A Geometric Approach to Optimal Experimental Design

Gavin Kerrigan
University of Oxford
kerrigan@stats.ox.ac.uk

Christian A. Naesseth
University of Amsterdam
c.a.naesseth@uva.nl

Tom Rainforth
University of Oxford
rainforth@stats.ox.ac.uk

Abstract

We introduce a novel geometric framework for optimal experimental design (OED). Traditional OED approaches, such as those based on mutual information, rely explicitly on probability densities, leading to restrictive invariance properties. To address these limitations, we propose the mutual transport dependence (MTD), a measure of statistical dependence grounded in optimal transport theory which provides a geometric objective for optimizing designs. Unlike conventional approaches, the MTD can be tailored to specific downstream estimation problems by choosing appropriate geometries on the underlying spaces. We demonstrate that our framework produces high-quality designs while offering a flexible alternative to standard information-theoretic techniques.

1 INTRODUCTION

Effective experimental design is central to a wide range of scientific and industrial applications (Kuhfeld et al., 1994; Park et al., 2013; Melendez et al., 2021). Many such problems require a principled, model-based, approach, wherein we utilize a model over possible experimental outcomes to directly optimize our design decisions. This can be particularly effective in *adaptive* design settings, where frameworks like sequential Bayesian experimental design (BED) allow us to iterate between using our model to make design decisions and updating our model with the collected data (DeGroot, 1962; MacKay, 1992; Sebastiani and Wynn, 2000; Rainforth et al., 2024). Many of these approaches are grounded in *information theory*, where the value of an experiment is quantified using the values of an associated probability density.

For instance, a popular and principled approach in the BED literature is optimizing the mutual information (MI) (Lindley, 1956, 1972; Bernardo, 1979)

$$\mathcal{I}(d) = \mathsf{KL}\left[p(\theta, y \mid d) \,||\, p(\theta)p(y \mid d)\right] \tag{1}$$

$$= \mathbb{E}_{p(\theta,y|d)}\left[\log\left(\frac{p(\theta, y \mid d)}{p(\theta)p(y \mid d)}\right)\right] \tag{2}$$

where $p(\theta)$ is the prior over the quantity of interest θ, and $p(y \mid \theta, d)$ models the experiment outcome y under design d. Notably, the MI is an expectation of log-density ratios, and is thus a *unitless* quantity. Consequently, the MI has strong invariance properties: any injective transformation of θ or y leaves $\mathcal{I}(d)$ unchanged.

We highlight that this invariance, while often attractive, can also be detrimental for optimal experimental design (OED). Experimental goals should be defined in terms of downstream errors (Lindley, 1972) and many common error metrics, such as mean squared error between true and estimated parameters, are inherently geometric and dependent on the space in which they are measured. Thus, the MI, being purely informational, cannot be naturally aligned with task-specific error metrics and has no mechanism by which it may be targeted to a particular geometric distance on predictions: it implicitly assumes errors are measured by log loss. In turn, this inflexibility can be problematic for various applications. For example, in financial settings, we often inherently care about the variance in future returns and not the entropy, noting that the two can take arbitrary values with respect to one another.

The MI also poses several practical challenges (Rainforth et al., 2024). Foremost, it is a doubly intractable quantity, in general requiring a nested estimation (Rainforth et al., 2018) of either the posterior $p(\theta \mid y, d)$ or marginal $p(y \mid d)$. Second, in implicit settings, where the likelihood can only be sampled but not evaluated, the MI faces additional challenges due to its explicit reliance on densities. While there exist approaches for estimating the MI in implicit settings

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

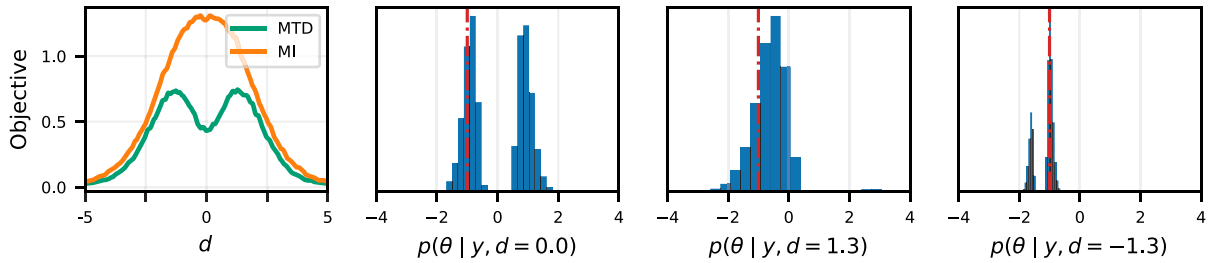

Figure 1: Comparison of MI and MTD on the 1D source location-finding problem, with the true source at $\theta = -1$ (dashed red line). MI is maximized at the origin, reflecting the prior mode, whereas MTD is maximized near $d \approx \pm 1.3$. The posterior for $d = 0$ is bimodal, while for $d = \pm 1.3$ it becomes unimodal or sharply concentrated. In practice, d is optimized directly, and MTD breaks the posterior symmetry even when initialized unfavorably.

(Kleinegesse and Gutmann, 2019; Ivanova et al., 2021), these amount to learning the unknown density ratio, a task that becomes increasingly difficult in high dimensions. This problem is also not specific to MI, with the core BED formalism inherently having nested dependence on the posterior (Lindley, 1972; Chaloner and Verdinelli, 1995).

We address these shortcomings by proposing the *mutual transport dependence (MTD)*, a novel class of geometric criteria for experimental design. The MTD measures the dependency between θ and y in terms of an optimal transport discrepancy (Villani et al., 2008; Feydy et al., 2019) between the joint distribution $p(\theta, y \mid d)$ and its product of marginals $p(\theta)p(y \mid d)$. The MTD depends explicitly on a choice of *sample-level* cost function, enabling practitioners to encode domain knowledge or downstream objectives directly into the design criterion. This cost function can be defined either directly or through a transformation of the underlying space, creating a family of flexible design criteria.

Moreover, the MTD offers practical benefits. It does not involve any nested expectations, can be estimated without likelihood evaluations, and can be directly optimized using gradient-based methods whenever differentiable sampling of $p(y \mid \theta, d)$ is possible. This makes it particularly well-suited to simulation-based or implicit scenarios where the likelihood $p(y \mid \theta, d)$ is unknown or intractable, an increasingly common setting in OED (Kleinegesse and Gutmann, 2019, 2021; Ivanova et al., 2021; Encinar et al., 2025),

Optimizing MTD results in qualitatively different design behaviour than MI (see Figure 1). We illustrate the effectiveness of MTD-optimal designs on standard experimental design benchmarks, comparing directly with MI-based designs. Our results show that the MTD can outperform traditional information-theoretic approaches, while also allowing experimenters to tailor the geometry of the underlying spaces to the problem at hand. In sum, our framework introduces a principled and practical new class of criteria for optimal experimental design, overcoming the rigidity of information-theoretic methods and enabling experiments that better reflect real-world objectives.

2 BACKGROUND AND NOTATION

We use $\theta \in \Theta$ to represent the unknown target quantity of interest that we wish to learn about through our experiments. This could correspond to a real world quantity, model parameters, or something abstract like downstream predictions. We further use $d \in \mathcal{D}$ to represent an experimental design and $y \in \mathcal{Y}$ to represent an observed outcome of an experiment. Akin to standard BED approaches, we specify a prior $p(\theta)$ representing our beliefs about θ before performing any experiments and a likelihood $p(y \mid \theta, d)$ capturing the data generating process. Our goal is now to select d in a way that will allow us to best estimate θ once the experiment's outcome is observed.

2.1 OED with Mutual Information

From an information-theoretic point of view, it is natural to seek designs which result in data y that reduces our uncertainty about the unknown quantity θ. That is, we consider the reduction in entropy (Lindley, 1956)

$$\mathcal{I}(d) = \mathbb{E}_{p(y|d)} \left[\mathsf{H}[\theta] - \mathsf{H}[\theta \mid y, d] \right] \tag{3}$$

which can straightforwardly be shown to yield the mutual information (1). The design choice is $d^* = \operatorname{argmax}_{d \in \mathcal{D}} \mathcal{I}(d)$, maximizing the mutual information. In the context of OED, $\mathcal{I}(d)$ is often called the *expected information gain* (EIG).

In all but the simplest cases, computing $\mathcal{I}(d)$ is non-trivial, as it requires estimating both the outer expectation and the integrand (i.e., either the posterior $p(\theta \mid y, d)$ or marginal likelihood $p(y \mid d)$). Often,

one resorts to nested estimators like nested Monte Carlo (NMC) (Rainforth et al., 2018), or variational approaches (Foster et al., 2019, 2020). Importantly, many techniques assume that the likelihood $p(y \mid \theta, d)$ is known explicitly, where the corresponding distribution can be evaluated pointwise.

OED is often particularly useful in adaptive scenarios, where experiments are designed sequentially based on the data $h_t = \{(d_k, y_k)\}_{k=1}^t$ gathered in previous trials. In this setting, we replace the prior $p(\theta)$ with our updated beliefs using the posterior $p(\theta \mid h_t)$, and consider the *incremental* MI

$$\mathcal{I}^{(t+1)}(d) = \mathsf{KL}\left[p(\theta, y|d, h_t)||p(\theta|h_t)p(y|d, h_t)\right]. \quad (4)$$

We refer to Chaloner and Verdinelli (1995); Rainforth et al. (2024) for more comprehensive surveys of BED.

2.2 Optimal Transport

Optimal transport (OT) is a mathematical toolkit which allows us to compare two arbitrary probability distributions $p(x)$ and $q(x')$ in terms of the amount of work required to transform one distribution into the other (Villani et al., 2008; Peyré et al., 2019).

In the Kantorovich formulation of OT (Kantorovich, 1942), we specify a non-negative cost function $c(x, x') \geq 0$ encoding the cost of transporting a unit of mass from location x to x'. A *coupling* is a joint distribution $\gamma(x, x')$ whose marginals are p, q respectively. We write $\Pi(p, q)$ for the set of all valid couplings. Given a coupling $\gamma \in \Pi(p, q)$, its associated cost is

$$K_c(\gamma) = \int c(x, x')\, \mathrm{d}\gamma(x, x') = \mathbb{E}_{\gamma(x,x')}\left[c(x, x')\right] \quad (5)$$

which can be interpreted as the average sample-level transport cost under this coupling. An optimal coupling minimizes this cost, and we write

$$\mathsf{OT}_c\left[p, q\right] = \min_{\gamma \in \Pi(\mu,\nu)} K_c(\gamma) \quad (6)$$

for the minimum value attained. In short, $c(x, x')$ defines a sample-level cost function and $\mathsf{OT}_c[p, q]$ is the associated optimal transport discrepancy.

3 MUTUAL TRANSPORT DEPENDENCE

The MI is an inherently density-based objective, where the value of an experiment is considered a purely informational quantity that is based directly on the *density* of the posterior, rather than the actual *values* that θ and y can take. As such, the MI is unable to incorporate properties of the underlying sample spaces $\Theta, \mathcal{Y}$ into its design objective, such as an error metric on θ. This induces strong invariance properties in the MI (Polyanskiy and Wu, 2025, Theorem 3.7), such that it remains fixed under injective transformations of θ and/or y. Thus, if we are interested in measuring errors in terms of some transformation $\phi = f(\theta)$, the optimal design remains fixed with changes in f, even though our intuitive notion of error itself will change. For example, while θ may be the natural variables for parametrizing a model, we may be interested in a one-to-one transformation of θ which is more interpretable. Under MI, these scenarios are indistinguishable: any unit of entropy reduction is equally valuable, regardless of its practical implications.

These shortcomings motivate the need for a practical objective which is fully defined in terms of *geometric* notions on the underlying sample spaces. In other words, we seek a criterion which is determined by the values taken on by random variables themselves.

3.1 Mutual Transport Dependence

One interpretation of MI is that it measures the KL divergence between the joint, $p(\theta, y|d)$, under which θ and y are dependent, and the product of marginals, $p(\theta)p(y|d)$, under which they are independent. The KL, though, depends only on density ratios and is thus unsuitable for a geometric measure of information.

To provide a geometric, sample-based criterion, we propose to instead measure the dependency between θ and y via an optimal transport dependency between the same joint and product of marginals. Optimal transport, relying on an expected sample-level cost function, is a natural choice for such a geometric measure as it allows customisation of our notion of distance through the cost function. This yields our proposed mutual transport dependence (MTD), defined as follows.[1]

Definition 1. *Given a cost function* $c : \Theta^2 \times \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$*, the* mutual transport dependence *(MTD) is*

$$\begin{aligned}\mathcal{T}_c(d) :&= \mathsf{OT}_c\left[p(\theta, y \mid d), p(\theta)p(y \mid d)\right] \\ &= \min_{\gamma \in \Pi(p(\theta,y|d), p(\theta)p(y|d))} \mathbb{E}_\gamma\left[c(\theta, y, \theta', y')\right]. \end{aligned} \quad (7)$$

Note that the MTD depends only on expectations with respect to our model, with no direct dependency on the density functions themselves. This is critically important from a computational and scaling perspective, as there is no need to perform (nested) estimation of the posterior or marginal likelihood densities, which is typically challenging, especially in high dimensions.

High values of $\mathcal{T}_c(d)$ indicate that the parameter and

[1] We work under standard measure-theoretic assumptions regarding the spaces $\Theta, \mathcal{Y}$ as well as the cost c, which we discuss in Appendix A.

data are strongly coupled under this design, and conversely, $\mathcal{T}_c(d) = 0$ if and only if θ and y are independent under d. As $\mathcal{T}_c(d)$ can be also seen as the average (θ, y) displacement under the optimal transport plan, measured by the cost $c(\theta, y, \theta', y')$, it is a *geometric* notion, and by choosing the cost function c in an appropriate way the experimenter can incorporate downstream preferences directly into the experimental design process.

3.2 Alternative Transport-Based Measures

The MI, being the KL between the joint and product of marginals, can be equivalently understood as an expected KL discrepancy on either Θ or $\mathcal{Y}$ alone:

$$\mathcal{I}(d) = \mathbb{E}_{p(\theta)}\mathsf{KL}\left[p(y \mid \theta, d) \mid\mid p(y \mid d)\right] \tag{8}$$
$$= \mathbb{E}_{p(y|d)}\mathsf{KL}\left[p(\theta \mid y, d) \mid\mid p(\theta)\right]. \tag{9}$$

This equivalence does not translate to the MTD, as the transport distances in Θ and $\mathcal{Y}$ are inherently different. However, we can derive transport-based analogues of these interpretations of the MI as well by considering costs defined on Θ or $\mathcal{Y}$ alone as follows.

Definition 2. *Given a cost function $c : \Theta^2 \to \mathbb{R}_{\geq 0}$, the* target transport dependence (TTD) *is*

$$\mathcal{T}_c^{(\theta)}(d) := \mathbb{E}_{p(y|d)}\left[\mathsf{OT}_c\left[p(\theta \mid y, d), p(\theta)\right]\right]. \tag{10}$$

Definition 3. *Given a cost function $c : \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$, the* expected data transport dependence (DTD) *is*

$$\mathcal{T}_c^{(y)}(d) := \mathbb{E}_{p(\theta)}\left[\mathsf{OT}_c\left[p(y \mid \theta, d), p(y \mid d)\right]\right]. \tag{11}$$

In some ways, these transport dependencies are perhaps more intuitive than the MTD, $\mathcal{T}_c(d)$, which relies on a cost defined on the joint space $\Theta \times \mathcal{Y}$. In particular, given that our ultimate goal is the estimation of θ, the TTD arguably provides the more natural of the three objectives as it is directly measuring changes in beliefs in the space of Θ.

However, unlike the MTD, these objectives have reintroduced a nesting into the problem as the maximization over couplings is now inside an expectation. This makes them, at least in principle, computationally more problematic as they appear to require solving an OT problem for any given sample of the outer expectation.

In practice, though, both $\mathcal{T}_c^{(\theta)}(d)$ and $\mathcal{T}_c^{(y)}(d)$, can be approximated in a way that only requires us to solve only a single OT problem. Leveraging recent work on conditional optimal transport (Carlier et al., 2016; Kerrigan et al., 2024; Chemseddine et al., 2024; Baptista et al., 2024), we show in the following result that when the cost function is given by a norm, these quantities are recovered as a limiting special case of $\mathcal{T}_c(d)$ under a particular choice of c, the proof for which is given in Appendix A.

Theorem 1. *For a fixed $1 \leq p < \infty$, consider the cost*

$$c_\eta(\theta, y, \theta', y') = \eta|\theta - \theta'|_\Theta^p + |y - y'|_\mathcal{Y}^p \tag{12}$$

for $\eta > 0$. Then, $\eta^{-1}\mathcal{T}_{c_\eta}(d) \to \mathcal{T}_{c_\Theta}^{(\theta)}(d)$ as $\eta \to 0^+$, where $c_\Theta(\theta, \theta') = |\theta - \theta'|_\Theta^p$. Similarly, for

$$c_\psi(\theta, y, \theta', y') = |\theta - \theta'|_\Theta^p + \psi|y - y'|_\mathcal{Y}^p \tag{13}$$

we have $\psi^{-1}\mathcal{T}_{c_\psi}(d) \to \mathcal{T}_{c_\mathcal{Y}}^{(y)}(d)$ as $\psi \to 0^+$ where $c_\mathcal{Y}(y, y') = |y - y'|_\mathcal{Y}^p$.

As the MTD can be defined using any choice of cost function, we can thus think of it as a more general objective which subsumes the TTD and DTD as limiting cases in the choice of cost function. Thus, in practice, we can implement these alternatives simply by using a weighted cost function (c_η or c_ψ) in the MTD with a small $\eta > 0$ or $\psi > 0$ for TTD and DTD respectively. We, therefore, focus our subsequent discussion on the MTD.

We emphasize that while we have introduced various notions of transport dependence in terms of designing a single experiment, all constructions readily generalize to the sequential setting by updating the prior as discussed in Section 2. While our experiments will focus on this sequential case, our notions of transport dependence can also be extended to the policy setting (Foster et al., 2021; Ivanova et al., 2021) by considering optimal transport over the entire experimental rollout.

3.3 Estimation and Optimization

To use the MTD in practice, we must be able to efficiently estimate and optimize $\mathcal{T}_c(d)$. Here, we briefly discuss the computational tools used later in Section 6, but note that other algorithmic approaches to leverage transport dependence for experimental design may be viable.

We emphasize that our estimators for the MTD are purely sample-based. Assuming that we can draw samples $\theta \sim p(\theta)$ and $y|\theta \sim p(y \mid \theta, d)$, we can estimate the $\mathcal{T}_c(d)$ without ever evaluating densities. By contrast, estimators for the MI either require direct access to the likelihood density, or rely on learning approximations for density ratios (Kleinegesse and Gutmann, 2020, 2021; Ivanova et al., 2021), introducing substantial extra modelling and optimization complexity.

Estimation. For discrete measures, the optimal transport problem becomes a linear program (LP) for which a wide range of numerical solvers have been proposed (Bonneel et al., 2011; Peyré et al., 2019). When the distributions $p(\theta, y \mid d)$ and $p(\theta)p(y \mid d)$ are known and discrete, the OT problem may be solved directly using these distributions.

In practice, however, our distributions are often continuous or only accessible through sampling. In such cases, we approximate them using empirical measures based on samples. That is, we sample $(\theta_j, y_j) \overset{\text{i.i.d.}}{\sim} p(\theta, y \mid d)$ and $(\theta'_k, y'_k) \overset{\text{i.i.d.}}{\sim} p(\theta)p(y \mid d)$, followed by the approximations

$$p(\theta, y \mid d) \approx \frac{1}{n}\sum_{j=1}^{n} \delta_{(\theta_j, y_j)}, \; p(\theta)p(y \mid d) \approx \frac{1}{n}\sum_{k=1}^{n} \delta_{(\theta'_k, y'_k)}. \tag{14}$$

This procedure yields the plug-in estimator (Boissard and Gouic, 2014; Fournier and Guillin, 2015)

$$\widehat{\mathcal{T}}_c(d) = \mathsf{OT}_c\left[\frac{1}{n}\sum_{j=1}^{n} \delta_{(\theta_j, y_j)}, \frac{1}{n}\sum_{k=1}^{n} \delta_{(\theta'_k, y'_k)}\right] \tag{15}$$

which is asymptotically consistent (Dudley, 1969) as $n \to \infty$ and concentrates around its mean at an exponential rate (Bolley et al., 2007; Boissard, 2011; Weed and Bach, 2019), albeit with a positive bias that decreases with n (Papp and Sherlock, 2025).

Optimisation. Optimizing $\mathcal{T}_c(d)$ poses a further challenge, assuming we cannot simply enumerate over possible designs. We therefore now show how $\mathcal{T}_c(d)$ can be optimized using stochastic gradient ascent provided the design space is continuous.

The only additional assumption needed for this is the existence of a differentiable reparameterization of y with respect to d. Namely, using the noise outsourcing lemma (Kallenberg, 1997), then for any fixed noise distribution with appropriate reference measure, $q(\eta)$, there exists (subject to extremely weak assumptions) a function h such that $y = h(\eta; \theta, d)$ for $\eta \sim q(\eta)$. If we further assume that $d \mapsto h(\eta; \theta, d)$ is differentiable for our chosen $q(\eta)$, then the LP approach above enables the calculation of $\nabla_d \widehat{\mathcal{T}}_c(d)$ via automatic differentiation. Hence, we may perform gradient-based design optimization in this setting. We refer to Peyré et al. (2019, Chapter 9) for a further discussion of the differentiability of optimal transport discrepancies.

Note that this differentiable reparameterisation assumption is the same as in implicit MI methods and is often satisfied even when evaluating $p(y \mid \theta, d)$ is itself intractable: many, if not most, intractable likelihood models are based on stochastic simulators, with the intractability coming from deterministic mappings of stochastic variables or stochastic differential equations (Cranmer et al., 2020).

4 COMPARISON AGAINST MUTUAL INFORMATION

We now analyse how our proposed transport-based criteria relate to the expected information gain (MI). Although the MTD and MI originate from different principles, there are interesting links between the two as we now show. In particular, under quadratic costs, transport dependencies can be upper-bounded by the MI.

Theorem 2. *Suppose the prior $p(\theta)$ is strictly log-concave, i.e., there exists some $\lambda_\theta > 0$ with $-\nabla^2 \log p(\theta) \succeq \lambda_\theta I$. For $c(\theta, \theta') = |\theta - \theta'|^2$, we have*

$$\lambda_\theta \mathcal{T}_c^{(\theta)}(d) \leq 2\mathcal{I}(d). \tag{16}$$

Similarly, if the marginal $p(y \mid d)$ is strictly log-concave with parameter $\lambda_{y|d}$, and $c(y, y') = |y - y'|^2$, then

$$\lambda_{y|d} \mathcal{T}_c^{(y)}(d) \leq 2\mathcal{I}(d). \tag{17}$$

When both the prior and likelihood satisfy these assumptions, under cost $c(\theta, \theta', y, y') = \eta|\theta - \theta'|^2 + |y - y'|^2$, for $\lambda = \max\{\lambda_\theta/\eta, \lambda_{y|d}\}$ we have $\lambda \mathcal{T}_c(d) \leq 2\mathcal{I}(d)$.

See Appendix B for a proof and extended discussion. We note that Theorem 2 holds for quadratic costs on any space, and in particular remains valid under any transformations of Θ and $\mathcal{Y}$ when the quadratic cost is computed in the new coordinates. Thus, if the MTD is large under *any* such transformation, the MI must necessarily also be large. This suggests a robustness to the selected cost function, as selecting for designs under a given particular cost ensures a minimum level in the MI.

We further derive a closed-form expression for the MTD for a linear-Gaussian model under quadratic costs. We allow for the possibility of the observation noise $\sigma^2_{d,\theta}$ to vary with design to demonstrate how the value of the MI diverges as the likelihood approaches a deterministic outcome. ($\sigma^2_{d,\theta} \to 0$). In contrast, $\mathcal{T}_c(d)$ is bounded for all designs d and all noise $\sigma^2_{d,\theta}$, making the MTD a quantitative and stable measure of dependence even in scenarios approaching determinism.

Theorem 3. *Suppose $\theta \in \Theta = \mathbb{R}^n$ has a standard normal prior $p(\theta) = \mathcal{N}(0, I_n)$, designs are vectors $d \in \mathcal{D} = \mathbb{R}^n$, and $y \in \mathcal{Y} = \mathbb{R}$ has likelihood $p(y \mid \theta, d) = \mathcal{N}\left(\langle d, \theta\rangle, \sigma^2_{d,\theta}\right)$. Under the quadratic cost, we have*

$$\mathcal{T}_c(d) = 2\Bigg(1 + \sigma^2_{d,\theta} + |d|^2 - \sqrt{1 + (|d|^2 + \sigma^2_{d,\theta})^2 + 2\sqrt{|d|^2 + \sigma^2_{d,\theta}}}\Bigg). \tag{18}$$

Moreover, $\mathcal{T}_c(d) \leq 2$. On the other hand, the MI is

$$\mathcal{I}(d) = \frac{1}{2}\log\left(1 + |d|^2/\sigma^2_{d,\theta}\right) \tag{19}$$

which is unbounded as $\sigma^2_{d,\theta} \to 0$.

Table 1: Metrics for the CES model after $T = 10$ design iterations, averaged over 50 random seeds ($\pm$ one standard error). Designs produced by the MTD yield lower RMSEs than PCE on average for all parameters.

	ρ	α	u	σ	β	τ
Random	$0.251_{\pm 0.025}$	$0.116_{\pm 0.016}$	$36.365_{\pm 7.341}$	$317.219_{\pm 81.866}$	$0.727_{\pm 0.088}$	$0.740_{\pm 0.106}$
PCE	$0.047_{\pm 0.012}$	$0.036_{\pm 0.013}$	$8.902_{\pm 5.749}$	$24.942_{\pm 15.697}$	$0.201_{\pm 0.060}$	$0.100_{\pm 0.048}$
MTD	$0.018_{\pm 0.005}$	$0.009_{\pm 0.001}$	$3.671_{\pm 2.810}$	$1.767_{\pm 1.009}$	$0.058_{\pm 0.008}$	$0.049_{\pm 0.012}$
MTD ($\mathcal{T}_{c_\dagger}$)	$0.022_{\pm 0.008}$	$0.012_{\pm 0.004}$	$10.941_{\pm 10.107}$	$0.534_{\pm 0.195}$	$0.069_{\pm 0.013}$	$0.053_{\pm 0.013}$

5 RELATED WORK

Classical optimal experimental design criteria trace back to frequentist approaches based on the Fisher information matrix (Fisher, 1935; Wald, 1943; Kiefer, 1959, 1974; Pukelsheim, 2006). While powerful in some settings, these methods often rely on asymptotic approximations and are limited when models are nonlinear (Ryan et al., 2016; Rainforth et al., 2024). As they depend only on local (second-order) information about θ, they can lose fidelity compared to criteria based on the full joint distribution $p(\theta, y \mid d)$.

Bayesian experimental design (BED) addresses many of these limitations by evaluating designs using objectives that can be viewed as measuring expected reduction in uncertainty on θ (DeGroot, 1962; Dawid, 1998; Bickford Smith et al., 2025). Here this uncertainty is typically measured using (differential) entropy to produce the MI or expected information gain (Lindley, 1956), especially in the contemporary literature (Huan and Marzouk, 2014; Foster, 2021; Foster et al., 2021; Ao and Li, 2024; Iollo et al., 2024b). The trace or determinant of the posterior covariance matrix have also occasionally be used instead (Vanlier et al., 2012; Ryan et al., 2016; Huan et al., 2024), but this requires expensive nested inference procedures to be performed that are typically even more costly than MI optimization.

Concurrent work by Helin et al. (2025) also studies the target transport dependence $\mathcal{T}_c^{(\theta)}(d)$ under the specific choice $c(\theta, \theta') = |\theta - \theta'|^p$, $p \in [1, \infty)$, as an objective for experimental design. In light of Theorem 1, this can be seen as a limiting case of our more general MTD criterion under a Euclidean cost assumption. Their work is primarily theoretical with no quantitative comparisons against the MI, and they do not propose a practical method for optimizing the TTD and in particular overcoming its double intractability. Our work, by contrast, develops a sample-based, differentiable framework applicable for general costs and empirically demonstrates its efficacy for sequential OED.

Optimal-transport based notions of statistical dependency have also been considered in areas such as representation learning (Ozair et al., 2019) independence testing (Warren, 2021; Wiesel, 2022; Nies et al., 2025), and fairness (Leteno et al., 2023). These works, however, are not concerned with experimental design and also focus exclusively on Euclidean costs. Our work adds to this growing literature of geometric dependency measures by introducing the mutual transport dependence for OED under general cost functions.

6 EXPERIMENTS

We now evaluate the proposed methodology on both standard benchmark experimental design tasks and variations on these that have particular error desiderata in our final estimates. In each setting, we use the MTD as the design criterion, sequentially selecting experiments by optimizing $\mathcal{T}_c(d)$ as explained in Section 3.3. After each design is chosen, we perform posterior inference over θ and proceed to the next experimental iteration. All results are reported over either 25 or 50 random seeds, where each random seed constitutes a different ground-truth value of θ. We compare against the MI throughout, using PCE (Foster et al., 2019) as a well-known estimator for this quantity. We note that unlike our MTD approach, PCE is an explicit estimator that requires direct access to the likelihood density, thereby providing a stronger baseline than more directly comparable, but also more complex, implicit MI approaches. See Appendix D for details.[2]

CES. The first problem we consider is Constant Elasticity of Substitution (CES) (Arrow et al., 1961; Foster et al., 2020; Blau et al., 2022; Iollo et al., 2024b; Hedman et al., 2025), arising from behavioral economics. In this problem, a participant compares two baskets $d_1, d_2 \in [0, 100]^3$ consisting of various amounts of three different goods. Given two baskets, the participant provides a scalar response $y \in [0, 1]$ indicating their subjective preference between the baskets. The design variable $d = (d_1, d_2)$ is thus six-dimensional, and the goal is to recover the latent parameters $\theta = (\rho, \alpha_1, \alpha_2, \alpha_3, u) \in \mathbb{R}^5$ governing the participant's preferences. This is a particularly challenging design problem, as large regions of the design space result in uninformative outcomes $y \in \{0, 1\}$. We sequentially design $T = 10$ experiment iterations.

[2]Experiment code: github.com/GavinKerrigan/mtd

Source Location Finding. Our second problem is source location finding (LF) (Sheng and Hu, 2005; Foster et al., 2021; Ivanova et al., 2021; Blau et al., 2022; Iollo et al., 2024a). In this, our goal is to estimate the spatial location of two sources $\theta_1, \theta_2 \in \mathbb{R}^2$. Each source emits a signal which decays according to an inverse square law. At each experiment iteration, a sensor is placed at a location $d \in \mathbb{R}^2$ which records a noisy measurement $y \in \mathbb{R}$ of the total signal intensity at the sensor location. Here, we design $T = 25$ experiments.

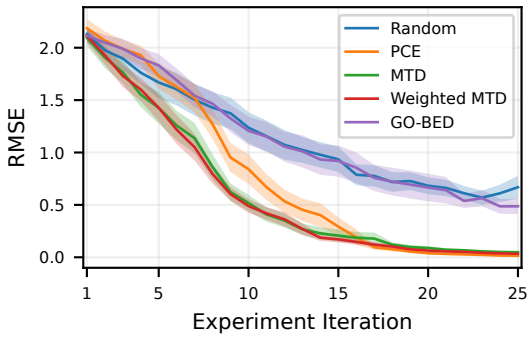

Figure 2: RMSE between posterior θ samples and ground-truth on the location finding problem, averaged over 25 seeds ($\pm$ one standard error).

6.1 MTD under Euclidean Costs

While one of the main appeals of the MTD is that it allows for flexible cost functions, in this section we first consider the quadratic cost $c(\theta, y, \theta', y') = |\theta - \theta'|^2 + |y - y'|^2$ as a reasonable default choice. Our first set of experiments demonstrates that even under this default setting, MTD-optimal designs can exceed the performance of MI-optimal designs in terms of recovering an unknown parameter.

In Table 1, we evaluate the MTD on the CES problem in terms of the final RMSE between posterior samples after $T = 10$ experiment iterations and the true value of θ. Designs produced by optimizing MTD achieve lower RMSEs than those produced by optimizing MI.

Similarly, in Figure 2, we plot the RMSE between posterior samples and the true θ value on the LF problem over the course of $T = 25$ design iterations. We observe that the MTD yields lower RMSEs throughout most of the iterations, but designs produced by optimizing the MI yield similar RMSEs at the final iteration.

For the LF problem, both MI and MTD are optimized using five random restarts, i.e., we generate five candidate designs and retain the best under the given objective. This approach serves not only to mitigate sensitivity to initialization but also to systematically improve design quality. In particular, for later iterations of the LF problem, where the posterior over θ becomes highly concentrated, the restart strategy provides a simple yet effective mechanism to ensure robustness against poor initializations.

In terms of runtime, for either problem optimizing a single design under MTD requires approximately 30 seconds of wall-clock time, whereas optimizing the same design with PCE takes roughly 120 seconds, with both objectives run to convergence. While these runtimes are sensitive to implementation choices and could likely be reduced through more careful tuning or normalization of compute budgets, the key observation is that MTD is comparably fast to previous approaches and potentially faster.

6.2 MTD under Transformations

While we may explicitly specify a cost for MTD, costs can also be defined *implicitly* through transformations of the underlying sample spaces. Concretely, if $f : \Theta \to \Theta^\dagger, g : \mathcal{Y} \to \mathcal{Y}^\dagger$ are two transformations, we may define $c_\dagger(\theta, \theta', y, y') = |f(\theta) - f(\theta')|^2 + |g(y) - g(y')|^2$, i.e., quadratic cost in transformed coordinates. This is useful when we wish to measure errors in a particular space. Note that the MI between $f(\theta)$ and $g(y)$ is equal to that between θ and y if f and g are injective, prohibiting the same trick from being meaningfully employed.

We illustrate this on CES using the transformations

$$\sigma = 1/(1-\rho) \qquad \beta_i = \log(\alpha_i)/g(\alpha) \qquad \tau = \log(u) \tag{20}$$

where $g(\alpha)$ is the geometric mean of α. These transformations are interpretable: σ is the elasticity (Arrow et al., 1961), β the centered-log-ratio of α, capturing relative importance of goods, and $\tau = \log u$ a natural reparametrization under its lognormal prior.

To evaluate this approach, we generate designs using $\mathcal{T}_{c_\dagger}(d)$ in the transformed variables, implicitly altering the cost. Table 1 reports posterior RMSEs for PCE, MTD on the original scale $\mathcal{T}_c(d)$, and MTD with the transformed cost $\mathcal{T}_{c_\dagger}(d)$. On the original parameters, $\mathcal{T}_{c_\dagger}(d)$ performs comparably to $\mathcal{T}_c(d)$ for ρ and α but somewhat worse for u, suggesting the untransformed version remains preferable when evaluation is performed directly on the original parameters.

On the transformed scale, $\mathcal{T}_{c_\dagger}(d)$ and $\mathcal{T}_c(d)$ perform similarly for β and τ. However, there are clearer differences in σ. In particular, we see that PCE exhibits high RMSE in σ. This is because PCE occasionally yields poor designs which are unable to identify that $\rho \neq 1$, leading to high errors in $\sigma = (1-\rho)^{-1}$. $\mathcal{T}_c(d)$ generally yields higher quality designs which reliably identify ρ and thus obtain low errors in σ. The transformed

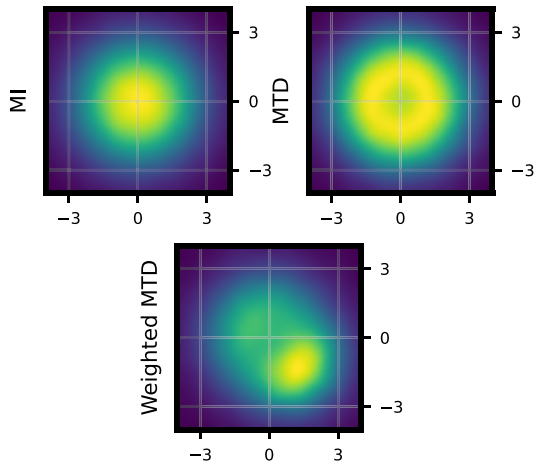

Figure 3: Estimated values of the EIG (via PCE) and the MTD for the 2D location-finding model with two sources. *Top row*: the EIG is maximized at the origin, whereas the MTD favors off-center designs, illustrating its symmetry-breaking behavior. *Bottom row*: with additional weighting of the cost function, the MTD can be tuned to favor specific regions of the design space.

$\mathcal{T}_{c_\dagger}(d)$, though, implicitly upweights designs where σ is large, leading to low RMSE values. Notably, there is no natural analogue of changing PCE to target RMSE in σ directly. Overall, this provides evidence that the MTD can be tailored to particular downstream metrics.

6.3 Weighted Cost Functions

We next evaluate the MTD on a variation of the LF problem which highlights its ability to incorporate downstream objectives through an appropriate choice of cost function. Here, the goal is not only to localize the sources, but also to rapidly determine whether a source lies in a critical region $\mathcal{R} \subset \Theta$.

To capture this preference, we define a weighted cost $c_w(\theta, y, \theta', y') = w(\theta)\left(|\theta - \theta'|^2 + |y - y'|^2\right)$ where $(\theta, y) \sim p(\theta, y \mid d)$ and $(\theta', y') \sim p(\theta')p(y' \mid d)$ and the weight is given by $w(\theta) = b + \sum_{k=1}^{2} g(\theta_k - \mu)$, with $b > 0$ a bias and g a bump function supported on $\mathcal{R}$, a ball of radius 1.5 centered at $\mu = (1.5, -1.5)$. See Appendix D for details. Intuitively, the cost is up-weighted whenever the "true" θ has a source in $\mathcal{R}$. In such cases, the MTD $\mathcal{T}_{c_w}(d)$ increases, thus yielding designs that prioritize detecting whether a source is present in $\mathcal{R}$. We stress that this represents only one possible weighting scheme, and other choices could be used to encode different downstream preferences.

In Figure 3, we plot $\mathcal{T}_{c_w}(d)$ under the weighted cost for the LF task. As intended, the objective is upweighted in

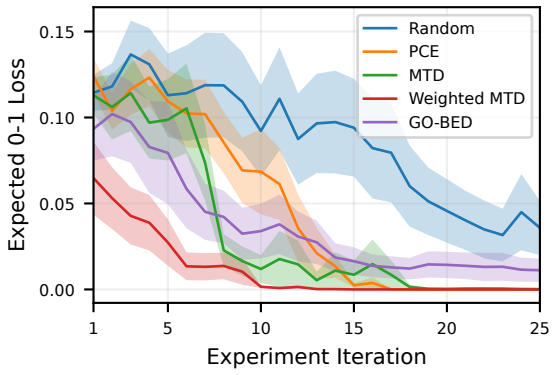

Figure 4: Expected zero-one loss for the 2D location-finding model. The weighted MTD rapidly determines whether $\theta \in \mathcal{R}$.

$\mathcal{R}$, encouraging designs to be placed in this region. We also note that the unweighted MTD (with the quadratic cost) exhibits a symmetry breaking behavior, whereas the MI favors designs at the origin, thereby preserving symmetry after posterior updates (as in Figure 1).

We then design a sequence of $T = 25$ experiments using the weighted cost c_w. In Figure 4, we plot the mean zero-one loss of $\theta \in \mathcal{R}$, i.e., $\mathbb{E}_{p(\theta|h_t)}[|\mathbf{1}[\theta_{\text{true}} \in \mathcal{R}] - \mathbf{1}[\theta \in \mathcal{R}]|]$ which directly measures if we have detected $\theta \in \mathcal{R}$. As an additional natural baseline, we compare against the goal-oriented BED (GO-BED) method of Zhong et al. (2026), which directly targets the expected information gain in $\mathbf{1}[\theta \in \mathcal{R}]$.

While PCE and the unweighted MTD eventually determine if $\theta \in \mathcal{R}$, the weighted MTD achieves a much faster reduction in the zero-one loss. Further, GO-BED is initially more successful than standard MI at determining if $\theta \in R$, but it does maintain this advantage. We believe this behavior stems from a signal-to-noise ratio issue. GO-BED's signal-to-noise ratio decays rapidly over iterations as GO-BED is targeting mutual information in a discrete variable, causing the signal to vanish as the posterior concentrates, thereby making effective design selection increasingly difficult.

We also emphasize that the MI cannot be easily adapted to the task of first identifying if $\theta \in \mathcal{R}$ before exploring the rest of the space. Despite GO-BED's relative success in determining if $\theta \in \mathcal{R}$, its performance in determining the precise source locations (as demonstrated by the RMSEs in Figure 2) is no better than random. In contrast, our framework's flexibility to encode task-specific preferences via the cost allows us to solve this task, where Figure 2 shows the RMSE under $\mathcal{T}_{c_w}$ matches that of $\mathcal{T}_c(d)$, confirming that we have not sacrificed performance in terms of identifying θ for this auxiliary objective.

6.4 Active Learning

As a final illustration of how to select the cost function in the MTD, we consider an active learning problem. We aim to perform regression on a positive function $f(d) > 0$ using a Gaussian Process (GP), emphasizing high accuracy when $f(d)$ is large. Design variables d correspond to inputs at which we observe $f(d)$, plus observation noise. This setting arises in a variety of real-world applications, including air pollution mapping (Patel et al., 2022) and protein property prediction (Notin et al., 2024). To enforce non-negativity, we model $y = \log f(d)$ with a GP. We evaluate our method on a synthetic one-dimensional dataset where f consists of a mixture of Gaussian bumps.

In this experiment, our objective is not to learn the GP itself, but rather to make accurate predictions y_* at test inputs $d_* \sim p_*(d_*)$ drawn from a target distribution $p_*(d_*)$. We thus apply the MTD with $\theta = y_*$, resulting in a transport-based generalisation of the expected predicted information gain (EPIG) (Smith et al., 2023). See Appendix Appendix D.3 for additional details on our task and objective.

Importantly, errors are evaluated on the original scale of the function, i.e., in terms of $\exp(y)$, rather than on the logarithmic scale. As this is an invertible transformation, MI-based approaches are invariant to it and cannot account for this distinction. In contrast, the MTD allows us to explicitly incorporate the relevant error scale through the choice of cost function. In particular, we consider the squared error on the original scale, $|\exp(y) - \exp(y')|^2$, which places greater emphasis on accuracy when $f(d)$ is large. This mirrors the transformed cost construction discussed in Section 6.2. Since f is a mixture of Gaussian bumps, this effectively prioritizes accurate modelling near the peaks of the bumps.

Figure 5 reports average RMSEs on the original scale (i.e., in terms of $\exp(y)$) as a function of acquired labels T. The MTD with the transformed cost consistently outperforms MI, as the cost function is aligned with the downstream loss. For comparison, we include MTD with the Euclidean cost $|y - y'|^2$ (denoted MTD (log)) which underperforms. This is expected, since this cost targets points with large absolute y values, corresponding to regions where $f(d) \approx 0$ on the original scale, thereby actively working against the intended objective. This illustrates the importance of choosing an appropriate cost. We also compare against EPIG (Smith et al., 2023), which is itself based on the MI, and so has similar invariance properties. EPIG exhibits space-filling behaviour on this problem, leading to RMSEs that decay slowly early on until it eventually explores regions where $f(d)$ is large.

Overall, these results highlight how the choice of cost function in the MTD enables practitioners to align the objective with their downstream goals. While our focus here is not on developing state-of-the-art active learning methods, this example shows the practical flexibility of our framework, and suggests promising directions for its application to active learning problems.

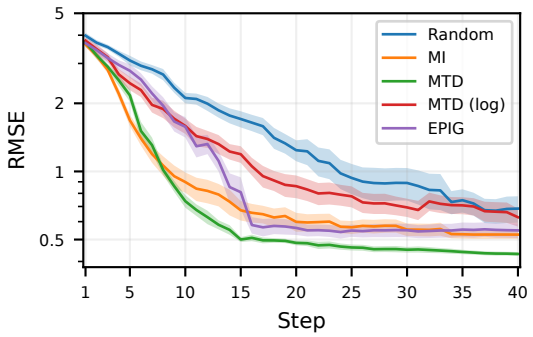

Figure 5: Average RMSE on the original scale (i.e., $\exp(y)$) for the active regression task, averaged over 25 seeds ($\pm$ one standard error). MTD with the transformed cost outperforms both MI and MTD (log), which uses a cost function that is misaligned with the downstream evaluation metric.

7 CONCLUSION

We introduce the mutual transport dependence (MTD), a novel class of geometric criteria for optimal experimental design. By quantifying the value of an experiment through an optimal transport divergence with an explicit sample-level cost, the MTD allows us to encode domain knowledge and task-specific objectives directly into the design criterion. We show that optimizing the MTD produces highly effective designs on standard benchmarks, and that tailoring the cost function enables alignment with particular experimental goals. Overall, the MTD offers a flexible, geometry-aware objective for OED, providing a practical tool for designing experiments that reflect both statistical dependence and the experimenter's real-world priorities.

Limitations While we have demonstrated several strategies for selecting the cost function, this choice ultimately remains problem-dependent and must be specified by the practitioner. Developing principled guidelines for choosing cost functions in different settings, e.g., promoting robustness to model misspecification, remains an important direction for future work. Additionally, our focus has been on sequential experimental design. Further improvements may be possible by integrating the MTD with policy-based approaches, which we leave to future investigation.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback in improving this work. GK and TR are supported by the UK EPSRC grant EP/Y037200/1.

References

Ambrosio, L. and Gigli, N. (2012). A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle*, pages 1–155. Springer.

Ao, Z. and Li, J. (2024). On estimating the gradient of the expected information gain in Bayesian experimental design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20311–20319.

Arrow, K. J., Chenery, H. B., Minhas, B. S., and Solow, R. M. (1961). Capital-labor substitution and economic efficiency. *The review of Economics and Statistics*, pages 225–250.

Baptista, R., Pooladian, A.-A., Brennan, M., Marzouk, Y., and Niles-Weed, J. (2024). Conditional simulation via entropic optimal transport: Toward non-parametric estimation of conditional Brenier maps. *arXiv preprint arXiv:2411.07154*.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, pages 686–690.

Bertsekas, D. P. (1971). *Control of uncertain systems with a set-membership description of the uncertainty*. PhD thesis, Massachusetts Institute of Technology.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Bickford Smith, F., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., and Rainforth, T. (2025). Rethinking aleatoric and epistemic uncertainty. In *Forty-second International Conference on Machine Learning*.

Blau, T., Bonilla, E. V., Chades, I., and Dezfouli, A. (2022). Optimizing sequential experimental design with deep reinforcement learning. In *International Conference on Machine Learning*, pages 2107–2128. PMLR.

Blower, G. (2003). The Gaussian isoperimetric inequality and transportation. *Positivity*, 7(3):203–224.

Boissard, E. (2011). Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance. *Electronic Journal of Probability*, 16:2296 – 2333.

Boissard, E. and Gouic, T. L. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539 – 563.

Bolley, F., Guillin, A., and Villani, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3):541–593.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. (2011). Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Carlier, G., Chernozhukov, V., and Galichon, A. (2016). Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44(3):1165 – 1192.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical science*, pages 273–304.

Chemseddine, J., Hagemann, P., Steidl, G., and Wald, C. (2024). Conditional Wasserstein distances with applications in Bayesian OT flow matching. *arXiv preprint arXiv:2403.18705*.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungaria*, 2:299–318.

Danskin, J. (1967). *The Theory of Max-min and Its Applications to Weapons Allocation Problems*. Econometrics and operations research. Springer.

Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. *Department of Statistical Science, University College London. http://www. ucl. ac. uk/Stats/research/abs94. html, Tech. Rep*, 139.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419.

Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer.

Dudley, R. M. (1969). The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

Encinar, P. C., Schröder, T., Yatsyshin, P., and Duncan, A. B. (2025). Deep optimal sensor placement for black box stochastic simulations. In *Frontiers in Probabilistic Inference: Learning meets Sampling*.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 2681–2690. PMLR.

Fisher, R. A. (1935). *The Design of Experiments*. The Design of Experiments. Oliver and Boyd.

Foster, A., Ivanova, D. R., Malik, I., and Rainforth, T. (2021). Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning*, pages 3384–3395. PMLR.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. (2019). Variational Bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, 32.

Foster, A., Jankowiak, M., O'Meara, M., Teh, Y. W., and Rainforth, T. (2020). A unified stochastic gradient approach to designing Bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pages 2959–2969. PMLR.

Foster, A. E. (2021). *Variational, Monte Carlo and policy-based approaches to Bayesian experimental design*. PhD thesis, University of Oxford.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738.

Gozlan, N. and Léonard, C. (2010). Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*.

Hedman, M., Ivanova, D. R., Guan, C., and Rainforth, T. (2025). Step-dad: Semi-amortized policy-based bayesian experimental design. *arXiv preprint arXiv:2507.14057*.

Helin, T., Marzouk, Y., and Rojo-Garcia, J. R. (2025). Bayesian optimal experimental design with Wasserstein information criteria. *arXiv preprint arXiv:2504.10092*.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hosseini, B., Hsu, A. W., and Taghvaei, A. (2025). Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338.

Huan, X., Jagalur, J., and Marzouk, Y. (2024). Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840.

Huan, X. and Marzouk, Y. M. (2014). Gradient-based stochastic optimization methods in Bayesian experimental design. *International Journal for Uncertainty Quantification*, 4(6).

Iollo, J., Heinkelé, C., Alliez, P., and Forbes, F. (2024a). Bayesian experimental design via contrastive diffusions. *arXiv preprint arXiv:2410.11826*.

Iollo, J., Heinkelé, C., Alliez, P., and Forbes, F. (2024b). PASOA - particle based Bayesian optimal adaptive design. *arXiv preprint arXiv:2402.07160*.

Ivanova, D. R., Foster, A., Kleinegesse, S., Gutmann, M. U., and Rainforth, T. (2021). Implicit deep adaptive design: Policy-based experimental design without likelihoods. *Advances in Neural Information Processing Systems*, 34:25785–25798.

Kallenberg, O. (1997). *Foundations of modern probability*. Springer.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Kerrigan, G., Migliorini, G., and Smyth, P. (2024). Dynamic conditional optimal transport through simulation-free flows. *Advances in Neural Information Processing Systems*, 37:93602–93642.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, pages 849–879.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinegesse, S. and Gutmann, M. U. (2019). Efficient Bayesian experimental design for implicit models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 476–485. PMLR.

Kleinegesse, S. and Gutmann, M. U. (2020). Bayesian experimental design for implicit models by mutual information neural estimation. In *International Conference on Machine Learning*, pages 5316–5326. PMLR.

Kleinegesse, S. and Gutmann, M. U. (2021). Gradient-based Bayesian experimental design for implicit models using mutual information lower bounds. *arXiv preprint arXiv:2105.04379*.

Kuhfeld, W. F., Tobias, R. D., and Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, 31(4):545–557.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE transactions on Information Theory*, 13(1):126–127.

Leteno, T., Gourru, A., Laclau, C., Emonet, R., and Gravier, C. (2023). Fair text classification with Wasserstein independence. *arXiv preprint arXiv:2311.12689*.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

Lindley, D. V. (1972). *Bayesian statistics: A review*. SIAM.

MacKay (1992). Information-based objective functions for active data selection. *Neural Computation*.

Massart, P. (2007). *Concentration inequalities and model selection*. Springer.

Melendez, J., Furnstahl, R., Grießhammer, H., McGovern, J., Phillips, D., and Pratola, M. (2021). Designing optimal experiments: An application to proton Compton scattering. *The European Physical Journal A*, 57(3):81.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

Nies, T. G., Staudt, T., and Munk, A. (2025). Transport dependency: Optimal transport based dependency measures. *The Annals of Applied Probability*, 35(4):2292 – 2362.

Notin, P., Rollins, N., Gal, Y., Sander, C., and Marks, D. (2024). Machine learning for functional protein design. *Nature Biotechnology*, 42(2):216–228.

Ozair, S., Lynch, C., Bengio, Y., Van den Oord, A., Levine, S., and Sermanet, P. (2019). Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32.

Papp, T. and Sherlock, C. (2025). Centered plug-in estimation of Wasserstein distances. *arXiv preprint arXiv:2203.11627*.

Park, M., Nassar, M., and Vikalo, H. (2013). Bayesian active learning for drug combinations. *IEEE transactions on biomedical engineering*, 60(11):3248–3255.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Patel, Z. B., Purohit, P., Patel, H. M., Sahni, S., and Batra, N. (2022). Accurate and scalable Gaussian processes for fine-grained air quality inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12080–12088.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Pinsker, M. S. (1964). Information and information stability of random variables and processes. *Holden-Day*.

Polyanskiy, Y. and Wu, Y. (2025). *Information theory: From coding to learning*. Cambridge university press.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. (2018). On nesting Monte Carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. (2024). Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84(1):128–154.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Sebastiani, P. and Wynn, H. P. (2000). Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157.

Sheng, X. and Hu, Y.-H. (2005). Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE transactions on signal processing*, 53(1):44–53.

Smith, F. B., Kirsch, A., Farquhar, S., Gal, Y., Foster, A., and Rainforth, T. (2023). Prediction-oriented bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7331–7348. PMLR.

Vanlier, J., Tiemann, C. A., Hilbers, P. A., and van Riel, N. A. (2012). A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142.

Villani, C. et al. (2008). *Optimal transport: Old and new*, volume 338. Springer.

Wald, A. (1943). On the efficient design of statistical investigations. *The Annals of Mathematical Statistics*, 14(2):134–140.

Warren, A. (2021). Wasserstein conditional independence testing. *arXiv preprint arXiv:2107.14184*.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

Wiesel, J. C. (2022). Measuring association with Wasserstein distances. *Bernoulli*, 28(4):2816 – 2832.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Zhong, S., Shen, W., Catanach, T., and Huan, X. (2026). Goal-oriented Bayesian optimal experimental design for nonlinear models using Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 14(1):19–47.

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; Section 3 and Section 4 contain the necessary mathematical and algorithmic details.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes; Section 3 discusses the time complexity of estimating the MTD.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Source code has been released with the camera-ready version of the paper and a link provided within this manuscript.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes; all theoretical claims have their assumptions stated.
   (b) Complete proofs of all theoretical results. Yes; all proofs are included in the appendix.
   (c) Clear explanations of any assumptions. Yes; any non-standard assumptions are explained.

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; details for the experiments are described in the Appendix. Code is provided.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes; details for the experiments are described in the appendix.
   (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; experimental settings are described in Section 6 and the Appendix.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes; provided in the Appendix.

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. Yes; citations are appropriately provided throughout.
   (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; we do not provide any new assets.
   (d) Information about consent from data providers/curators. Not Applicable;
   (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 Geometric Approach to Optimal Experimental Design: Supplementary Materials

A TRANSPORT DEPENDENCIES

In this section, we provide a more formal discussion of the MTD. We work under standard assumptions throughout, which are sufficient for guaranteeing that a solution to the OT problem exists.

A1. *The spaces Θ and $\mathcal{Y}$ are Polish spaces.*

A2. *All cost functions are lower-semicontinuous and non-negative.*

Under Assumptions A1-A2, minimizers to the OT problem in Equation (6) defined on either Θ or $\mathcal{Y}$ are guaranteed to exist (Ambrosio and Gigli, 2012, Theorem 1.5). Similarly, if $\Theta, \mathcal{Y}$ are Polish, then $\Theta \times \mathcal{Y}$ is Polish when equipped with the product topology, so that minimizers to an OT problem on the product spaces also exist under Assumptions A1-A2. Additional assumptions on c are necessary, though, to guarantee that $\mathcal{T}_c(d)$ (and the DTD/TTD) is finite. A trivially sufficient condition is that c is bounded from above. Other sufficient conditions can be given under moment assumptions of the corresponding densities. See Lemma 1 and Theorem 4 below.

As discussed in the main paper, the TTD and DTD are closely related to notions arising from conditional optimal transport (Carlier et al., 2016; Hosseini et al., 2025; Kerrigan et al., 2024; Chemseddine et al., 2024; Baptista et al., 2024). Viewing our transport dependencies from this lens is a fruitful avenue for theoretical analysis. We begin by recalling the notion of a triangular coupling (Hosseini et al., 2025), which gives a notion of couplings that fix certain variables.

Definition 4 (Triangular Couplings)**.** *A coupling $\gamma \in \Pi(p(\theta, y \mid d), p(\theta)p(y \mid d)$ is said to be $\mathcal{Y}$-triangular if draws $(\theta, y, \theta', y') \sim \gamma$ are such that $y = y'$ almost surely. Similarly, γ is said to be Θ-triangular if draws $(\theta, y, \theta', y') \sim \gamma$ are such that $\theta = \theta'$ almost surely.*

For the sake of brevity, we will write $\Pi := \Pi(p(\theta, y \mid d), p(\theta)p(y \mid d)$ for the set of all couplings, $\Pi_{\mathcal{Y}} := \Pi_{\mathcal{Y}}(p(\theta, y \mid d), p(\theta)p(y \mid d))$ for the set of $\mathcal{Y}$-triangular couplings, and $\Pi_{\Theta} := \Pi_{\Theta}(p(\theta, y \mid d), p(\theta)p(y \mid d))$ for the set of Θ-triangular couplings.

We begin with a lemma which allows us to bound the MTD by the DTD and TTD. This result shows that if the MTD is large, then both corresponding transport divergences on Θ or $\mathcal{Y}$ must also be large. This is particularly interpretable in terms of the TTD $\mathcal{T}_c^{(\theta)}(d)$, where we see that large MTD implies that there is a large transport divergence between the posterior and prior, on average across the marginal $p(y \mid d)$.

Lemma 1. *Fix $p \in [1, \infty)$. Suppose $\Theta, \mathcal{Y}$ are separable Hilbert spaces. Consider the cost function $c(\theta, y, \theta', y') = \eta|\theta - \theta'|^p + |y - y'|^p$ for a given $\eta > 0$. Write $c_\Theta(\theta, \theta') = |\theta - \theta'|^p$ and $c_\mathcal{Y}(y, y') = |y - y'|^p$. Assume that $p(\theta, y \mid d)$ has finite pth moment for a given $d \in \mathcal{D}$. Then,*

$$\mathcal{T}_c(d) \leq \eta \mathcal{T}_{c_\Theta}^{(\theta)}(d) \qquad \text{and} \qquad \mathcal{T}_c(d) \leq \mathcal{T}_{c_\mathcal{Y}}^{(y)}(d). \tag{21}$$

Furthermore, both $\mathcal{T}_{c_\Theta}^{(\theta)}(d)$ and $\mathcal{T}_{c_\mathcal{Y}}^{(y)}(d)$ are finite.

Proof. First consider the $\eta = 1$ case. Observe that $p(\theta, y \mid d)$ having finite pth moments immediately implies that both $p(\theta)$ and $p(y \mid d)$ also have finite pth moments. Note further that $p(\theta, y \mid d)$ and $p(\theta)p(y \mid d)$ have the same marginals in both Θ and $\mathcal{Y}$ space. It follows that both $\mathcal{T}_{c_\Theta}^{(\theta)}(d)$ and $\mathcal{T}_{c_\mathcal{Y}}^{(y)}$ are pth powers of conditional Wasserstein metrics (Kerrigan et al., 2024, Definition 2). By Kerrigan et al. (2024, Prop. 2), both $\mathcal{T}_{c_\Theta}^{(\theta)}(d)$ and $\mathcal{T}_{c_\mathcal{Y}}^{(y)}(d)$ are finite, and furthermore conditional Wasserstein metrics upper bound the Wasserstein metric on the corresponding joint measure. This proves the claim for $\eta = 1$.

For the general $\eta > 0$ case, observe that we can equip Θ with the alternative inner product $\langle \theta, \theta' \rangle_\eta = \eta^{1/p} \langle \theta, \theta' \rangle_\Theta$ which yields the norm $|\theta|_\eta = \eta^{1/p} |\theta|_\Theta$. Since the preceding argument applies to general separable Hilbert spaces, it also holds when Θ is equipped with the alternative norm, yielding

$$\mathcal{T}_c(d) \leq \mathcal{T}^{(\theta)}_{c_{\eta,\Theta}}(d) \tag{22}$$

for $c_{\Theta,\eta}(\theta, \theta') = \eta |\theta - \theta'|^p$. Further, observe that

$$\begin{aligned}
\mathcal{T}^{(\theta)}_{c_\Theta,\eta}(d) &= \int_{\mathcal{Y}} \left[\min_{\gamma \in \Pi(p(\theta), p(\theta \mid y,d))} \int_{\Theta^2} \eta |\theta - \theta'|^2 \, \mathrm{d}\gamma(\theta, \theta') \right] p(y \mid d) \, \mathrm{d}y && (23) \\
&= \eta \int_{\mathcal{Y}} \left[\min_{\gamma \in \Pi(p(\theta), p(\theta \mid y,d))} \int_{\Theta^2} |\theta - \theta'|^2 \, \mathrm{d}\gamma(\theta, \theta') \right] p(y \mid d) \, \mathrm{d}y && (24) \\
&= \eta \mathcal{T}^{(\theta)}_{c_\Theta}(d). && (25)
\end{aligned}$$

This yields the desired claim. □

A.1 Moment Bounds

In this section, we prove several upper bounds on our transport dependencies which rely on moments of the underlying distributions. In particular, this theorem shows that when there is a link between the DTD $\mathcal{T}^{(y)}_c(d)$ (and thus also the MTD by Lemma 1) and the predictive variance of $p(y \mid d)$. Intuitively, this means that we should expect that maximizing the DTD (and MTD) should select for designs for which there is a high amount of variance in the experimental outcome.

Theorem 4. *Suppose $\Theta, \mathcal{Y}$ are separable Hilbert spaces. Fix $p \in [1, \infty)$ and suppose that $p(\theta, y \mid d)$ has finite pth moment. For $c_\Theta(\theta, \theta') = |\theta - \theta|^p$, we have*

$$\mathcal{T}^{(\theta)}_{c_\Theta}(d) \leq 2^p \mathbb{E}_{p(\theta)} |\theta - \mathbb{E}[\theta]|^p. \tag{26}$$

Similarly, for $c_{\mathcal{Y}}(y, y') = |y - y'|^p$,

$$\mathcal{T}^{(y)}_{c_{\mathcal{Y}}}(d) \leq 2^p \mathbb{E}_{p(y|d)} |y - \mathbb{E}[y \mid d]|^p. \tag{27}$$

Proof. We begin with the bound on $\mathcal{T}^{(\theta)}_c(d)$. Observe that $\gamma(\theta, y, \theta', y') = p(\theta, y \mid \xi) p(\theta') \delta[y' = y]$ is a valid triangular coupling of $p(\theta, y \mid d)$ and $p(\theta') p(y' \mid d)$. By the convexity of $x \mapsto x^p$, we have

$$\begin{aligned}
\mathcal{T}^{(\theta)}_{c_\Theta}(d) &\leq \mathbb{E}_\gamma |\theta - \theta'|^p = \mathbb{E}_\gamma |(\theta - \mathbb{E}_{p(\theta)}[\theta]) - (\theta' - \mathbb{E}_{p(\theta)}[\theta])|^p && (28) \\
&\leq 2^{p-1} \mathbb{E}_\gamma \left[|\theta - \mathbb{E}\theta|^p + |\theta' - \mathbb{E}\theta|^p \right] && (29) \\
&= 2^{p-1} \left(\int_{\Theta^2 \times \mathcal{Y}} |\theta - \mathbb{E}\theta|^p \, \mathrm{d}p(\theta') \, \mathrm{d}p(\theta, y \mid d) + \int_{\Theta^2 \times \mathcal{Y}} |\theta' - \mathbb{E}\theta|^p \, \mathrm{d}p(\theta') \, \mathrm{d}p(\theta, y \mid d) \right) && (30) \\
&= 2^p \mathbb{E}_{p(\theta)} |\theta - \mathbb{E}\theta|^p. && (31)
\end{aligned}$$

where the last line follows by marginalization. The proof for $\mathcal{T}^{(y)}_{c_{\mathcal{Y}}}(d)$ is analogous. □

We note that in the case $p = 2$, one may use the identity

$$\begin{aligned}
\mathbb{E}_\gamma |\theta - \theta'|^2 &= \mathbb{E}_\gamma |(\theta - \mathbb{E}\theta) - (\theta' - \mathbb{E}\theta)|^2 && (32) \\
&= 2\mathbb{E}_{p(\theta)} |\theta - \mathbb{E}\theta|^2 - 2\mathbb{E}_\gamma \langle \theta - \mathbb{E}\theta, \theta' - \mathbb{E}\theta \rangle && (33)
\end{aligned}$$

rather than convexity to obtain a sharper constant.

A.2 Limiting Cases of the MTD

In this section, we prove that the TTD and DTD can be obtained as limiting cases of the MTD under a particular choice of cost. We refer to Section 3.2 for a discussion of this result and its implications for OED. The following is a formal restatement of Theorem 1.

Theorem 5. *Suppose $\Theta, \mathcal{Y}$ are separable Hilbert spaces. Fix $p \in [1, \infty)$ and assume that $p(\theta, y \mid d)$ has finite pth moment. Consider the cost*

$$c_\eta(\theta, y, \theta', y') = \eta|\theta - \theta'|_\Theta^p + |y - y'|_\mathcal{Y}^p \tag{34}$$

for $\eta > 0$. Then, $\eta^{-1}\mathcal{T}_{c_\eta}(d) \to \mathcal{T}_{c_\Theta}^{(\theta)}(d)$ as $\eta \to 0^+$, where $c_\Theta(\theta, \theta') = |\theta - \theta'|_\Theta^p$. Similarly, for

$$c_\psi(\theta, y, \theta', y') = |\theta - \theta'|_\Theta^p + \psi|y - y'|_\mathcal{Y}^p \tag{35}$$

we have $\psi^{-1}\mathcal{T}_{c_\psi}(d) \to \mathcal{T}_{c_\mathcal{Y}}^{(y)}(d)$ as $\psi \to 0^+$ where $c_\mathcal{Y}(y, y') = |y - y'|_\mathcal{Y}^p$.

Proof. We begin with the first claim. Let $\gamma^* \in \Pi_\mathcal{Y}$ be an optimal $\mathcal{Y}$-triangular coupling for the cost $c_\Theta(\theta, \theta') = |\theta - \theta'|^p$ and let $\gamma_\eta \in \Pi$ be an optimal coupling for the cost c_η. Such optimal couplings exist and yield finite costs due to our moment assumption and Theorem 4.

Since $\Pi_\mathcal{Y} \subset \Pi$, using the $\mathcal{Y}$-triangularity of γ^* we may upper bound the MTD under the cost c_η by

$$\begin{aligned}
\mathcal{T}_{c_\eta}(d) &= \int c_\eta \, \mathrm{d}\gamma_\eta \leq \int c_\eta \, \mathrm{d}\gamma^* && (36)\\
&= \int \eta|\theta - \theta'|^p \, \mathrm{d}\gamma^* + \int |y - y'|^p \, \mathrm{d}\gamma^* && (37)\\
&= \eta \int |\theta - \theta'|^p \, \mathrm{d}\gamma^*. && (38)
\end{aligned}$$

Consequently, by expanding out the definition of $\mathcal{T}_{c_\eta}(d)$ we see that

$$0 \leq \eta^{-1} \int |y - y'|^p \, \mathrm{d}\gamma_\eta \leq \int |\theta - \theta'|^p \, \mathrm{d}(\gamma^* - \gamma_\eta). \tag{39}$$

This yields $\int |\theta - \theta'|^p \, \mathrm{d}\gamma_\eta \leq \int |\theta - \theta'|^p \, \mathrm{d}\gamma^*$, so that $\limsup_{\eta \to 0^+} \int |\theta - \theta'|^p \, \mathrm{d}\gamma_\eta \leq \int |\theta - \theta'|^p \, \mathrm{d}\gamma^*$, which is finite as we assume $p(\theta, y \mid d)$ has finite pth moment. Hosseini et al. (2025, Prop. 3.11) show that as $\eta \to 0^+$, we have $\gamma_\eta \to \gamma^*$ in the weak sense. By the Portmanteau theorem, this weak convergence implies $\liminf_{\eta \to 0^+} \int |\theta - \theta'|^p \, \mathrm{d}\gamma_\eta \geq \int |\theta - \theta'|^p \, \mathrm{d}\gamma^*$. We have thus shown

$$\lim_{\eta \to 0^+} \int |\theta - \theta'|^p \, \mathrm{d}\gamma_\eta = \int |\theta - \theta'|^p \, \mathrm{d}\gamma^*. \tag{40}$$

By Equation (39), we thus also have

$$\lim_{\eta \to 0^+} \eta^{-1} \int |y - y'|^p \, \mathrm{d}\gamma_\eta = 0. \tag{41}$$

Together, Equation (40) and Equation (41) imply that

$$\begin{aligned}
\lim_{\eta \to 0^+} \eta^{-1}\mathcal{T}_{c_\eta}(d) &= \lim_{\eta \to 0^+} \left(\int |\theta - \theta'|^p \, \mathrm{d}\gamma_\eta + \eta^{-1} \int |y - y'|^p \, \mathrm{d}\gamma \right) && (42)\\
&= \int |\theta - \theta'|^p \, \mathrm{d}\gamma^* = \mathcal{T}_{c_\Theta}^{(\theta)}(d). && (43)
\end{aligned}$$

The second claim can be shown with a directly analogous argument, interchanging the roles of θ and y. □

A.3 Estimation and Optimization

Estimation. Here we include further details regarding the estimation and optimization of $\mathcal{T}_c(d)$. We focus on the setting where θ, y are continuous. In principle, to form our plug-in estimate $\widehat{\mathcal{T}}_c(d)$ in Equation (15), we require samples

$$(\theta_j, y_j) \overset{\text{i.i.d.}}{\sim} p(\theta, y \mid d) \quad j = 1, \ldots, n \qquad (\theta'_k, y'_k) \overset{\text{i.i.d.}}{\sim} p(\theta)p(y \mid d) \quad k = 1, \ldots, n. \tag{44}$$

The product of marginals $p(\theta)p(y \mid d)$ can be sampled by drawing $\theta'_k, \theta''_k \sim p(\theta)$ followed by sampling $y'_k \sim p(y \mid \theta''_k, d)$. In principle, this requires $2n$ draws from the prior and simulations from the likelihood. However, in practice we reduce this to n draws by first obtaining $(\theta_j, y_j) \overset{\text{i.i.d.}}{\sim} p(\theta, y \mid d)$ followed by choosing a derangement σ (i.e., a permutation with no fixed points) and defining $\theta'_k = \theta_j, y'_k = y_{\sigma(j)}$, breaking the dependency. This allows for a computational speedup (particularly when simulating the likelihood is expensive) and further can serve to reduce the bias of our estimator. We will write μ_n, ν_n for these two empirical measures, i.e.,

$$\mu_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{(\theta_j, y_j)} \qquad \nu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{(\theta'_k, y'_k)}. \tag{45}$$

This yields the plug-in estimator,

$$\mathcal{T}_c(d) \approx \widehat{\mathcal{T}}_c(d) = \mathsf{OT}_c\left[\mu_n, \nu_n\right], \tag{46}$$

which can be solved using efficient linear-programming techniques (Bonneel et al., 2011; Peyré et al., 2019; Papp and Sherlock, 2025). In particular, this requires forming the cost matrix $C(d) \in \mathbb{R}^{n\times n}$ with entries $C_{j,k}(d) = c(\theta_j, y_j, \theta'_k, y'_k)$. Note that $C(d)$ is a function of d as y_j, y'_k depend on d through the likelihood. Further, the value of $\widehat{\mathcal{T}}_c(d)$ is determined entirely by this cost matrix $C(d)$. We will write $\mathcal{G}(C)$ for the minimal transport cost obtained for a given cost matrix C.

Optimization. We require an estimate of $\nabla_d \mathcal{T}_c(d)$, which we obtain by computing the gradient $\nabla_d \widehat{\mathcal{T}}_c(d)$ of our plug-in estimator. Key to computing this is the envelope theorem (Danskin, 1967; Bertsekas, 1971, 1997), which shows that we can obtain the gradient of $C \mapsto \mathcal{G}(C)$ in terms of the optimal transport plan. To be more precise, this mapping is not differentiable, but we use $\partial\mathcal{G}(C)$ to represent the superdifferential (Boyd and Vandenberghe, 2004) of $\mathcal{G}$ at C, i.e., the set of all $v \in \mathbb{R}^{n\times n}$ satisfying

$$\mathcal{G}(C') - \mathcal{G}(C) \leq \langle v, C' - C\rangle_F \qquad \forall C' \in \mathbb{R}^{n\times n} \tag{47}$$

for the Frobenius inner product $\langle\cdot,\cdot\rangle_F$. The following theorem shows that $\partial\mathcal{G}(C)$ is precisely given by the set of optimal plans (Peyré et al., 2019, Prop. 9.2), which in general may be non-unique.

Theorem 6. *For the mapping $C \mapsto \mathcal{G}(C)$, we have*

$$\partial\mathcal{G}(C) = \left\{\gamma^* \in \Pi(\mu_n, \nu_n) : \gamma^* \in \operatorname*{argmin}_{\Pi(\mu_n,\nu_n)} K_c(\gamma)\right\}. \tag{48}$$

Thus, computing a supergradient of $\partial\mathcal{G}(C)$ requires no more computation than solving the OT problem itself, as it is simply the value of the optimal plan found when solving the linear program.

The (super-)gradient $\nabla_d \widehat{\mathcal{T}}_c(d)$ then requires us to use the chain rule to compute $\nabla_d \mathcal{G}(C(d))$:

$$\nabla_d \widehat{\mathcal{T}}_c(d) = \sum_{j,k=1}^{n} \gamma^*_{jk} \nabla_d C_{jk}(d) \tag{49}$$

$$= \sum_{j,k=1}^{n} \gamma^*_{jk} \left(\nabla_d y_j(\eta,\theta,d)^T \partial_2 c(\theta_j, y_j, \theta'_k, y'_k) + \nabla_d y'_k(\eta,\theta,d)^T \partial_4 c(\theta_j, y_j, \theta'_k, y'_k)\right) \tag{50}$$

where the second equality follows from directly computing $\nabla_d C_{jk}(d)$. Here, we use $\partial_2 c$ and $\partial_4 c$ to denote the partial derivatives of c with respect to its second and fourth arguments, and $\nabla_d y(\eta,\theta,d) \in \mathbb{R}^{d_y \times d_d}$ is the Jacobian of y with respect to d. In practice, $\nabla_d C_{jk}(d)$ can be computed via automatic differentiation when we have a differentiable cost function and a differentiable sampler. In particular, as discussed in Section 3.3, using the noise outsourcing lemma (Kallenberg, 1997), for any fixed noise distribution with appropriate reference measure, $q(\eta)$, there exists (subject to weak assumptions) a function h such that $y(\eta,\theta,d) = h(\eta;\theta,d)$ for $\eta \sim q(\eta)$. If we further assume that $d \mapsto h(\eta;\theta,d)$ is differentiable for our chosen $q(\eta)$, we may use automatic differentiation to compute $\nabla_d\, y(\eta,\theta,d)$. We refer to Peyré et al. (2019, Chapter 9) for a further discussion of the differentiability of optimal transport discrepancies. While in principle $\nabla_d \widehat{\mathcal{T}}_c(d)$ is merely a supergradient, this is sufficient for the purposes of performing stochastic gradient-ascent based procedures on the MTD.

B TRANSPORT DEPENDENCIES AND MUTUAL INFORMATION

In this section, we give a proof for Theorem 2, as well as an extended discussion of this result. In addition, we show that the total variation distance between $p(\theta, y \mid d)$ and $p(\theta)p(y \mid d)$ may be obtained as a special case of the MTD, and provide an upper bound analogous to Theorem 2.

B.1 The Euclidean Case

While the MTD can be defined for general Polish spaces, we work under a Euclidean assumption throughout this section to facilitate the analysis, and also because this is a highly practically relevant scenario. We turn our attention towards bounding the transport dependencies by the mutual information. The key assumption we rely on is a strong log-concavity assumption.

Definition 5 (Strong Log-Concavity)**.** *Let $p \in C^2(\mathbb{R}^m)$ be a twice continuously differentiable probability density function. We say that p is strongly log-concave if there exists $\lambda > 0$ such that for all x with $p(x) > 0$, we have*

$$-\nabla_x^2 \log p(x) \succeq \lambda I_n. \tag{51}$$

The greatest such λ is called the parameter of log-concavity for $p(x)$.

A generalized form of Talagrand's inequality yields an upper bound on the MTD in terms of the mutual information (Villani et al., 2008, Section 9.3), (Blower, 2003, Theorem 4.1).

Theorem 7 (Talagrand's Inequality)**.** *Suppose $\mathcal{X} = \mathbb{R}^m$ is a Euclidean space and $p(x), q(x)$ are two probability densities over $\mathcal{X}$. Suppose $p(x)$ is strongly log-concave with parameter λ and $\mathsf{KL}[q(x)||p(x)] < \infty$. For the cost function $c(x, x') = |x - x'|^2$, we have*

$$\mathsf{OT}_c[p(x), q(x)] \leq 2\lambda^{-1}\mathsf{KL}\left[q(x) \mid\mid p(x)\right]. \tag{52}$$

Using Talagrand's inequality, we may relate our MTD and other transport discrepancies to the mutual information. We provide a more formal statement of Theorem 2 here, as well as a proof.

Theorem 8. *Suppose $\Theta = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$.*

1. *Assume the prior $p(\theta)$ is strongly log-concave with parameter λ_θ. For $c(\theta, \theta') = |\theta - \theta'|^2$, we have*
$$\lambda_\theta \mathcal{T}_c^{(\theta)}(d) \leq 2\mathcal{I}(d). \tag{53}$$
2. *Assume the marginal $p(y \mid d)$ is strongly log-concave with parameter $\lambda_{y|d}$. For $c(y, y') = |y - y'|^2$, we have*
$$\lambda_{y|d} \mathcal{T}_c^{(y)}(d) \leq 2\mathcal{I}(d). \tag{54}$$
3. *Let $\eta > 0$ be given. When both the prior and likelihood satisfy these assumptions, under cost $c(\theta, \theta', y, y') = \eta|\theta - \theta'|^2 + |y - y'|^2$, for $\lambda = \max\{\lambda_\theta/\eta, \lambda_{y|d}\}$ we have*
$$\lambda \mathcal{T}_c(d) \leq 2\mathcal{I}(d). \tag{55}$$

Proof. We start with the first claim. If $\mathcal{I}(d)$ is infinite, then there is nothing to show. When $\mathcal{I}(d)$ is finite and $p(\theta)$ is strongly log-concave, by Theorem 7 we have

$$\mathsf{OT}_c[p(\theta \mid y, d), p(\theta)] \leq 2\lambda_\theta^{-1}\mathsf{KL}[p(\theta \mid y, d)||p(\theta)]. \tag{56}$$

Integrating both sides of this with respect to $p(y \mid d)$ yields

$$\begin{aligned}
\mathcal{T}_c^{(\theta)}(d) &= \int_{\mathcal{Y}} \mathsf{OT}_c[p(\theta \mid y, d), p(\theta)]p(y \mid d)\,\mathrm{d}y && (57)\\
&\leq 2\lambda_\theta^{-1} \int_{\mathcal{Y}} \mathsf{KL}[p(\theta \mid y, d)||p(\theta)]p(y \mid d)\,\mathrm{d}y && (58)\\
&= 2\lambda_\theta^{-1}\mathcal{I}(d) && (59)
\end{aligned}$$

as claimed. The proof for the second claim is analogous.

For the last claim, consider $c_{\Theta,\eta}(\theta,\theta') = \eta|\theta-\theta'|^2$. As Lemma (1) applies in arbitrary separable Hilbert spaces, we obtain

$$\mathcal{T}_c(d) \le \min\{\mathcal{T}^{(\theta)}_{c_{\Theta},\eta}(d), \mathcal{T}^{(y)}_{c_{\mathcal{Y}}}(d)\} \tag{60}$$
$$= \min\{\eta\mathcal{T}^{(\theta)}_{c_{\Theta}}(d), \mathcal{T}^{(y)}_{c_{\mathcal{Y}}}(d)\} \tag{61}$$
$$\le 2\min\{\eta\lambda_{\theta}^{-1}, \lambda_{y|d}^{-1}\}\mathcal{I}(d). \tag{62}$$

where the final inequality follows from the first two claims. Rearranging this inequality yields the result. □

While the bounds in this previous section apply for Euclidean spaces, generalizations of Talagrand's inequality exist for general metric spaces (Gozlan and Léonard, 2010), although resulting in a more complex relationship between information-theoretic quantities and optimal transport divergences.

B.2 Total Variation and Hamming Distance

One particular special case that may be of interest is when the cost function is defined via the Hamming distance. In this case, we obtain the total variation distance between $p(\theta, y \mid d)$ and $p(\theta)p(y \mid d)$ as a special case of our MTD framework, and moreover, obtain an upper bound in terms of the mutual information. However, we note that as this cost function is not differentiable, the MTD under this cost function cannot be directly optimized with gradient-based methods.

Theorem 9. *Suppose $\Theta\times\mathcal{Y}$ is a metric space. Consider the Hamming distance $c_H(\theta,y,\theta',y') = \mathbf{1}[(\theta,y)\neq(\theta',y')]$. Then,*

$$\mathcal{T}_{c_H}(d) = |p(\theta,y\mid d) - p(\theta)p(y\mid d)|_{TV} = \sup_{A\in\mathcal{B}}\left\{\int_A p(\theta,y\mid d)\,\mathrm{d}\theta\;dy - \int_A p(\theta)p(y\mid d)\,\mathrm{d}\theta\,\mathrm{d}y\right\} \tag{63}$$

where $\mathcal{B}$ is the set of all Borel subsets of $\Theta\times\mathcal{Y}$ and $|\cdot|_{TV}$ is the total variation metric. Moreover,

$$\mathcal{T}_{c_H}(d) \le \sqrt{\frac{1}{2}\mathcal{I}(d)}. \tag{64}$$

Proof. It is well-known that the Hamming cost in the OT problem yields the total variation distance (Massart, 2007, Lemma 2.20). By the Csiszár-Kullback-Pinsker inequality (Pinsker, 1964; Csiszár, 1967; Kullback, 1967; Gozlan and Léonard, 2010), we immediately obtain the desired upper bound. □

C LINEAR-GAUSSIAN MODEL

Here we investigate the behavior of the gain $\mathcal{T}_c(d)$ in the linear-Gaussian setting under quadratic costs. In particular, we may obtain a closed-form solution for the MTD in this case, allowing for a direct comparison against the MI. See Figure 6 for an illustration.

Theorem 10. *Suppose $\theta\in\Theta=\mathbb{R}^n$ has a standard normal prior $p(\theta)=\mathcal{N}(0,I_n)$, designs are vectors $d\in\mathcal{D}=\mathbb{R}^n$, and $y\in\mathcal{Y}=\mathbb{R}$ has likelihood $p(y\mid\theta,d)=\mathcal{N}\left(\langle d,\theta\rangle,\sigma^2_{d,\theta}\right)$. Under the quadratic cost, we have*

$$\mathcal{T}_c(d) = 2\left(1+\sigma^2_{d,\theta}+|d|^2 - \sqrt{1+(|d|^2+\sigma^2_{d,\theta})^2+2\sqrt{|d|^2+\sigma^2_{d,\theta}}}\right).$$

Moreover, $\mathcal{T}_c(d)\le 2$. On the other hand, the MI is

$$\mathcal{I}(d) = \frac{1}{2}\log\left(1+|d|^2/\sigma^2_{d,\theta}\right) \tag{65}$$

which is unbounded as $\sigma^2_{d,\theta}\to 0$.

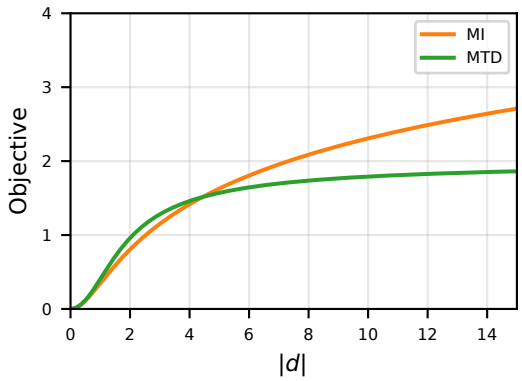

Figure 6: Values of the MI and MTD for the linear-Gaussian model for $\sigma^2_{d,\theta} = 1$.

Proof. Define $s = |d|^2 + \sigma^2_{d,\theta}$. Under this setting we may explicitly calculate the joint and product of marginals as

$$p(\theta, y \mid d) = \mathcal{N}(0, \Sigma_J) \qquad \Sigma_J = \begin{bmatrix} I & d \\ d^T & s \end{bmatrix} \tag{66}$$

$$p(\theta)p(y \mid d) = \mathcal{N}(0, \Sigma_P) \qquad \Sigma_P = \begin{bmatrix} I & 0 \\ 0 & s \end{bmatrix}. \tag{67}$$

Under quadratic costs, $\mathcal{T}_c(d)$ is the squared 2-Wasserstein between these distributions, which admits a closed-form via the (squared) Bures-Wasserstein metric

$$\mathcal{T}_c(d) = \operatorname{tr}\left(\Sigma_J + \Sigma_P - 2\sqrt{\Sigma_P^{1/2}\Sigma_J\Sigma_P^{1/2}}\right). \tag{68}$$

Observe that $\operatorname{tr}(\Sigma_J) = \operatorname{tr}(\Sigma_P) = n + s$ and thus we turn our attention to the term $B = \sqrt{\Sigma_P^{1/2}\Sigma_J\Sigma_P^{1/2}}$. Since Σ_P is diagonal its square-root is simply

$$\Sigma_P^{1/2} = \begin{bmatrix} I & 0 \\ 0 & \sqrt{s} \end{bmatrix} \tag{69}$$

and an explicit calculation of B^2 yields

$$B^2 = \Sigma_P^{1/2}\Sigma_J\Sigma_P^{1/2} = \begin{bmatrix} I & \sqrt{s}\xi \\ \sqrt{s}\xi^T & s^2 \end{bmatrix}. \tag{70}$$

We require $\operatorname{tr}(B)$. Let (λ_i) be the eigenvalues of B^2, so that $\operatorname{tr}(B) = \sum_i \sqrt{\lambda_i}$. Using the expression for the determinant of a block matrix, we see that the characteristic polynomial of B^2 is

$$p(\lambda) = (1-\lambda)^n\left(s^2 - \lambda - \frac{sr^2}{1-\lambda}\right) = (1-\lambda)^{n-1}\left(\lambda^2 - (1+s^2)\lambda + s\right). \tag{71}$$

Thus, B^2 has eigenvalues $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-2} = 1$ and eigenvalues λ_{n-1}, λ_n which are the roots of the quadratic part. In particular, $\lambda_{n-1} + \lambda_n = 1 + s^2$ and $\lambda_{n-1}\lambda_n = s$. This yields

$$\sqrt{\lambda_{n-1}} + \sqrt{\lambda_n} = \sqrt{1 + s^2 + 2\sqrt{s}} \tag{72}$$

$$\operatorname{tr}(B) = (n-1) + \sqrt{1 + s^2 + 2\sqrt{s}}. \tag{73}$$

Putting everything together via the additivity of the trace, we have

$$\mathcal{T}_c(d) = 2\left(1 + \sigma^2_{d,\theta} + |d|^2 - \sqrt{1 + (|d|^2 + \sigma^2_{d,\theta})^2 + 2\sqrt{|d|^2 + \sigma^2_{d,\theta}}}\right). \tag{74}$$

Using a computer algebra system one can verify that $\mathcal{T}_c(d) \leq 2$ and that $\lim_{|d|\to\infty} \mathcal{T}_c(d) = 2$.

On the other hand, the mutual information between two Gaussians admits a closed form. This immediately gives

$$\mathcal{I}(d) = \frac{1}{2} \log \left(1 + \frac{|d|^2}{\sigma_{d,\theta}^2} \right). \tag{75}$$

which is unbounded as $\sigma_{d,\theta}^2 \to 0$. □

D EXPERIMENT DETAILS

This section provides additional details for our experiments. Unless specified otherwise, our experiments in Section 6 were performed with the following settings. Experiments were performed on an Apple M4 Pro chip with 24 GB of unified memory and a 14-core CPU and primarily implemented in `pytorch` (Paszke et al., 2019).

Designs are optimized for 250 gradient steps with a learning rate of 2e-2 using the Adam optimizer (Kingma, 2014). For MTD, we draw 1 000 samples from $p(\theta, y \mid d, h_t) = p(\theta \mid h_t)p(y \mid \theta, d)$ per gradient step, shuffled following Appendix A to yield samples from $p(\theta)p(y \mid d)$. For PCE, we draw $N = 1\,000$ samples $(\theta_n, y_n) \overset{\text{i.i.d.}}{\sim} p(\theta, y \mid d, h_t)$ to form the approximation (Foster et al., 2020)

$$\mathcal{I}^{(t)}(d) \approx \frac{1}{N} \sum_{n=1}^{N} \left[\log p(y_n \mid \theta_n, d) - \log \left(\frac{1}{L+1} \left(p(y_n \mid \theta_n, d) + \sum_{\ell=1}^{L} p(y_n \mid \theta_{\ell,n}, d) \right) \right) \right]. \tag{76}$$

where $L = 1\,000$ and $\theta_{\ell,n} \overset{\text{i.i.d.}}{\sim} p(\theta \mid h_t)$.

Figure 1 and Figure 3 are plotted with additional Gaussian smoothing for visualization purposes.

D.1 Location Finding

In the location finding task, our goal is to estimate the spatial location of $K \geq 1$ sources $\theta_k \in \mathbb{R}^{d_\theta}$. The number of sources K is assumed to be known, so that $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$. Each source θ_k emits a signal which decays according to an inverse square law. In each step, a sensor is placed at a location $d \in \mathbb{R}^{d_\theta}$ which records a noisy measurement $y \in \mathbb{R}$ of the total signal intensity at the sensor location (Sheng and Hu, 2005). We note that this task has become a standard benchmark for BED (Foster et al., 2021; Ivanova et al., 2021; Blau et al., 2022; Iollo et al., 2024a).

The total (noiseless) intensity at a location d is given by

$$\mu(\theta, d) = b + \sum_{k=1}^{K} \frac{\alpha_k}{m + |\theta_k - d|^2}. \tag{77}$$

Here, $b, \alpha_k, m \geq 0$ are known constants. The variable b represents a background signal level and m controls the (inverse) maximum signal strength. We assume an independent standard Gaussian prior over each θ_k,

$$p(\theta_k) = \mathcal{N}(\theta_k \mid 0, I_{d_\theta}) \qquad k = 1, \ldots, K. \tag{78}$$

We further assume that we observe the logarithm of the total signal intensity with Gaussian noise, i.e.,

$$y \mid \theta, d \sim \mathcal{N}\left(y \mid \log \mu(\theta, d), \sigma^2\right) \tag{79}$$

where $\sigma^2 > 0$ is assumed to be known. In all experiments we follow prior work (Foster et al., 2021) and set $b = 0.1, \alpha_k = 1, m = 10^{-4}, \sigma^2 = 0.25$. Further, we consider $K = 2$ sources in $d_\theta = 2$ dimensions, so that θ is four dimensional.

Inference. For inference, we use the NUTS MCMC sampler implemented in `pymc3` (Salvatier et al., 2016) with 1e4 warm-up steps and four independent chains to draw a total of 1e5 posterior samples at each experiment iteration. In LF, posteriors are complex and multi-modal, necessitating accurate (but expensive) inference. See Figure 7 for an illustration.

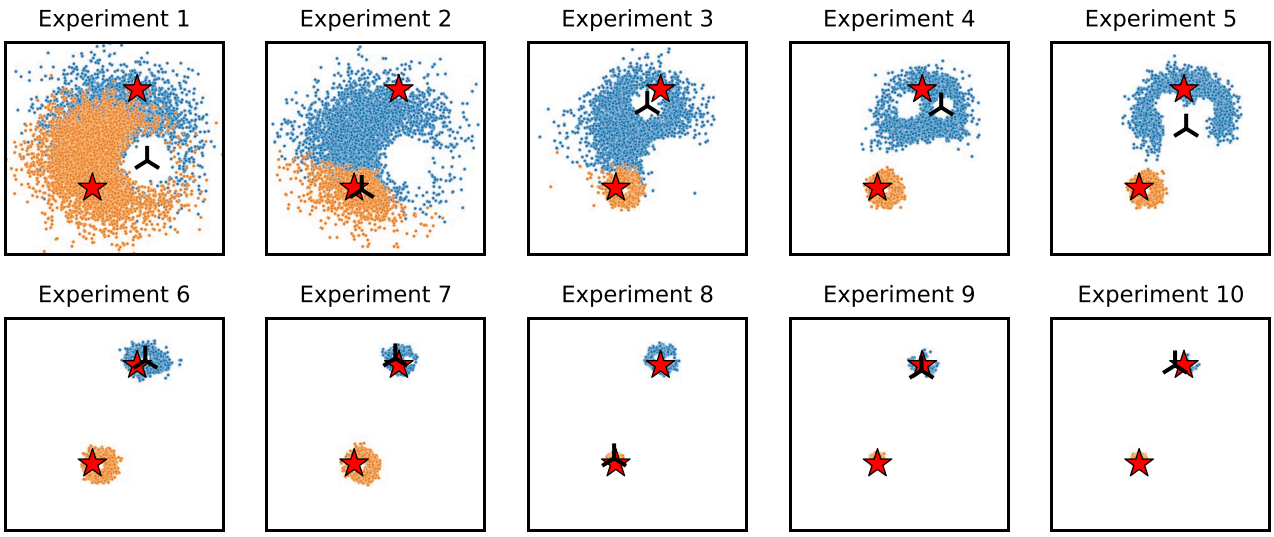

Figure 7: An illustration of the LF problem. Red stars indicate the true, unknown source locations θ_1, θ_2. At each experimental iteration, a measurement location (black triangle) is selected by maximizing the MTD. Posterior samples $\theta \sim p(\theta \mid h_t)$ (orange and blue points) depict the updated beliefs about the source positions after each measurement. The MTD adaptively selects measurement locations that rapidly determine both sources.

RMSE. During posterior sampling, the model exhibits non-identifiability with respect to the ordering of the two sources, θ_1 and θ_2. As a result, the correspondence between estimated and true sources may be swapped across samples. To address this when computing the RMSE, we evaluate both possible orderings of each posterior sample and use the ordering that yields the lower RMSE.

Weighting Function. In Section 6.3 we modify the cost function in MTD using a weighting function. In particular, this takes the form

$$w(\theta) = b + g(\theta_1 - \mu) + g(\theta_2 - \mu) \tag{80}$$

where $b = 1$ is a bias and $\mu = (1.5, -1.5)$ controls the location of the bump, and

$$g(x) = k\left(1 - \text{sigmoid}\left(s\left(\frac{|x|^2 - \alpha^2}{\beta^2 - \alpha^2}\right)\right)\right) \tag{81}$$

is a bump-like function. Here, $\beta = 1$ approximately controls the radius of its support, $\alpha = 0.5$ controls an inner radius where g is approximately maximized, $s = 0.3$ controls the slope of the bump, and $k = 1\text{e}4$ controls its amplitude. See Figure 8 for a visualization. The weighting $w(\theta)$ is selected to depend only on θ which is drawn from the joint.

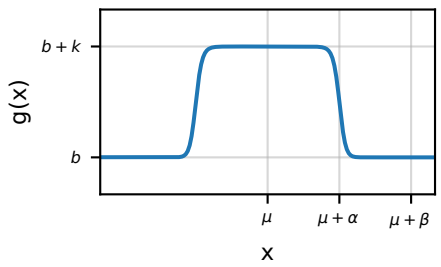

Figure 8: A visualization of the bump function $g(x)$.

D.2 Constant Elasticity of Substitution (CES)

The Constant Elasticity of Substitution (CES) model, arising from behavioral economics, asks a participant to compare two baskets $d_1, d_2 \in [0, 100]^3$ consisting of various amounts of three different goods. Given two baskets, the participant provides a scalar response $y \in [0, 1]$ indicating subjective preference between the baskets. This model has previously served as a benchmark for several recent BED works (Foster et al., 2019, 2020; Blau et al., 2022; Iollo et al., 2024b).

The experimental goal is to choose a design $d = (d_1, d_2) \in [0, 100]^6$ consisting of two baskets in order to infer the participant's latent utility function. This utility function is assumed to be parametrized by $\theta = (\rho, \alpha, u)$ where $\rho \in [0, 1], \alpha \in \Delta_3, u \in \mathbb{R}_{\geq 0}$ and Δ_3 is the 3-simplex. Thus, $\theta \in \mathbb{R}^5$ is a five-dimensional unknown parameter of interest.

Following previous work, we assume the following priors:

$$\rho \sim \beta(1, 1) \tag{82}$$

$$\alpha \sim \mathrm{Dir}(1, 1, 1) \tag{83}$$

$$\log u \sim \mathcal{N}(1, 3). \tag{84}$$

The likelihood for the participant's response is modeled as

$$U(d) = \left(\sum_{i=1}^{3} d_i^{\rho} \alpha_i \right)^{1/\rho} \tag{85}$$

$$\mu = (U(d_1) - U(d_2))u \tag{86}$$

$$\sigma = (1 + |d_1 - d_2|)\tau u \tag{87}$$

$$\eta \sim \mathcal{N}(\mu, \sigma^2) \tag{88}$$

$$y = \begin{cases} \mathrm{sigmoid}(\eta) & \epsilon < \mathrm{sigmoid}(\eta) < 1 - \epsilon \\ \epsilon & \mathrm{sigmoid}(\eta) \leq \epsilon \\ 1 - \epsilon & \mathrm{sigmoid}(\eta) \geq 1 - \epsilon \end{cases} \tag{89}$$

where $\epsilon = 2^{-22}, \tau = 5\mathrm{e}{-3}$ are fixed constants and

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{90}$$

is the usual sigmoid function. Thus, the participant's response depends on the difference in utilities $U(d_1) - U(d_2)$ between the two baskets.

Log-Likelihood. The CES model can present numerical challenges for methods that require evaluating the likelihood of observations (e.g., PCE). We follow the recommendations in (Foster et al., 2020; Iollo et al., 2024b) to evaluate this quantity. In particular, since y is censored, we have that

$$p(y \mid \theta, d) = p_0 \delta[y = \epsilon] + p_1 \delta[y = 1 - \epsilon] + (1 - p_0 - p_1) q(y \mid \theta, d) \tag{91}$$

is a mixture of Dirac deltas at the boundaries and a density on the interior, where the density

$$q(y \mid \theta, d) = \frac{1}{\sigma y (1 - y) \sqrt{2\pi}} \exp\left(-\frac{(\mathrm{logit}(y) - \mu)^2}{2\sigma^2} \right) \tag{92}$$

represents a logit-normal distribution away from the boundary with $\mathrm{logit}(y) = \mathrm{sigmoid}^{-1}(y) = \log(y/(1 - y))$. The quantities p_0, p_1 are defined via

$$p_0 = \Phi\left(\frac{\mathrm{logit}(y) - \mu}{\sigma} \right) \tag{93}$$

$$p_1 = 1 - \Phi\left(\frac{\mathrm{logit}(1 - \epsilon) - \mu}{\sigma} \right) \tag{94}$$

where Φ is the standard normal CDF. When $p_0, p_1 \ll 1$, computing their logarithms becomes challenging, in which case we approximate Φ by a first-order Taylor expansion, i.e.,

$$\Phi(x) \approx \frac{1}{-x\sqrt{2\pi}} \exp(-x^2/2) \qquad x \ll -1 \tag{95}$$

$$1 - \Phi(x) \approx \frac{1}{x\sqrt{2\pi}} \exp(-x^2/2) \qquad x \gg 1. \tag{96}$$

We perform additional clipping as necessary before taking logarithms in our implementation to further improve stability.

For inference, we perform importance resampling (Doucet et al., 2001) where 1e7 proposal samples are drawn from the prior, weighted according to the likelihood, and 1e5 proposal samples are re-drawn with replacement with probability proportional to their weight.

D.3 Active Learning

In our active learning task, our goal is to learn a GP regression model of an underlying, unknown function $f(d)$ which is assumed to be non-negative. We take the true $f(d)$ to be a mixture of two Gaussian bumps of the form

$$f(d) = \sum_{k=1}^{2} \alpha_k \exp\left(-\frac{|d - \mu_k|^2}{2\sigma_k^2}\right) \tag{97}$$

where $\alpha_1 = 5, \mu_1 = 3$, $\alpha_2 = 2, \mu_2 = -3$ and $\sigma_1 = \sigma_2 = 0.6$. We assume that observations are of the form $y = \log f(d) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$ is i.i.d. observation noise with $\sigma_y = 0.25$. In Figure 9, we plot samples $\exp(y)$ from the true underlying model, as well as $\log f(d)$ and $f(d)$. We take the test-time distribution $p_*(d_*)$ to be a uniform distribution on $[-7, 7]$ throughout. To simulate an active learning setup, we draw a pool of 500 samples from $p_*(d_*)$ and acquire labels for 3 of these pool points to initialise each method. To acquire new labels, we select the point from the pool having the highest criteria value under the respectve method (MI, EPIG, MTD).

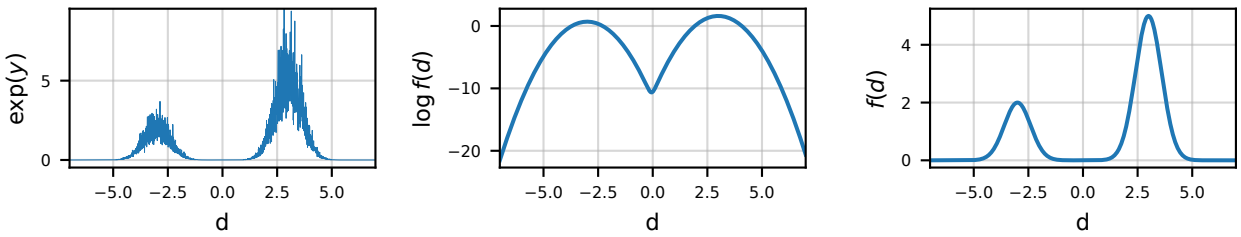

Figure 9: An illustration of the noisy observations and true underlying $f(d)$ in our active learning task.

Gaussian Process Model Towards modelling $f(d)$, we work on a logarithmic scale and posit a mean-zero GP prior for $g = \log f$ with kernel k, i.e., $g \sim \mathcal{GP}(0, k)$ with observation likelihood $p(y \mid d, g) = \mathcal{N}(g(d), \sigma_y^2)$. Here, $\theta = g$ plays the role of our unknown quantity of interest. We assume σ_y^2 is known and use an RBF kernel

$$k_{\text{RBF}}(d, d') = a \cdot \exp\left(-\frac{1}{2\ell^2}|d - d'|^2\right) \tag{98}$$

with amplitude $a = 3.5$ and lengthscale $\ell = 0.5$.

Suppose we have observed some data $h_t = \{(d_k, y_k)\}_{k=1}^t$. We write $D_t = [d_1, \ldots, d_t]^T$ and $Y_t = [y_1, \ldots, y_t]^T$. Then, the posterior $p(g \mid h_t)$ is again a Gaussian process with mean function μ_{post} and covariance operator C_{post} which are readily computed in closed-form (Williams and Rasmussen, 2006). In particular for any designs $d_\circ, d_*$, we have that the posterior mean and covariance (in the latent g) is given by

$$\mu_{\text{post}}(d_*) = K_{*D_t}^{\mathsf{T}} \left(K_{D_t D_t} + \sigma_y^2 I\right)^{-1} Y_t \tag{99}$$

$$C_{\text{post}}(d_\circ, d_*) = K_{\circ *} - K_{\circ D_t}^T (K_{D_t D_t} + \sigma_y^2 I)^{-1} K_{* D_t} \tag{100}$$

where $K_{* D_t} = [k(d_*, d_1), \ldots, k(d_*, d_t)]^T \in \mathbb{R}^t, K_{D_t D_t} = (k(d_i, d_j))_{i,j=1}^t \in \mathbb{R}^{t \times t}$ and $K_{\circ *} = k(d_\circ, d_*) \in \mathbb{R}$ are the associated kernel matrices. This yields the posterior predictive distributions

$$p(y_* \mid d_*, h_t) = \mathcal{N}(\mu_{\text{post}}(d_*), C_{\text{post}}(d_*, d_*) + \sigma_y^2). \tag{101}$$

$$p(y_*, y_\circ \mid d_*, d_\circ, h_t) = \mathcal{N}\left(\begin{bmatrix} \mu_{\text{post}}(d_*) \\ \mu_{\text{post}}(d_\circ) \end{bmatrix}, \begin{bmatrix} C_{\text{post}}(d_*, d_*) + \sigma_y^2 & C_{\text{post}}(d_*, d_\circ) \\ C_{\text{post}}(d_\circ, d_*) & C_{\text{post}}(d_\circ, d_\circ) + \sigma_y^2 \end{bmatrix} \right). \tag{102}$$

On the original scale, the likelihood and posterior predictive are then both log-normal distributed.

Mutual Information We can, in closed-form, compute the MI of this model after observing data h_t. Recall that the MI (Equation (4)) is the reduction in entropy:

$$\mathcal{I}^{(t+1)}(d) = H_{p(y|d,h_t)}[y] - \mathbb{E}_{p(g|h_t)} H_{p(y|g,d,h_t)}[y]. \tag{103}$$

Since everything is Gaussian, these entropies are available in closed form, i.e., $H[\mathcal{N}(m, \sigma^2)] = \frac{1}{2}(1 + \log(2\pi\sigma^2))$. This yields

$$\mathcal{I}^{(t+1)}(d) = \frac{1}{2}\left(1 + \log(2\pi(C_{\text{post}}(d, d) + \sigma_y^2))\right) - \mathbb{E}_{p(g|h_t)}\left[\frac{1}{2}\left(1 + \log(2\pi\sigma_y^2)\right)\right] \tag{104}$$

$$= \frac{1}{2}\log\left(1 + \frac{C_{\text{post}}(d, d)}{\sigma_y^2}\right). \tag{105}$$

Thus, we see that that maximising the MI is equivalent to maximizing $C_{\text{post}}(d, d)$, the posterior variance in the latent g at design d.

EPIG The EPIG objective (Smith et al., 2023), i.e., the average mutual information between y_*, y given d_*, d, is given by

$$\text{EPIG}^{(t+1)}(d) = \mathbb{E}_{p_*(x_*)}\left[\mathsf{KL}[p(y, y_* \mid d, d_*, h_t) \,||\, p(y \mid d, h_t) p(y_* \mid d_*, h_t)]\right] \tag{106}$$

$$= \mathsf{KL}[p(y, y_*, d_* \mid d, h_t), p(y \mid d, h_t) p(y_* \mid d_*, h_t) p_*(d_*)] \tag{107}$$

As shown in (Smith et al., 2023, Appendix E.2), for a GP model with a Gaussian likelihood, this is

$$\text{EPIG}^{(t+1)}(d) = \frac{1}{2}\mathbb{E}_{p_*(d_*)}\left[\log \frac{\widetilde{C}_{\text{post}}(d, d)\widetilde{C}_{\text{post}}(d_*, d_*)}{\widetilde{C}_{\text{post}}(d, d)\widetilde{C}_{\text{post}}(d_*, d_*) - C_{\text{post}}(d, d_*)^2}\right] \tag{108}$$

where $\widetilde{C}_{\text{post}}(d, d) = C_{\text{post}}(d, d) + \sigma_y^2$ is the posterior predictive variance at d. The outer expectation over $p_*(d_*)$ can be approximated using a Monte Carlo estimate.

Design Objective: Predictive Transport Dependence As discussed in Section 6.4, we use a variant of our MTD targetting $\theta = y_*$, the label associated with a downstream input $d_* \sim p_*(d_*)$ from a test-time input distribution $p_*(d_*)$. Our design objective takes the form

$$\mathcal{T}_c^{\text{pred}}(d) = \mathsf{OT}_c[p(y, y_*, d_* \mid d), p(y \mid d) p(y_* \mid d_*) p_*(d_*)] \tag{109}$$

which is a prediction-space generalisation of the MTD defined in Definition 1 to account for the additional input distribution $p_*(x_*)$. From Equation (107) we see that $\mathcal{T}_c^{\text{pred}}$ is an analogue of EPIG, where we use a transport-based divergence in place of the KL. Although T_c^{pred} is not equivalent to $\mathbb{E}_{p_*(d_*)}[\mathsf{OT}_c[p(y, y_* \mid d, d_*), p(y \mid d) p(y_* \mid d_*)]]$ in general, this alternative objective can be obtained by an appropriate weighting of c (as in the equivalences between the MTD and the TTD/DTD in Theorem 1).

In Section 6.4, we use the cost function

$$c(y, y_*, d_*, y', y'_*, d'_*) = |d_* - d'_*|^2 + |\exp(y) - \exp(y')|^2 + |\exp(y_*) - \exp(y'_*)|^2 \tag{110}$$

i.e. where distances in $\mathcal{Y}$-space are measured on the original scale, prior to log-transforming the function. Our ablation MTD (log) uses the cost function

$$\tilde{c}(y, y_*, d_*, y', y'_*, d'_*) = |d_* - d'_*|^2 + |y - y'|^2 + |y_* - y'_*|^2 \tag{111}$$

which is simply the Euclidean distance where y is on a logarithmic scale.

E ADDITIONAL EXPERIMENTS

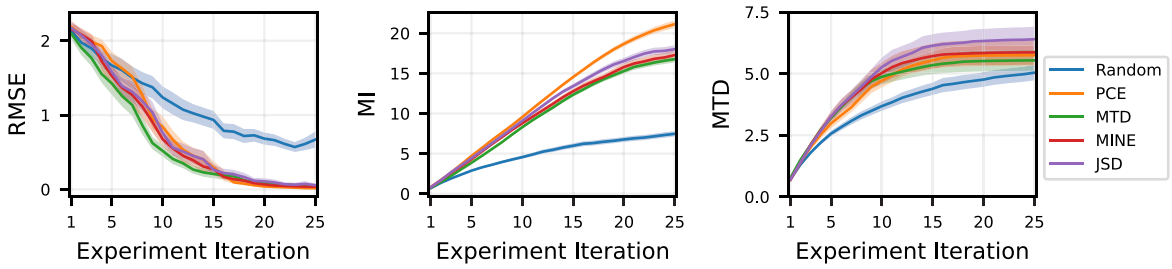

Figure 10: Quantitative results for the location finding problem, averaged across 25 seeds (± one standard error). MTD achieves the lowest RMSE across most of the rollout, even though its corresponding MI values are slightly lower. All methods obtain comparable MTD scores within one standard error, reflecting the high variance of this quantity.

E.1 Location Finding

One appeal of the MTD is that it is naturally an implicit method, as it only relies on our ability to draw samples. To highlight this, we compare against two methods for MI-based experimental design in implicit settings: MINEBED (Kleinegesse and Gutmann, 2020) and JSD (Kleinegesse and Gutmann, 2021).

MINEBED is based on the NWJ (Nguyen et al., 2010) lower bound for the mutual information:

$$\mathcal{I}(d) \geq \mathbb{E}_{p(\theta,y|d)}\left[T(\theta,y)\right] - e^{-1}\mathbb{E}_{p(\theta)p(y|d)}\left[\exp(T(\theta,y))\right] \tag{112}$$

where $T : \Theta \times \mathcal{Y} \to \mathbb{R}$ is an arbitrary measurable function. This lower bound is, in fact, tight. In practice, we take $T(\theta, y) = T_\psi(\theta, y)$ to be a neural network parametrized by ψ, yielding a strict lower bound to $\mathcal{I}(d)$ when the considered class of neural networks does not contain the true optimum. In the continuous setting, the network parameters ψ and design d may be optimized simultaneously via gradient-based procedures.

On the other hand, JSD is based on a lower bound of the Jensen-Shannon divergence (Hjelm et al., 2018)

$$\begin{aligned}
\mathcal{J}(d) &= \frac{1}{2}\mathsf{KL}\left[p(\theta,y \mid d)||m(\theta,y \mid d)\right] + \frac{1}{2}\mathsf{KL}\left[p(\theta)p(y \mid d)||p(\theta)p(y \mid d)\right] && (113)\\
&= \log 2 - \frac{1}{2}\mathbb{E}_{p(\theta,y|d)}\left[\log\left(1 + \frac{p(\theta)p(y \mid d)}{p(\theta,y \mid d)}\right)\right] - \frac{1}{2}\mathbb{E}_{p(\theta)p(y|d)}\left[\log\left(1 + \frac{p(\theta,y \mid d)}{p(\theta)p(y \mid d)}\right)\right] && (114)\\
&\geq \log 2 + \frac{1}{2}\left(\mathbb{E}_{p(\theta,y|d)}\left[-\log(1+\exp(-T(\theta,y))\right] - \mathbb{E}_{p(\theta)p(y|d)}\left[\log(1+\exp T(\theta,y))\right]\right) && (115)
\end{aligned}$$

where $m(\theta, y \mid d) = \frac{1}{2}\left(p(\theta, y \mid d) + p(\theta)p(y \mid d)\right)$ is a mixture distribution. Again we take $T(\theta, y) = T_\psi(\theta, y)$ to be a neural network, yielding a potentially strict lower bound. Kleinegesse and Gutmann (2021) motivate the JSD as an objective for OED by noting that it behaves similarly to the mutual information while potentially offering more stable training, as there is no exponential of the network unlike in MINEBED.

For both MINEBED and JSD, we parametrize T_ψ by an MLP with two hidden layers of size 200, trained at a learning rate of 3e−3 using Adam (Kingma, 2014). Designs are simultaneously optimized via Adam, but at a higher learning rate of 2e−2.

Results. In Figure 10, we report the RMSE, MI, and MTD values for our method, PCE, MINE, and JSD, with all designs optimized using five random restarts. The MTD-based approach consistently achieves the lowest RMSE across most of the experimental rollout, even though it attains slightly lower MI values, which is expected, since it does not explicitly optimize this objective. All methods obtain comparable MTD scores; this is largely attributable to the high variance of the MTD estimate and to the nature of the location-finding task, where, once the posterior becomes highly concentrated, the set of effective designs narrows substantially, so that any design within this region yields similarly strong MTD values.

E.2 CES

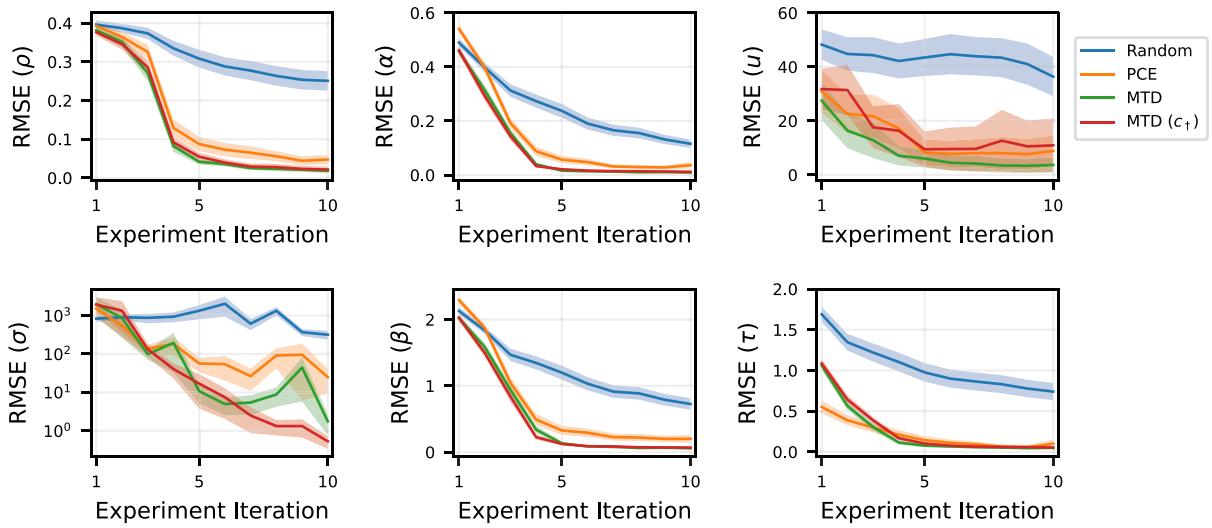

Figure 11: Quantitative results for the CES problem, averaged over 50 seeds (± one standard error). On the original scale (top row), MTD achieves the lowest RMSE, while the transformed variant $\mathcal{T}_{c_\dagger}$ performs slightly worse on u. On the transformed scale (bottom row), $\mathcal{T}_{c_\dagger}$ attains the lowest RMSEs, reflecting its alignment with the evaluation variables.

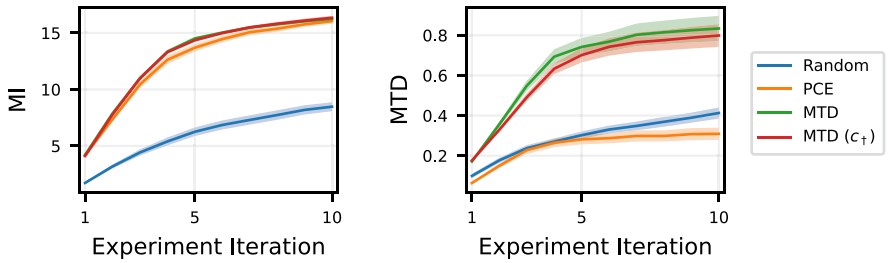

Figure 12: MI and MTD values for the CES problem, averaged over 50 seeds (± one standard error). MTD achieves similar MI values to PCE, while PCE attains worse MTD values than random designs.

To supplement the results in Section 6 for the CES problem, we provide additional figures visualizing our results.

Figure 11 plots the RMSE values for the CES problem across the experimental iterations. When evaluating performance on the original parameter scale (top row of Table 1), MTD achieves the lowest RMSE, while the transformed variant $\mathcal{T}_{c_\dagger}$ performs slightly worse on u. This is expected, as the transformed cost emphasizes errors in a different coordinate system. Conversely, when performance is assessed in the transformed space (bottom row), $\mathcal{T}_{c_\dagger}$ attains the lowest RMSEs, illustrating that defining the cost in terms of the relevant variables can effectively guide design selection toward the aspects of the parameters that are most important for downstream evaluation. These results highlight the flexibility of MTD: by choosing an appropriate cost, either directly or via transformations, experimenters can align the design criterion with the metric that truly matters for their task.

Figure 12 shows the MI and MTD values for CES. We observe that designs optimized using the MTD achieve comparable MI values to PCE. On the other hand, the designs produced by PCE yield low MTD values, performing worse than random. Notably, this serves as a verification of our Theorem 2, which can be interpreted as showing that designs which have high MTD tend to have large MI as well, but not the converse.