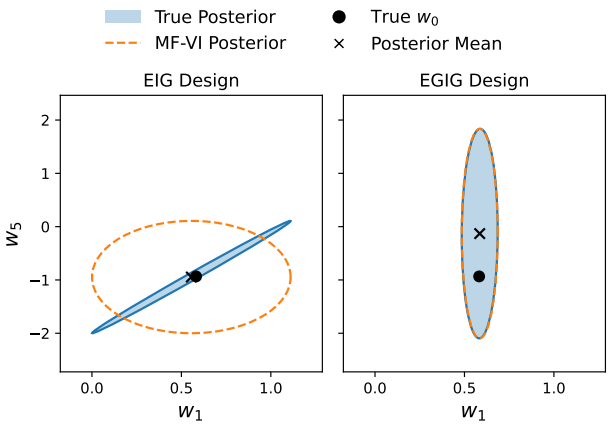

Figure 4: Representative 2D marginal posterior. (Left) The EIG design produces a strongly correlated true posterior that axis-aligned MF-VI fails to capture. (Right) The EGIG design produces a posterior well-approximated by MF-VI.

6.3 Neural Network Surrogate for a Coupled Spring–Mass System

We examine misspecification from a neural network surrogate for Bayesian calibration. True data are generated from a coupled spring-mass system with six parameters $(m_1, m_2, b_1, b_2, k_1, k_2)$ having independent log-normal priors ($\mu = 0, \sigma = 1$). The dynamics are

$$\begin{aligned} x_1' &= v_1, \quad x_2' = v_2 \\ v_1' &= \frac{-b_1 v_1 - (k_1 + k_2)x_1 + k_2 x_2}{m_1}, \\ v_2' &= \frac{-b_2 v_2 + k_2(x_1 - x_2) + f(\gamma, t)}{m_2}, \end{aligned} \tag{21}$$

with forcing $f(\gamma, t) = 5\sin(\gamma t)\exp(-t/5)$. We observe position x_1 at 100 time points with Gaussian noise ($\sigma = 0.025$). To infer parameters, we replace the costly true dynamics with a trained neural network surrogate, inducing posterior misspecification.

The design problem is selecting the working model noise variance σ^2_{noise} for different dataset sizes, creating a trade-off: lower noise increases sensitivity to data features but also to surrogate error.

Figure 5 shows the three criteria offer conflicting advice. From the working model's perspective, lower noise always appears more informative, so EIG (solid lines, circles) always prefers the smallest noise ($\sigma_{noise} = 0.1$) regardless of surrogate fidelity. Both EGIG (dashed lines, squares) and $I^{\star}_{\text{KL}}$ (dotted lines, triangles) correctly diagnose the issue: with an inaccurate surrogate (small training set), the lowest-noise designs yield negative utility, signaling they are harmful and will create bias-dominated posteriors. As estab-

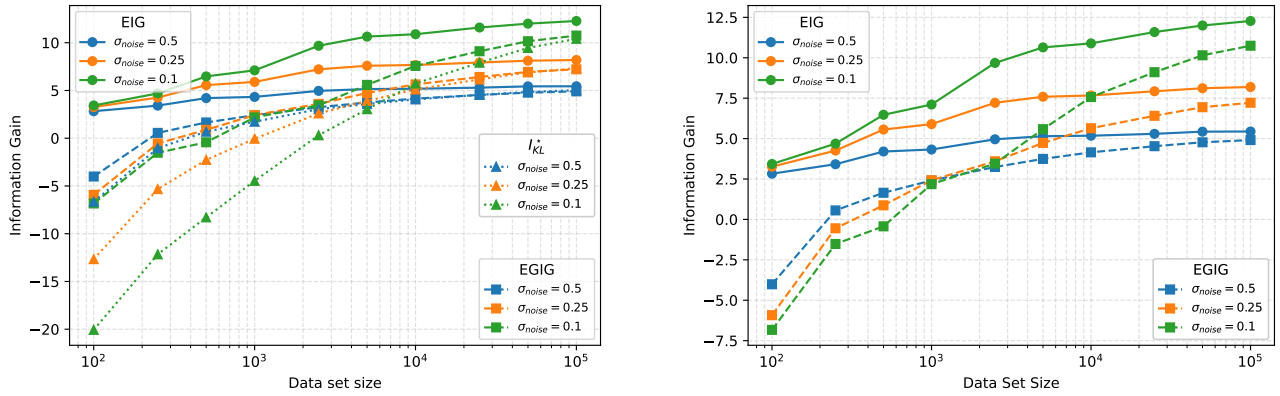

Figure 5: Comparison of EIG, EGIG, and $I^{\star}_{\mathrm{KL}}$ on the neural network surrogate experiment. The left panel shows a full comparison of information measures, while the right panel shows just EIG and EGIG for easier parsing. EIG (solid lines, circles) monotonically increases, consistently favoring the most aggressive noise model. EGIG (dashed lines, squares) reveals this design is initially harmful with negative utility and only becomes optimal as the surrogate improves. The most conservative criterion is $I^{\star}_{\mathrm{KL}}$ (dotted lines, triangles), requiring much higher surrogate fidelity than EGIG before yielding positive utility.

lished in Lemma 4.4, $I^{\star}_{\mathrm{KL}}$ is a strictly more conservative criterion than EGIG. This is empirically observed at $\sigma_{noise} = 0.1$ with 1,000 training samples, where EGIG is positive (2.18 nats) but $I^{\star}_{\mathrm{KL}}$ remains significantly negative (-4.46 nats). Because the gap between these metrics is exactly the error in the marginal data distribution, this large discrepancy indicates that the marginal $p_{\mathcal{M}}(y|d)$ is highly misspecified. This suggests the surrogate learns the relative sensitivities needed for posterior inference much faster than it learns to accurately predict the absolute data distribution; while EGIG rewards the former, $I^{\star}_{\mathrm{KL}}$ strictly penalizes the latter. As surrogate fidelity improves, both robust measures increase for all designs, eventually favoring the lowest noise.

Negative utility values signal the misspecified posterior is less reliable than the prior. The zero-crossing determines minimum surrogate fidelity (training dataset size) required for a design to be useful. For example, $\sigma_{\mathrm{noise}} = 0.1$ needs over 1,000 training samples for positive information gain under EGIG, while more conservative designs are useful sooner. This offers a principled criterion for deciding if a surrogate is "good enough".

7 CONCLUSION

In this work we develop a general framework for experimental design under model misspecification by extending the classical Ginebra criteria. This provides a theoretical foundation for robust utilities that account for model error, from which Expected Generalized Information Gain (EGIG) emerges as a natural special case. Our framework unifies robust design and can motivate other robust criteria beyond EGIG, tailored to different modeling and computational contexts.

Our experiments show that EGIG provides reliable guidance where classical criteria fail. It successfully navigated structural model error in a spring-mass system, managed the trade-off between surrogate accuracy and likelihood noise for a neural network emulator, and avoided pathological designs for approximate variational inference. In each case, EGIG made robust design choices that were inaccessible to traditional methods.

Acknowledgements

Sandia National Laboratories (SNL) is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2026-18312C. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

Ali, A., Caruana, R., and Kapoor, A. (2014). Active Learning with Model Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Ali, S. M. and Silvey, S. D. (1966). A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142. Publisher: [Royal Statistical Society, Oxford University Press].

Attia, A., Leyffer, S., and Munson, T. S. (2025). Robust A-Optimal Experimental Design for Sensor Placement in Bayesian Linear Inverse Problems. *SIAM/ASA Journal on Uncertainty Quantification*, 13(2):744–774. Publisher: Society for Industrial and Applied Mathematics.

Blackwell, D. (1951). Comparison of Experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 93–103. University of California Press.

Blackwell, D. (1953). Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272. Publisher: Institute of Mathematical Statistics.

Brynjarsdóttir, J. and O'Hagan, A. (2014). Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30(11):114007.

Catanach, T. A. and Das, N. (2023). Metrics for bayesian optimal experiment design under model misspecification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 7707–7714. IEEE.

Dong, J., Jacobsen, C., Khalloufi, M., Akram, M., Liu, W., Duraisamy, K., and Huan, X. (2025). Variational Bayesian optimal experimental design with normalizing flows. *Computer Methods in Applied Mechanics and Engineering*, 433:117457.

Duersch, J. and Catanach, T. (2020). Generalizing Information to the Evolution of Rational Belief. *Entropy*, 22(1):108. Num Pages: 108 Place: Basel, Switzerland Publisher: MDPI AG.

Duong, D.-L., Helin, T., and Rojo-Garcia, J. R. (2023). Stability estimates for the expected utility in Bayesian optimal experimental design. *Inverse Problems*, 39(12):125008.

Feng, C. (2015). *Optimal Bayesian experimental design in the presence of model error*. Thesis, Massachusetts Institute of Technology. Accepted: 2015-07-17T19:46:48Z.

Forster, A., Ivanova, D. R., and Rainforth, T. (2025). Improving robustness to model misspecification in bayesian experimental design. In *7th Symposium on Advances in Approximate Bayesian Inference-Workshop Track*.

Foster, A., Ivanova, D. R., Malik, I., and Rainforth, T. (2021). Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3384–3395. PMLR. ISSN: 2640-3498.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. (2020). Variational Bayesian Optimal Experimental Design. arXiv:1903.05480 [stat].

Ginebra, J. (2007). On the measure of the information in a statistical experiment. *Bayesian Analysis*, 2(1).

Go, J. and Isaac, T. (2022). Robust Expected Information Gain for Optimal Bayesian Experimental Design Using Ambiguity Sets. arXiv:2205.09914 [stat].

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian Calibration of Computer Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(3):425–464.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kleinegesse, S. and Gutmann, M. U. (2020). Bayesian Experimental Design for Implicit Models by Mutual Information Neural Estimation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5316–5326. PMLR. ISSN: 2640-3498.

Lei, B., Kirk, T. Q., Bhattacharya, A., Pati, D., Qian, X., Arroyave, R., and Mallick, B. K. (2021). Bayesian optimization with adaptive surrogate models for automated experimental design. *npj Computational Materials*, 7(1):194.

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005. Publisher: Institute of Mathematical Statistics.

Overstall, A. and McGree, J. (2022). Bayesian Decision-Theoretic Design of Experiments Under an Alternative Model. *Bayesian Analysis*, 17(4).

Petsagkourakis, P. and Galvanin, F. (2021). Safe model-based design of experiments using Gaussian processes. *Computers & Chemical Engineering*, 151:107339.

Rainforth, T., Foster, A., Ivanova, D. R., and Smith, F. B. (2023). Modern Bayesian Experimental Design. arXiv:2302.14545 [stat].

Ryan, K. J. (2003). Estimating Expected Information Gains for Experimental Designs with Application to the Random Fatigue-Limit Model. *Journal of*

Computational and Graphical Statistics, 12(3):585–603. Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America].

Sloman, S. J., Oppenheimer, D. M., Broomell, S. B., and Shalizi, C. R. (2022). Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias. arXiv:2205.13698 [stat].

Sugiyama, M. (2005). Active Learning for Misspecified Models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Tang, R., Sloman, S. J., and Kaski, S. (2025). Generalization Analysis for Bayesian Optimal Experiment Design under Model Misspecification. arXiv:2506.07805 [stat].

Tsirpitzi, R. E., Miller, F., and Burman, C.-F. (2023). Robust optimal designs using a model misspecification term. *Metrika*, 86(7):781–804.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] The mathematical framework is detailed in Sections 2, 3, and 4. All models used for experiments are described in Section 6.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] The paper does not present algorithms that require complexity analysis.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Code for replicating the experiments is provided in the supplemental material file.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Assumptions are stated for all propositions (e.g., Props. 3.1, 4.1, and 4.3) and the proposed framework itself (Section 3.2).
   (b) Complete proofs of all theoretical results. [Yes] All proofs can be found in Appendix C.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The supplemental material contains code to reproduce all experiments, which are based on synthetic data.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Details for the VI and NN surrogate experiments, including model structures and training setups, are described in Section 6.2 and 6.3.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] The utility measures are defined in Sections 2 and 4. Uncertainty is represented explicitly with error bars in Figure 2. Figure 5 omits error bars as the uncertainty from the Monte Carlo procedure is negligible on the scale of the figures. Other figures represent uncertainty by displaying full posterior or predictive distributions.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] This information is provided in Appendix A.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable] All data used in this work are synthetic.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A NESTED MONTE CARLO

The utility curves in Figure 2 were estimated using Nested Monte Carlo (NMC) procedures.

The EGIG objective is an expectation over the true joint distribution. We estimate it by drawing N outer-loop samples (θ_i, y_i) from the true joint distribution $p_{\mathcal{M}^\star}$ and using an M-sample inner loop to estimate the intractable marginal likelihood $p_{\mathcal{M}}(y_i \mid d)$ for each sample. The estimator is given by:

$$\widehat{EGIG}(d) = \frac{1}{N}\sum_{i=1}^{N} \log p_{\mathcal{M}}(y_i \mid \theta_i, d) - \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M}\sum_{j=1}^{M} p_{\mathcal{M}}(y_i \mid \theta_j, d)\right) \tag{22}$$

where $(\theta_i, y_i) \sim p_{\mathcal{M}^\star}(y \mid \theta, d)p(\theta)$ and $\theta_j \sim p(\theta)$.

The $I^\star_{KL}$ utility was also estimated via NMC. From its definition, $I^\star_{KL}$ can be simplified to the expectation $I^\star_{KL}(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(\theta,y|d)}\left[\log p_{\mathcal{M}}(y \mid \theta, d) - \log p_{\mathcal{M}^\star}(y \mid d)\right]$. We approximate this by drawing N outer-loop samples $(\theta_i, y_i) \sim p_{\mathcal{M}^\star}(\theta, y \mid d)$. For each sample, we compute the working likelihood term $\log p_{\mathcal{M}}(y_i \mid \theta_i, d)$ directly. The true marginal likelihood, $p_{\mathcal{M}^\star}(y_i \mid d)$, is intractable and is estimated with an M-sample inner Monte Carlo loop: $p_{\mathcal{M}^\star}(y_i \mid d) \approx \frac{1}{M}\sum_{j=1}^{M} p_{\mathcal{M}^\star}(y_i \mid \theta_j, d)$, where $\theta_j \sim p(\theta)$.

All experiments were conducted in a Jupyter Notebook environment. The computations were performed on a node of a high-performance computing cluster. For the experiments requiring significant computation, a single NVIDIA A100 40GB GPU was utilized, allocated from a node with 32 CPU cores and 512GB of RAM.

B MORE EXAMPLES

B.1 Blackwell Monotonicity Violation: A Concrete Example

We construct a discrete example where a more informative experiment has lower misspecified utility than a less informative one.

Setup: Consider a binary parameter $\theta \in \{A, B\}$ with uniform prior $p(\theta = A) = p(\theta = B) = 0.5$, and binary data $Y \in \{y_1, y_2\}$.

The true model $\mathcal{M}^\star$ for each design is

$$\begin{aligned} p_{\mathcal{M}^\star}(y_1|\theta, d_0) &= \begin{cases} 0.6 & \theta = A \\ 0.4 & \theta = B \end{cases} \quad \text{(weakly informative)}, \\ p_{\mathcal{M}^\star}(y_1|\theta, d_1) &= \begin{cases} 0.9 & \theta = A \\ 0.1 & \theta = B \end{cases} \quad \text{(strongly informative)}. \end{aligned} \tag{23}$$

Clearly d_1 is sufficient for d_0 under $\mathcal{M}^\star$, so d_1 is more informative.

The working model $\mathcal{M}$ is defined for each design as:

$$\begin{aligned} p_{\mathcal{M}}(y_1|\theta, d_0) &= \begin{cases} 0.6 & \theta = A \\ 0.4 & \theta = B \end{cases} \text{(correctly specified)}, \\ p_{\mathcal{M}}(y_1|\theta, d_1) &= \begin{cases} 0.5 & \theta = A \\ 0.5 & \theta = B \end{cases} \quad \text{(believes } Y \perp \theta\text{)}. \end{aligned} \tag{24}$$

Computing misspecified utilities (with $\varphi = D_{KL}(\cdot\|p(\theta))$):

For d_0, both models agree, so the working and true posteriors coincide:

$$\begin{aligned} I^{\text{true}}_{\varphi}(d_0) &= \mathbb{E}_{y\sim p_{\mathcal{M}^\star}(y|d_0)}[D_{KL}(p_{\mathcal{M}^\star}(\theta|y, d_0)\|p(\theta))] \\ &= 0.5[0.6\ln(1.2) + 0.4\ln(0.8)] + 0.5[0.4\ln(0.8) + 0.6\ln(1.2)] \\ &= 0.6\ln(1.2) + 0.4\ln(0.8) \approx 0.020 \text{ nats} \end{aligned}$$

Since models agree: $\Delta_\varphi(d_0) = 0$, thus $I_\varphi^{mis}(d_0) = 0.020$ nats.

For d_1, the true model provides substantial information:

$$I_\varphi^{\text{true}}(d_1) = 0.9\ln(1.8) + 0.1\ln(0.2) \approx 0.368 \text{ nats}$$

However, the working model believes $Y \perp \theta$, so its posterior equals the prior for all y, giving $\varphi(p_{\mathcal{M}}(\theta|y, d_1)) = 0$. The penalty is:

$$\begin{aligned}\Delta_\varphi(d_1) &= \mathbb{E}_y[\varphi(p_{\mathcal{M}^\star}(\theta|y, d_1)) - \varphi(p_{\mathcal{M}}(\theta|y, d_1))] \\ &= \mathbb{E}_y[0.368 - 0] = 0.368 \text{ nats}\end{aligned}$$

This yields $I_\varphi^{mis}(d_1) = 0.368 - 0.368 = 0$ nats.

Therefore, we have $I_\varphi^{mis}(d_1) = 0 < 0.020 = I_\varphi^{mis}(d_0)$, even though d_1 is sufficient for (more informative than) d_0 under the true model. The misspecified utility violates Blackwell monotonicity, preferring the less informative experiment because the working model's complete failure to recognize the signal in d_1 incurs a penalty equal to the true information gain.

B.2 Amortized Variational Inference

Figure 6 displays all fifteen 2D marginal posterior distributions for the six-parameter Bayesian polynomial regression problem. For both the EIG-optimal and EGIG-optimal designs, we show the true posterior (computed via exact Bayesian inference) and the mean-field variational approximation. The EIG design consistently induces strong correlations in the true posterior across parameter pairs, which the axis-aligned mean-field approximation cannot capture. In contrast, the EGIG design produces posteriors with weaker correlations that are well-approximated by the factorized variational posterior.

C PROOFS

C.1 Proof of the Blackwell-Sherman-Stein Theorem

The proof of the Blackwell-Sherman-Stein Theorem (which corresponds to Proposition 3.2 in Ginebra (2007)) relies on showing its equivalence to the original formulation of the theorem (Proposition 3.1 in Ginebra (2007)). This first proposition is stated in terms of the likelihood-ratio statistic $T_p(y)$.

Proposition C.1 (Blackwell-Sherman-Stein, Prop. 3.1 in Ginebra (2007))**.** *Experiment d_1 is sufficient for experiment d_2 if and only if for some strictly positive prior density $p(\theta)$,*

$$\mathbb{E}_{p(y_1)}[\phi(T_p(y_1))] \geq \mathbb{E}_{p(y_2)}[\phi(T_p(y_2))]$$

for every convex functional $\phi(\cdot)$.

The proof of Proposition C.1 can be found in Blackwell (1953). We now show that Equation (3) is an equivalent formulation.

Proof. We show that Proposition C.1 and the statement in Section 2.2 are mathematically equivalent. The proof follows from a direct substitution based on the definitions in Ginebra (2007).

First, the likelihood-ratio statistic $T_p(y)$ is the posterior-to-prior ratio *function*

$$T_p(y)(\theta) = \frac{p(\theta \mid y)}{p(\theta)}.$$

Second, the functional $\varphi(\cdot)$ is defined in relation to $\phi(\cdot)$. For any convex function ϕ that takes a ratio function $u(\theta)$ as input, there exists an equivalent convex function φ that takes a posterior density function $h(\theta)$ as input. This relationship is given by the change of variables:

$$\phi(u) := \varphi(p \cdot u),$$

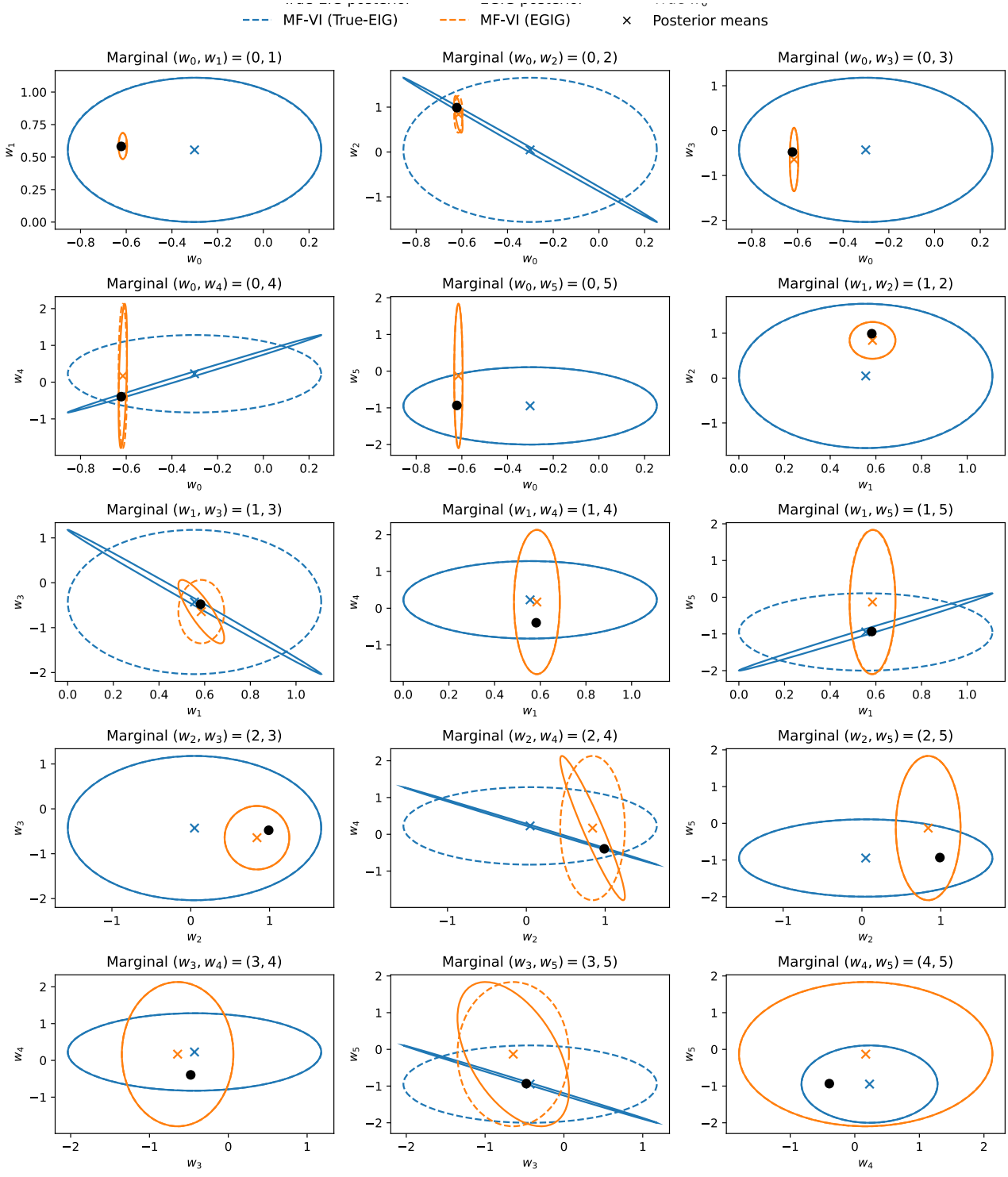

Figure 6: Complete set of 2D marginal posteriors for the variational inference experiment. Each subplot shows one pair of parameters (w_i, w_j) with posteriors from both designs overlaid. Blue: EIG design the solid line is the true posterior, and the dashed line is the MF-VI approximation. Orange: EGIG design. The solid line is the true posterior, and the dashed line is the MF-VI approximation. For the EIG design (blue), the true posterior exhibits strong correlations that the axis-aligned MF-VI approximation fails to capture, resulting in substantial discrepancies between solid and dashed blue contours. For the EGIG design (orange), the true posterior has less correlation and the MF-VI approximation matches it closely, with solid and dashed orange contours nearly coinciding. Black dot: true parameter value. X marks: posterior means.

where $(p \cdot u)$ is the function defined by the pointwise product $(p \cdot u)(\theta) = p(\theta)u(\theta)$.

By substituting the definition of $T_p(y)$ into this relationship, we see the equivalence. The argument to ϕ is the function $T_p(y)$, so

$$\phi(T_p(y)) = \varphi(p \cdot T_p(y)).$$

The argument to φ is the function $(p \cdot T_p(y))(\theta)$, which evaluates to

$$p(\theta) \cdot T_p(y)(\theta) = p(\theta) \cdot \frac{p(\theta \mid y)}{p(\theta)} = p(\theta \mid y).$$

Thus, the argument to φ is the posterior density function itself, giving the identity

$$\phi(T_p(y)) = \varphi(p(\theta \mid y)).$$

Therefore, the statement in Proposition C.1 equivalent to the statement in Section 2.2. This completes the proof. □

C.2 Proof of Proposition 4.1

Proof. **Decomposition.** Recall that the definition of EGIG is

$$EGIG(d) = \mathbb{E}_{p_{\mathcal{M}^\star}(y,\theta|d)}\left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p(\theta)}\right].$$

By adding and subtracting $\log p_{\mathcal{M}^\star}(\theta \mid y, d)$ inside the logarithm, we get

$$\begin{aligned}
\mathrm{EGIG}(d) &= \mathbb{E}_{p_{\mathcal{M}^\star}(\theta,y|d)}\left[\log \frac{p_{\mathcal{M}^\star}(\theta \mid y, d)}{p(\theta)} + \log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p_{\mathcal{M}^\star}(\theta \mid y, d)}\right] \\
&= \mathbb{E}_{p_{\mathcal{M}^\star}(\theta,y|d)}\left[\log \frac{p_{\mathcal{M}^\star}(\theta \mid y, d)}{p(\theta)}\right] + \mathbb{E}_{p_{\mathcal{M}^\star}(\theta,y|d)}\left[-\log \frac{p_{\mathcal{M}^\star}(\theta \mid y, d)}{p_{\mathcal{M}}(\theta \mid y, d)}\right] \\
&= \mathrm{EIG}^{\mathrm{true}}(d) - \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}\left[\mathbb{E}_{p_{\mathcal{M}^\star}(\theta|y,d)}\left[\log \frac{p_{\mathcal{M}^\star}(\theta \mid y, d)}{p_{\mathcal{M}}(\theta \mid y, d)}\right]\right] \\
&= \mathrm{EIG}^{\mathrm{true}}(d) - \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}\left[D_{KL}(p_{\mathcal{M}^\star}(\theta \mid y, d) \;||\; p_{\mathcal{M}}(\theta \mid y, d))\right] \\
&= \mathrm{EIG}^{\mathrm{true}}(d) - \Delta_{\mathrm{EGIG}}(d).
\end{aligned}$$

(R1) Real-Valued. We assume standard regularity conditions such that the expectations and KL divergences involved are finite, making $\mathrm{EGIG}(d)$ a finite real number.

(R2) Non-Informative. If an experiment is non-informative under $\mathcal{M}^\star$, then $Y \perp \theta$, which implies $p_{\mathcal{M}^\star}(\theta \mid y, d) = p(\theta)$. We can then write EGIG as:

$$\begin{aligned}
\mathrm{EGIG}(d) &= \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}\left[\mathbb{E}_{p_{\mathcal{M}^\star}(\theta|y,d)}\left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p(\theta)}\right]\right] \\
&= \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}\left[\mathbb{E}_{p(\theta)}\left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p(\theta)}\right]\right] \\
&= \mathbb{E}_{p_{\mathcal{M}^\star}(y|d)}\left[-D_{KL}(p(\theta) \;||\; p_{\mathcal{M}}(\theta \mid y, d))\right].
\end{aligned}$$

Since the KL divergence is always non-negative, its negative is non-positive. The expectation of a non-positive quantity is non-positive, so $\mathrm{EGIG}(d) \leq 0$. Equality holds if and only if $D_{KL}(p(\theta) \;||\; p_{\mathcal{M}}(\theta \mid y, d)) = 0$ for almost every y, which requires $p_{\mathcal{M}}(\theta \mid y, d) = p(\theta)$ almost everywhere.

(R3) Information-Penalty Structure. The decomposition proven above, $\mathrm{EGIG}(d) = \mathrm{EIG}^{\mathrm{true}}(d) - \Delta_{\mathrm{EGIG}}(d)$, matches the required structure.

First, $U^{\mathrm{true}}(d) = \mathrm{EIG}^{\mathrm{true}}(d)$ is the mutual information $I_{\mathcal{M}^\star}(\theta; Y \mid d)$, which is the standard measure of expected information gain under the true model. It is a known result that mutual information satisfies the classical Ginebra axioms (G1)–(G3).

Second, $\Delta_U(d) = \Delta_{\text{EGIG}}(d)$ is the expectation of a KL divergence. Since $D_{KL} \geq 0$, its expectation is also non-negative, so $\Delta_{\text{EGIG}}(d) \geq 0$.

(R4) Bounded Contraction Under Sufficiency. Let design d_1 be sufficient for design d_2. This implies a Markov chain $\theta \to Y_1 \to Y_2$ under the true model, where Y_1, Y_2 are the data from designs d_1, d_2 respectively.

True information contracts: By the Data Processing Inequality for mutual information, the Markov chain $\theta \to Y_1 \to Y_2$ implies $I_{\mathcal{M}^\star}(\theta; Y_2) \leq I_{\mathcal{M}^\star}(\theta; Y_1)$. This is equivalent to $\text{EIG}^{\text{true}}(d_2) \leq \text{EIG}^{\text{true}}(d_1)$.

Penalty contracts: If the sufficiency holds via a common kernel for both $\mathcal{M}^\star$ and $\mathcal{M}$, then we have the Markov chain $\theta \to Y_1 \to Y_2$ under both models. The Data Processing Inequality for conditional KL divergence states that this implies $D_{KL}(p_{\mathcal{M}^\star}(\theta \mid Y_2) \;||\; p_{\mathcal{M}}(\theta \mid Y_2)) \leq D_{KL}(p_{\mathcal{M}^\star}(\theta \mid Y_1) \;||\; p_{\mathcal{M}}(\theta \mid Y_1))$. Taking the expectation of both sides with respect to the true data distribution $p_{\mathcal{M}^\star}(y_1, y_2 \mid d_1)$ gives:

$$\mathbb{E}_{p_{\mathcal{M}^\star}(y_2 \mid d_2)}\left[D_{KL}(p_{\mathcal{M}^\star}(\theta \mid y_2) \;||\; p_{\mathcal{M}}(\theta \mid y_2))\right] \leq \mathbb{E}_{p_{\mathcal{M}^\star}(y_1 \mid d_1)}\left[D_{KL}(p_{\mathcal{M}^\star}(\theta \mid y_1) \;||\; p_{\mathcal{M}}(\theta \mid y_1))\right]$$

This is precisely $\Delta_{\text{EGIG}}(d_2) \leq \Delta_{\text{EGIG}}(d_1)$.

Thus, EGIG satisfies all robust axioms (R1)–(R4). □

C.3 Proof of Proposition 4.3

Throughout this proof we assume the standard measure-theoretic regularity conditions (absolute continuity and integrability) that ensure all f-divergences considered are finite.

Proof. **(R1) Real-Valuedness:** Under standard measurability and integrability conditions, both $D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \,||\, p(\theta)p_{\mathcal{M}^\star}(y \mid d))$ and $D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \,||\, p_{\mathcal{M}}(\theta, y \mid d))$ are finite. Hence $I_f^\star(d) \in \mathbb{R}$.

(R2) Non-Informative: Suppose $Y \perp \theta$ under $\mathcal{M}^\star$. Then $p_{\mathcal{M}^\star}(\theta, y \mid d) = p(\theta)p_{\mathcal{M}^\star}(y \mid d)$, which gives:

$$I_f^\star(d) = D_f(p(\theta)p_{\mathcal{M}^\star}(y \mid d) \,||\, p(\theta)p_{\mathcal{M}^\star}(y \mid d)) = 0.$$

For the penalty, the nonnegativity of f-divergences gives us

$$\Delta_f(d) = D_f(p(\theta)p_{\mathcal{M}^\star}(y \mid d) \,||\, p_{\mathcal{M}}(\theta, y \mid d)) \geq 0,$$

with equality if and only if $p_{\mathcal{M}}(\theta, y \mid d) = p(\theta)p_{\mathcal{M}^\star}(y \mid d)$ almost everywhere.

Therefore,

$$I_f^\star(d) = 0 - \Delta_f(d) \leq 0,$$

with equality if and only if $p_{\mathcal{M}}(\theta, y \mid d) = p(\theta)p_{\mathcal{M}^\star}(y \mid d)$, which occurs if and only if $p_{\mathcal{M}}(\theta \mid y, d) = p(\theta)$ almost everywhere. This establishes (R2).

(R3) Information-Penalty Structure: By definition:

$$I_f^\star(d) = I_f^{\text{true}}(d) - \Delta_f(d). \tag{25}$$

Both terms are non-negative by properties of f-divergences.

We now verify that $I_f^{\text{true}}(d)$ satisfies the classical Ginebra axioms (G1)–(G3).

(G1) Real-Valuedness: Under standard regularity conditions, $D_f(p_{\mathcal{M}^\star}(\theta, y \mid d) \,||\, p(\theta)p_{\mathcal{M}^\star}(y \mid d))$ is a finite real number.

(G2) Non-Informativeness: If $Y \perp \theta$ under $\mathcal{M}^\star$, then $p_{\mathcal{M}^\star}(\theta, y \mid d) = p(\theta)p_{\mathcal{M}^\star}(y \mid d)$, giving

$$I_f^{\text{true}}(d) = D_f(p(\theta)p_{\mathcal{M}^\star}(y \mid d) \,||\, p(\theta)p_{\mathcal{M}^\star}(y \mid d)) = 0. \tag{26}$$

(G3) Blackwell Monotonicity: If d_1 is sufficient for d_2 then by the data processing inequality for f-divergences,

$$I_f^{\text{true}}(d_2) \leq I_f^{\text{true}}(d_1).$$

Thus $I_f^{\text{true}}(d)$ is a valid information measure according to the Ginebra criteria. This establishes (R3).

(R4) Bounded Contraction Under Sufficiency:

Assume d_1 is sufficient for d_2 via a common transition kernel $K(y_2 \mid y_1)$ for both models.

Part (a): True information contracts.

Since d_1 is sufficient for d_2 via kernel $K(y_2 \mid y_1)$ under $\mathcal{M}^\star$, we have that

$$p_{\mathcal{M}^\star}(\theta, y_2 \mid d_2) = \int K(y_2 \mid y_1) p_{\mathcal{M}^\star}(\theta, y_1 \mid d_1) dy_1 \tag{27}$$

and

$$p(\theta) p_{\mathcal{M}^\star}(y_2 \mid d_2) = \int K(y_2 \mid y_1) p(\theta) p_{\mathcal{M}^\star}(y_1 \mid d_1) dy_1. \tag{28}$$

By the data processing inequality for f-divergences:

$$I_f^{\text{true}}(d_2) = D_f(p_{\mathcal{M}^\star}(\theta, y_2 \mid d_2) \,||\, p(\theta) p_{\mathcal{M}^\star}(y_2 \mid d_2)) \leq D_f(p_{\mathcal{M}^\star}(\theta, y_1 \mid d_1) \,||\, p(\theta) p_{\mathcal{M}^\star}(y_1 \mid d_1)) = I_f^{\text{true}}(d_1). \tag{29}$$

Part (b): Penalty contracts under common kernel.

Since the kernel K is common to both models, both the true joint and the working joint transform via the same Markov kernel:

$$p_{\mathcal{M}^\star}(\theta, y_2 \mid d_2) = \int K(y_2 \mid y_1) p_{\mathcal{M}^\star}(\theta, y_1 \mid d_1) dy_1 \tag{30}$$

and

$$p_{\mathcal{M}}(\theta, y_2 \mid d_2) = \int K(y_2 \mid y_1) p_{\mathcal{M}}(\theta, y_1 \mid d_1) dy_1. \tag{31}$$

By the data processing inequality for f-divergences,

$$\Delta_f(d_2) = D_f(p_{\mathcal{M}^\star}(\theta, y_2 \mid d_2) \,||\, p_{\mathcal{M}}(\theta, y_2 \mid d_2)) \leq D_f(p_{\mathcal{M}^\star}(\theta, y_1 \mid d_1) \,||\, p_{\mathcal{M}}(\theta, y_1 \mid d_1)) = \Delta_f(d_1). \tag{32}$$

This completes the proof of (R4). □

C.4 Proof of Lemma 4.4

Proof. Recall the definition of the Expected Generalized Information Gain (EGIG) from Equation (7):

$$\text{EGIG}(d) = \mathbb{E}_{p_{\mathcal{M}^*}(\theta, y \mid d)} \left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p(\theta)} \right] \tag{33}$$

and the Robust f-Information Measure instantiated with the Kullback-Leibler divergence ($I_{\text{KL}}^\star$) from Equation (3):

$$I_{\text{KL}}^\star(d) = D_{KL}(p_{\mathcal{M}^*}(\theta, y \mid d) \parallel p(\theta) p_{\mathcal{M}^*}(y \mid d)) - D_{KL}(p_{\mathcal{M}^*}(\theta, y \mid d) \parallel p_{\mathcal{M}}(\theta, y \mid d)). \tag{34}$$

By the chain rule of KL divergence, the first term in $I_{\text{KL}}^\star(d)$ is the true mutual information $I^{\text{true}}(d)$:

$$I^{\text{true}}(d) = \mathbb{E}_{p_{\mathcal{M}^*}(\theta, y \mid d)} \left[\log \frac{p_{\mathcal{M}^*}(\theta, y \mid d)}{p(\theta) p_{\mathcal{M}^*}(y \mid d)} \right]. \tag{35}$$

The second term is the penalty $\Delta_{KL}(d)$:

$$\Delta_{KL}(d) = \mathbb{E}_{p_{\mathcal{M}^*}(\theta, y \mid d)} \left[\log \frac{p_{\mathcal{M}^*}(\theta, y \mid d)}{p_{\mathcal{M}}(\theta, y \mid d)} \right]. \tag{36}$$

Subtracting the penalty from the true information yields:

$$I_{\text{KL}}^\star(d) = \mathbb{E}_{p_{\mathcal{M}^*}(\theta, y \mid d)} \left[\log \frac{p_{\mathcal{M}^*}(\theta, y \mid d)}{p(\theta) p_{\mathcal{M}^*}(y \mid d)} - \log \frac{p_{\mathcal{M}^*}(\theta, y \mid d)}{p_{\mathcal{M}}(\theta, y \mid d)} \right] \tag{37}$$

$$= \mathbb{E}_{p_{\mathcal{M}^*}(\theta, y \mid d)} \left[\log \frac{p_{\mathcal{M}}(\theta, y \mid d)}{p(\theta) p_{\mathcal{M}^*}(y \mid d)} \right]. \tag{38}$$

We now decompose the working joint distribution into $p_{\mathcal{M}}(\theta, y \mid d) = p_{\mathcal{M}}(\theta \mid y, d)p_{\mathcal{M}}(y \mid d)$:

$$I^{\star}_{\mathrm{KL}}(d) = \mathbb{E}_{p_{\mathcal{M}^*}(\theta,y|d)}\left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)p_{\mathcal{M}}(y \mid d)}{p(\theta)p_{\mathcal{M}^*}(y \mid d)}\right] \tag{39}$$

$$= \mathbb{E}_{p_{\mathcal{M}^*}(\theta,y|d)}\left[\log \frac{p_{\mathcal{M}}(\theta \mid y, d)}{p(\theta)}\right] + \mathbb{E}_{p_{\mathcal{M}^*}(y|d)}\left[\log \frac{p_{\mathcal{M}}(y \mid d)}{p_{\mathcal{M}^*}(y \mid d)}\right]. \tag{40}$$

The first term is exactly the definition of $\mathrm{EGIG}(d)$. The second term is the negative KL divergence between the true and working marginal data distributions:

$$I^{\star}_{\mathrm{KL}}(d) = \mathrm{EGIG}(d) - D_{KL}(p_{\mathcal{M}^*}(y \mid d) \parallel p_{\mathcal{M}}(y \mid d)). \tag{41}$$

Rearranging gives the desired decomposition:

$$\mathrm{EGIG}(d) = I^{\star}_{\mathrm{KL}}(d) + D_{KL}(p_{\mathcal{M}^*}(y \mid d) \parallel p_{\mathcal{M}}(y \mid d)). \tag{42}$$

Since the KL divergence is non-negative, $\mathrm{EGIG}(d) \geq I^{\star}_{\mathrm{KL}}(d)$, with equality if and only if $p_{\mathcal{M}^*}(y \mid d) = p_{\mathcal{M}}(y \mid d)$ almost everywhere. □