Learning Optimal Interventions

Jonas Mueller  David Reshef  George Du  Tommi Jaakkola
jonasmueller@csail.mit.edu, dnreshef@mit.edu, gdu@mit.edu, tommi@csail.mit.edu
MIT CSAIL

Abstract

Our goal is to identify beneficial interventions from observational data. We consider interventions that are narrowly focused (impacting few features) and may be tailored to each individual or globally enacted over a population. If the underlying relationship obeys an invariance condition, our approach can identify beneficial interventions directly from observational data, side-stepping causal inference. We provide theoretical guarantees for predicted gains when the relationship is governed by a Gaussian Process, even in settings involving unintentional downstream effects. Empirically, our approach outperforms causal inference techniques (even when our model is misspecified) and is able to discover good interventions in two practical applications: gene perturbation and writing improvement.

1 Introduction

In many applications, including medicine, the primary interest in causality has to do with identifying interventions that are likely to produce a desired change in some outcome of interest. Interventions may be tailored to each individual to improve their outcome with high probability or prescribed as a global policy applied to all individuals in the population. Optimal intervention in general requires understanding both the statistical relationship between features $X \in \mathbb{R}^d$ and the outcome $Y \in \mathbb{R}$ as well as the underlying causal structure. While existing methods for causal inference aim to learn both of these, they remain limited to large sample sizes and few dimensions. We explore an alternative paradigm to improve outcomes that dispenses with causal modeling, instead treating the underlying mechanisms as a black-box function to be optimized.

Strictly speaking, the validity of statistical modeling alone to identify beneficial interventions requires the following three assumptions:

(A1) $Y = f(X) + \varepsilon$ with $E[\varepsilon] = 0$, $\varepsilon \perp X$ for some underlying function $f$ that encodes the effects of causal mechanisms (ie. some features in $X$ causally affect $Y$, and not vice-versa).

(A2) It is clear how to precisely intervene upon certain features of an individual, and the set of feasible interventions is known (we do not consider “fat hand” settings where an intervention on one variable has unintended side-effects on other features [1]).

(A3) Invariance assumption [2]: The underlying relationship between $X$ and $Y$ is not altered by any of the interventions under consideration (ie. $P_{Y|X=x} = P_{Y'|X'=x}$ for any $X'$ produced by an intervention).

While many real-world tasks do not meet these criteria, there exist important applications in which these conditions can be assumed to approximately hold. Similar reasoning has been explored in [3,4], although these works do not address the problem of identifying the most beneficial intervention.

We use two applications to illustrate our methods. One is a writing improvement task where the data consists of documents labeled with associated outcomes (eg. grades, impact, popularity) and the goal
is to suggest beneficial changes to the author. Our second example is a gene perturbation task where the expression of certain regulatory genes can be up/down-regulated in a population (e.g. cells or bacteria) with the goal of inducing a particular phenotype or downstream gene expression pattern. These examples are settings where features cause outcomes (not vice-versa) and conditions (A1)-(A3) may hold approximately, depending on the types of external intervention used to alter the features.

2 Related Work

There exists a rich literature on causal discovery from observational data (cf. [5]) but the work is primarily theoretical and suffers under limited sample sizes / increased dimensionality [6, 7]. Recent work emphasizes a greater role for predictive rather than causal modeling in various decision-making settings [1, 3, 8], including identification of the underlying causal structure by leveraging the invariance assumption [2]. Broderson et al. use Bayesian predictive models under assumptions similar to ours in order to quantify the impact of an external intervention, after first collecting additional data under the intervention [9]. Bayesian deep learning has been used to predict the effects of hypothetical interventions [8], but the task of identifying optimal interventions is not addressed.

While our goals appear similar to Bayesian optimization [10] and we also adopt Gaussian process models, additional data is not acquired in our setting. Acquisition functions tailored for exploration of uncertain areas are therefore not appropriate. Instead, we seek interventions that lead to the greatest improvement in each individual’s outcome rather than finding a single answer (the maximum across the population). For example, in the writing improvement task, we wish to inform a particular author in each individual’s outcome rather than finding a single answer (the maximum across the population). For example, in the writing improvement task, we wish to inform a particular author in each individual’s outcome rather than finding a single answer (the maximum across the population). For example, in the writing improvement task, we wish to inform a particular author in each individual’s outcome rather than finding a single answer (the maximum across the population).

3 Methods

We assume the data $D_n := \{ (x^{(i)}, y^{(i)}) \}_{i=1}^n$ consists of i.i.d. samples from joint distribution $P_{XY}$ over features $X \in \mathbb{R}^d$ and outcomes $Y \in \mathbb{R}$. Our strategy is to first fit a Bayesian model for $Y \mid X$ whose posterior given the data encodes our beliefs about the underlying function $f$. Subsequently, this posterior for $f \mid D_n$ is used to identify a transformation of the features $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ which is likely to improve expected post-intervention outcomes ($f(T(x)) > f(x)$) according to our current beliefs. The posterior distribution $f \mid D_n$ may be summarized at any points $x, x' \in \mathbb{R}^d$ by its mean function $E[f(x)]$ and covariance function $\text{Cov}([f(x) \mid f(x')]$).

3.1 Intervening at the Individual Level

For $x \in \mathbb{R}^d$ representing the feature-measurements from an individual, we are given a set $C \subset \mathbb{R}^d$ that denotes constraints of possible interventions on $x$. Let $T(x) \in C$ denote the new feature-measurements of this individual after a particular intervention on $x$ which alters features according to transformation $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$. We first consider personalized interventions in which $T$ may be tailored to a particular $x$. Under the Bayesian perspective, $f \mid D_n$ is randomly distributed according to our posterior beliefs, and we define the individual expected gain function:

$$G_x(T) := f(T(x)) - f(x) \mid D_n$$

(1)

Since $f(x) = E_x[Y \mid X = x]$, random function $G_x$ evaluates the expected outcome-difference at the post vs. pre-intervention point $x$ (note the expectation here is over the noise $\varepsilon$, not our posterior). To infer the best personalized intervention, we use optimization (over vectors $T(x) \in \mathbb{R}^d$) to find:

$$T^*(x) = \arg\max_{T(x) \in C} F_{G_x(T)}^{-1}(\alpha)$$

(2)

where $F_{G_x(T)}^{-1}(\alpha)$ denotes the $\alpha$th quantile of our posterior distribution over $G(\cdot)$. We assume the choice $0 < \alpha < 0.5$, which implies the intervention that produces $T^*(x)$ should improve the expected outcome with probability $\geq 1 - \alpha$ under our posterior beliefs.
Defined based on known constraints of feasible interventions, the set $C_x \subset \mathbb{R}^d$ enumerates possible transformations that can be applied to an individual with feature values $x$. If the set of possible interventions is independent of $x$ (i.e. $C_x = C \forall x$), then our goal is similar to the optimal feature-configuration problem studied in Bayesian optimization. However, in many applications, $x$-independent transformations are not practically realizable. Consider gene perturbation, a scenario where it is impractical to simultaneously target more than a few genes due to technological limitations. If instead intervening on a quantity like caloric intake, it is only realistic to change an individual’s current value by at most a small amount. The choice $C_x := \{ z \in \mathbb{R}^d : |x - z|_0 \leq k \}$ reflects the constraint that at most $k$ features can be intervened upon. We can denote limits on the amount that the $s^{th}$ feature may be altered by $C_s := \{ z \in \mathbb{R}^d : |x_s - z_s| \leq \gamma_s \} \quad \text{for} \quad s \in \{1, \ldots, d\}$. In realistic settings, $C_s$ may be the intersection of many such sets reflecting other possible constraints such as boundedness, impossible joint configurations of multiple features, etc.

For any $x$, $T(x) \in \mathbb{R}^d$: the posterior distribution for $G_x(T)$ has mean $\mathbb{E}[f(T(x)) \mid D_n] - \mathbb{E}[f(x) \mid D_n]$ and variance $\text{Var}(f(T(x)) \mid D_n) + \text{Var}(f(x) \mid D_n) - 2\text{Cov}(f(T(x)), f(x) \mid D_n)$, which can be easily computed using the corresponding mean/covariance-functions of the posterior $f \mid D_n$. When $T(x) = x$, the objective in (2) takes value 0, so any optimum corresponds to an intervention we are confident will lead to expected improvement. If there is no good intervention in $C_x$ or too much uncertainty about $f$ given limited data, then this method simply returns $T^*(x) = x$ indicating no intervention should be performed.

### 3.2 Interventions for Entire Populations

The above discussion focused on personalized interventions tailored on an individual basis. In many applications, policy-makers are interested in designing a single intervention which will be applied to all individuals from the same underlying population as the data. Furthermore, relying on such a global policy is the only option in cases where we no longer observe feature-measurements of new individuals outside the data. In our gene perturbation example, gene expression may no longer be individually profiled in future cells/organisms that receive the decided-upon intervention due to cost/labor constraints.

In this context, the features $X$ are assumed distributed according to some underlying (pre-intervention) population, and we define the population expected gain function:

$$G_X(T) := \mathbb{E}_X[G_x(T)] = \mathbb{E}_X[f(T(x)) - f(x) \mid D_n]$$ \quad (3)

which is also randomly distributed based on our posterior (expectation $\mathbb{E}_X$ is with respect to the feature-distribution $X$ which is not modeled by $f \mid D_n$). Our goal is now to find a single transformation $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ corresponding to a population intervention which will (with high certainty under our posterior beliefs) lead to large outcome improvements on average across the population:

$$T^* = \arg \max_{T \in \mathcal{T}} F_{G_X(T)}^{-1}(\alpha)$$ \quad (4)

where the family of possible transformations $\mathcal{T}$ is constrained such that $T(x) \in C_x$ for all $T \in \mathcal{T}$, $x \in \mathbb{R}^d$. In practice, a good model of multivariate features is difficult to specify, so we instead simply work with an empirical estimate:

$$T^* = \arg \max_{T \in \mathcal{T}} F_{G_n(T)}^{-1}(\alpha)$$ \quad (5)

where

$$G_n(T) := \frac{1}{n} \sum_{i=1}^{n} [f(T(x^{(i)})) - f(x^{(i)})] \mid D_n$$ \quad (6)

is the empirical population expected gain, whose posterior mean and variance can again be easily evaluated using the mean and variance of $f\mid D_n$.

The population intervention objective in (5) is again 0 for the choice $T(x) = x$, so in the presence of high uncertainty or a dearth of good feasible interventions in $\mathcal{T}$, the policy produced by this method will again simply be to perform no intervention. The policy $T$ in this population intervention setting is designed under the assumption that future individuals will stem from the same underlying distribution as the samples in $D_n$. Furthermore, it is assumed that the desired transformation $T$ may be enacted exactly upon any individual from the population via external intervention (discussed further in §5.1).
Although $T$ is a function of $x$, the form of the transformation must be agnostic to the specific values of $x$ (so the intervention can be applied to new individuals without measuring their features).

We consider two types of transformations that we find widely applicable. Shift-interventions involve transformations of the form: $T(x) = x + \Delta$ where $\Delta \in \mathbb{R}^d$ represents a (sparse) shift that the policy applies to each individuals’ features (eg. always adding 3 to the value of the second feature corresponds to $T(x) = [x_1, x_2 + 3, \ldots, x_d]$). Uniform-interventions are policies which set certain features to a constant value for all individuals and involve transformations $T(x) = [z_1, \ldots, z_d]$ such that for some feature subset $I \subseteq \{1, \ldots, d\}$: $z_j = x_j \forall j \notin I$ (eg. always setting the first feature to 0 corresponds to $T(x) = [0, x_2, \ldots, x_d]$).

Figure 1: Contour plot of relationship $Y = X_1 \cdot X_2 + \varepsilon$ depicting outcomes $Y$ expected across feature space $[X_1, X_2]$. Black points: the underlying population. Red points: same population after global shift intervention $T(X) = X + [-3, 0]$. Gold diamond: optimal feature configuration if any transformation in the box is feasible. Light (or dark) green points (along border): best uniform intervention which can only set $X_2$ (or only $X_1$) to a fixed value. Blue, purple, and light blue points: individuals who receive a single-variable personalized intervention, the direction of the optimal transformation for each is shown.

Figure 1 depicts examples of these different interventions. Under a sparsity constraint, we must carefully model the underlying population in order to identify the best uniform intervention (for this population, setting $X_1$ to a large value is superior to intervening on $X_2$). Under the optimal sparse personalized interventions, different intervention-variables may be chosen for different individuals, and the direction of the transformation can vary significantly.

4 Algorithms

Throughout this work, we use Gaussian Process (GP) regression \[11\] to model $Y \mid X$ (the GP model is described in §S1 where references preceded by ‘S’ indicate locations in the Supplementary Material). This nonparametric method has been favored in many applications as it produces both accurate predictions and effective measures of uncertainty (with closed-form estimators available in the standard case). Furthermore, a variety of GP models exist for different settings including: non-Gaussian response variables \[11\], non-stationary relationships \[12\], deep architectures \[13\], measurement error \[14\], and heteroscedastic noise \[15\]. While these variants are not employed in this work, our methodology can be directly used in conjunction with such extensions, or more generally, any model which produces useful posterior estimates for $f \mid D_n$.

In the GP setting, the $\alpha$th quantile of our personalized gain is given by: $G_T(x) = \mathbb{E}[f(T(x)) - f(x) \mid D_n] + \Phi^{-1}(\alpha) \cdot \text{Var}[f(T(x)) - f(x) \mid D_n]$ (where $\Phi^{-1}$ is the $N(0, 1)$ quantile function), and the quantiles of the empirical population gain take a similar form. When our GP prior adopts a smooth covariance-function $k(\cdot, \cdot)$, derivatives of our intervention-objectives are easily computed with respect to $T(x)$. To find transformations which maximize our objectives, we employ the proximal gradient method \[16\] described in §S2. However, a greedy gradient method is susceptible to the possible presence of local maxima. To avoid severely suboptimal solutions, we propose a continuation technique that performs a series of optimizations over variants of our objective with tapering levels of exaggerated smoothness (details in §S2).

Due to cost/feasibility-constraints, an intervention that only affects a small subset of variables is desired in many practical settings. This implies we must consider the underlying distribution of the remaining untouched variables in the population when evaluating a particular intervention (as demonstrated in Figure 1). Note that identifying a sparse transformation of the features is different from feature selection in supervised learning (where the goal is to identify dimensions along which $f$ varies most). In contrast, we seek the dimensions $I \subseteq \{1, \ldots, d\}$ along which one of our feasible feature-transformations can produce the largest high-probability increase in $f$, assuming the other
features remain fixed at their initial pre-intervention values (in the case of personalized intervention) or follow the same distribution as the pre-intervention population (in the case of a global policy).

In shift-intervention settings where altering \( x_s \) (the \( s \)th feature for \( s \in \{1, \ldots, d\} \)) by one unit requires cost \( \gamma_s \), we take these costs into consideration by penalizing our intervention-objective with the following regularization term: \( \Omega_\lambda(T) = \lambda \sum_{s=1}^{d} \gamma_s |\Delta_s| \). Computing the full regularization path over different choices of \( \lambda \geq 0 \) provides insight on the trade-off between each unit of expected outcome improvement (with probability \( \geq 1 - \alpha \) according to our posterior belief) against the cost of intervening to produce this improvement. This can help analysts identify which variables to intervene upon, depending on the relative value of outcome-improvements.

In some applications, the per-feature costs of intervention may be unspecified, but it is clear that no more than \( k \) features should be intervened upon. Writing improvement software, for example, should not overwhelm authors with a multitude of desired changes, but rather present a concise list of the most important considerations. To find the best \( k \)-sparse shift intervention, we employ \( \ell_1 \)-relaxation\(^{17}\), again leveraging the regularization path for our \( \ell_1 \)-penalized objective (with all \( \gamma_s = 1 \)); details are provided in the Sparse Shift Algorithm of \( \S 2.1 \). Recall that in the case of personalized intervention, we simply optimize over vectors \( T(x) \in \mathcal{C}_x \) and thus any sparse \( T \) may be equivalently expressed as a sparse shift intervention. To find a uniform intervention which sets \( k \) of the features to particular fixed constants across all individuals from the population, we instead employ a forward step-wise selection algorithm (detailed in \( \S 2.2 \)), as the form of the optimization is not amenable to \( \ell_1 \)-relaxation in this case.

5 Theoretical Results

Consider the following basic conditions: (A4) all data lies in \( \mathcal{C} := [0,1]^d \). (A5) \( 0 < \alpha \leq 0.5 \). Throughout this section, we assume (A1)-(A5) hold, and the true underlying relationship is \( Y = f^*(X) + \epsilon \). Our results are with respect to the true improvement of an intervention \( G^*_{\chi}(T) := f^*(T(x)) - f^*(x) \), and thus any sparse \( T \) may be equivalently expressed as a sparse shift intervention. To find a uniform intervention which sets \( k \) of the features to particular fixed constants across all individuals from the population, we instead employ a forward step-wise selection algorithm (detailed in \( \S 2.2 \)), as the form of the optimization is not amenable to \( \ell_1 \)-relaxation in this case.

**Theorem 1.** Suppose we adopt a GP \( (0, k(x,x')) \) prior and the following conditions hold:

(A6) noise variables \( \epsilon^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \) (A7) there exist \( \rho > 0 \) such that the Hölder space \( C^\rho[0,1]^d \) has probability one under our prior (see \( \S 2.1 \)). (A8) \( f^* \) and any \( f \) supported by the prior are Lipschitz continuous over \( \mathcal{C} \) with constant \( L \) (A9) the density of our input features \( p_X \in [0,1]^d \) is bounded above and below over domain \( \mathcal{C} \).

Then, for all \( x, T(x) \in \mathcal{C} \): 
\[
\mathbb{E}_{\mathcal{D}_n} \left| F_{G^*_{\chi}(T)}^{-1}(\alpha) - G^*_\chi(T) \right| \leq \frac{C}{\alpha} \left( \frac{L + \frac{1}{\alpha}}{a} \right) \left[ \Psi_{f^*}(n) \right]^{1/(2(d+1))}
\]
where constant \( C \) depends on the prior and density \( p_X \) and we define:
\[
\Psi_{f}(n) := \begin{cases} 
\left[ \psi_{f^*}^{-1}(n) \right]^2 & \text{if } \psi_{f^*}^{-1}(n) \leq n^{d/(4\rho+2d)} \\
\left[ \psi_{f^*}^{-1}(n) \right]^{(4\rho+4d)/d} & \text{otherwise}
\end{cases}
\]
(7)

\( \psi_{f^*}^{-1}(n) \) is the (generalized) inverse of \( \psi_{f^*}(\epsilon) := \frac{\phi_{f^*}(\epsilon)}{\epsilon} \) which depends on the concentration function \( \phi_{f^*}(\epsilon) = \inf_{f \in \mathcal{H}_k} \|f\|_{\infty} \log (1 + \frac{\|f\|_{\infty}}{\epsilon}) \). \( \phi_{f^*} \) measures how well the RKHS of our GP prior \( \mathcal{H}_k \) approximates \( f^* \) (see \( \S 2.1 \) for more details). The expectation \( \mathbb{E}_{\mathcal{D}_n} \) is over the distribution of the data \( \{ (X^{(i)}, Y^{(i)}) \}_{i=1}^n \). Importantly, Theorem 1 does not assume anything about the true relationship \( f^* \), and the bound depends on the distance between \( f^* \) and our prior. When \( f^* \) is \( \rho \)-smooth, a typical bound is given by \( \psi_{f^*}^{-1}(n) = O(n^{-\min(\rho,\nu)}/(2\rho+4d)) \) if \( k \) is the Matérn kernel with smoothness parameter \( \nu \). When \( k \) is the squared exponential kernel and \( f^* \) is \( \beta \)-regular (in Sobolev sense), \( \psi_{f^*}^{-1}(n) = O((1/\log n)^{3\beta/2d}) \) \( \S 2.1 \).

**Theorem 2.** Under the assumptions of Theorem 1, for any \( T \) such that \( \Pr(T(X) \in \mathcal{C}) = 1 \):
\[
\mathbb{E}_{\mathcal{D}_n} \left| F_{G^*_{\chi}(T)}^{-1}(\alpha) - G^*_\chi(T) \right| \leq \frac{C}{\alpha} \left( \frac{d}{n} + \left( \frac{L + \frac{1}{\alpha}}{a} \right) \Psi_{f^*}(n) \right)^{1/(2(d+1))}
\]
Theorems 1 and 2 characterize the rate at which our personalized/population-intervention objectives are expected to converge to the true improvement (due to contraction of the posterior as \( n \) grows). Since these results hold for all \( T \), this implies the maximizer of our intervention-objectives will converge to the true optimal transformation as \( n \to \infty \) (under a reasonable prior). Complementing these results, Theorem 6 in §S6 ensures that optimizing our personalized intervention objective corresponds to improving a lower bound on the true improvement with high probability (regardless of \( n \)), when \( \alpha \) is small and \( f^* \) belongs to the RKHS of our prior. In this case, the optimal transformation inferred by our approach only worsens the actual expected outcome with low probability.

5.1 Imprecise Transformations and Causal Models

Until now, we assumed the desired transformation \( T \) is enacted exactly upon individuals from the population. When the transformation is only allowed to alter at most \( k < d \) features, this requires that we can intervene to alter only this subset without affecting the other feature-values. However, in some domains (such as our gene perturbation example when the profiled genes belong to the same regulatory network), the feature-transformation produced by a particular external intervention can only be roughly controlled. Let \( T_{I \to z} \) denote a uniform transformation which sets only a subset of features in \( I \subset \{1, \ldots , d\} \) to constant values \( z \in \mathbb{R}^{|I|} \) across all individuals in the population. We consider a scenario in which the intervention applied in hopes of achieving \( T_{I \to z} \) propagates outward to affect other features outside \( I \), formalized as the do operation of Pearl’s causal calculus [5]. Here, the initial distribution of \( X, Y \) is assumed to follow a (fully endogeneous) structural equation model (SEM) [5], and we suppose the population which arises post-intervention is distributed according to \( \text{do}(X_I = z_I) \) rather than our intended \( T_{I \to z}(X) \). The outcome \( Y \) is restricted to be a sink node of the causal DAG, so we can still write \( Y = f^*(X) + \varepsilon \).

**Theorem 3.** For some \( I \subseteq \{1, \ldots , d\} \), suppose the condition: (A10) \( \text{pa}(Y) \subseteq I \cup \text{desc}(I)^C \) holds. Then, for any uniform transformation \( T_{I \to z} \):

\[
E_{\text{do}(X_I = z_I)}\left[f^*(x)\right] - \mathbb{E}_X\left[f^*(x)\right] = \mathbb{E}_X\left[f^*(T_{I \to z}(x)) - f^*(x)\right]
\]

Here, \( \text{pa}(Y) \) denotes the variables which are parents of outcome \( Y \) in the underlying causal DAG, and \( \text{desc}(I)^C \) is the set of variables which are not descendants of variables in subset \( I \). For the next result, we define: \( I^* := \arg\min \{|I'| \text{ s.t. } \exists T_{I' \to z} \in \arg\max_{T_{I' \to z} : |I'| \leq k} \mathbb{E}_X\left[f^*(T_{I' \to z}(x)) - f^*(x)\right]\} \) as the intervention set corresponding to the optimal \( k \)-sparse uniform transformation (where in the case of ties, the set of smallest cardinality is chosen), if transformations were enacted exactly.

**Theorem 4.** Suppose the underlying DAG satisfies: (A11) No variable in \( \text{pa}(Y) \) is a descendant of other parents, ie. \( \exists j \in \text{pa}(Y) \text{ s.t. } j \in \text{desc}(\text{pa}(Y) \setminus \{j\}) \). Then, \( I^* \) satisfies (A10).

In the absence of extremely strong interactions between variables in \( \text{pa}(Y) \), the equality of Theorem 3 will also hold for \( I^* \) if \( |\text{pa}(Y)| \leq k \). For settings where sparse interventions elicit unintentional do-effects and the causal DAG meets condition (A11), Theorems 3 and 4 imply that, under complete certainty about \( f^* \), the (minimum cardinality) maximizer of our uniform intervention objective corresponds to an intervention that produces an equally good outcome change when applied to the population as a do-operation. Combined with Theorem 2, our results ensure that, even in this misspecified setting, the empirical maximizer of our sparse uniform intervention objective (5) produces (in expectation as \( n \to \infty \)) beneficial interventions for populations whose underlying causal relationships satisfy our conditions.

Despite its popularity, the standard do-calculus is only partly appropriate in many practical applications where off-target effects of interventions can be mitigated. Consider our example of modeling text documents using NLP features, where the desired transformation is to increase the polarity of the text. This can be accomplished by inserting additional positive adjectives, but such an intervention also increases articles’ word count. Alternatively, polarity may be increased by replacing certain words with more positive alternatives, an external intervention which would not affect word count. These examples illustrate two different external interventions to enact the same desired transformation, and neither introduces effects governed by the natural relationship between polarity and word-count (as the do-calculus would assume).
6 Results

§S3 contains an analysis of our approach on simulated data from simple outcome-feature relationships. The average improvement produced by our chosen interventions rapidly converges to the best possible value with increasing \( n \). In these experiments, sparse-interventions consistently alter the correct feature subset, and proposed transformations under our conservative \( \alpha = 0.05 \) criterion are much more rarely harmful than optima of the posterior mean function (ignoring uncertainty).

6.1 Performance in Causal Settings

Here, we investigate how effective our methods are in a causal SEM setting, where a proposed sparse population intervention is actually realized as a \( do \)-operation \[5\] and can therefore unintentionally affect other features in the post-intervention population. We generate data from an underlying linear non-Gaussian SEM, and select \( Y \) as a variable corresponding to a sink node in the corresponding causal DAG (see §S3.1 for details). Our approach to identify a beneficial sparse population intervention is compared with inferring the complete SEM using LinGAM \[20\] and subsequently identifying the optimal \( do \) in the inferred SEM. Note that LinGAM is explicitly designed for this setting, while both our method and the relied-upon Gaussian Process model are severely misspecified. Figures 2A and 2B demonstrate that the inferred best single-variable shift population intervention (under constraints on the magnitude of the shift) matches the performance of LinGAM (except for in rare cases with tiny sample sizes), despite the unfavorable nature of these experiments. Thus, we believe a supervised learning approach like ours is preferable in practical applications where interpreting the underlying causal structure is not as important as producing good outcomes (especially for higher dimensional data where causal inference is generally ineffective \[7\]).

![Figure 2: The average (solid) and 0.05\(^{th}\) quantile (dashed) expected outcome change produced by our method (red) vs LinGAM (blue) over 100 datasets drawn from two underlying SEMs chosen by \[20\]. The black dashed line indicates the best possible improvement in each case.](image)

6.2 Population Intervention for Gene Perturbation

Next, we applied our method to search for population interventions in observational yeast gene expression data \[21\]. We evaluated the effects of proposed interventions (restricted to single gene knockouts) over a set \( X \) of 10 transcription factors \( (n = 161) \) with the goal of down-regulating each of a set of 16 small molecule metabolism target genes, \( Y \). Results for all methods are compared to the actual expression change of the target gene found experimentally under individual knockouts of each transcription factor in \( X \). Compared to marginal linear regressions and multivariate linear regression, our method’s uncertainty prevents it from proposing harmful interventions, and the interventions it proposes are optimal or near optimal (Figure 3).

Insets (a) and (b) in Figure 3 show empirical marginal distributions between target gene \( TSL1 \) and members of \( X \) identified for knockout by our method (\( CIN5 \)) and marginal regression (\( GAT2 \)). From the linear perspective, these relationships are fairly indistinguishable, but only \( CIN5 \) displays a strong inhibitory effect in the knockout experiments. Inset (c) shows the empirical marginal for a harmful intervention proposed by multivariate regression for down-regulating \( GPH1 \), where the overall correlation is significantly positive, but the few lowest expression values (which influence our GP intervention objective the most) do not provide strong evidence of a large knockdown effect.
6.3 Personalized Intervention for Writing Improvement

Finally, we apply our personalized intervention methodology to the task of transforming a given news article into one which will be more widely-shared on social media. We use a dataset containing various features about individual Mashable articles along with their subsequent popularity in social networks [22] (detailed description / results of this analysis provided in §S5). We train a GP regressor on 5,000 articles labeled with popularity-annotations and evaluate sparse interventions on a held-out set of 300 articles based on changes they induce in article benchmark popularity, which we define in §S5. When $\alpha = 0.05$, the average benchmark popularity increase produced by our personalized intervention methodology is 0.59, whereas it statistically significantly decreases to 0.55 if $\alpha = 0.5$ is chosen. Thus, even given this large sample size, ignoring uncertainty appears detrimental for this application, and $\alpha = 0.5$ results in 4 articles whose benchmark popularity worsens post-intervention (compared to only 2 for $\alpha = 0.05$). Nonetheless, both methods generally produce very beneficial improvements in this analysis, as seen in Figure S3.

As an example of the personalization of proposed interventions, our method ($\alpha = 0.05$) generally proposes different sparse interventions for articles in the Business category vs. the Entertainment category. On average, the sparse transformation for business articles uniquely advocates decreasing global sentiment polarity and increasing word count (which are not commonly altered in the personalized interventions found for entertainment articles), whereas interventions to decrease title subjectivity are uniquely prevalent throughout the entertainment category. These findings appear intuitive (eg. critical business articles likely receive more discussion, and titles of popular entertainment articles often contain startling statements written non-subjectively as fact). Interestingly, the model also tends to advise shorter titles for business articles, but increasing the length for entertainment articles. Articles across all categories are universally encouraged to include more references to other articles / keywords that were historically popular.

7 Discussion

This work introduces methods for directly learning beneficial interventions from observational data without the intermediate causal inference. While this objective is, strictly speaking, only possible under stringent assumptions, our approach performs well in both intentionally-misspecified and complex real-world settings. Further theoretical and practical exploration of the intersection between predictive and causal modeling thus appears to be a valuable line of future research. Interesting extensions of this work might investigate stratified/budgeted interventions for populations or issues such as extrapolation beyond the data and robustness under imprecise transformation.

Acknowledgements

We thank David Gifford for helpful comments. DR was funded by an award from IARPA (under research contract 2015-15061000003).
References

[1] Duvenaud D, Eaton D, Murphy K, Schmidt M (2010) Causal learning without DAGs. *JMLR: Workshop and Conference Proceedings* 6: 177–190.

[2] Peters J, Bühlmann P, Meinshausen N (2015) Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*.

[3] Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z (2015) Prediction Policy Problems. *American Economic Review: Papers & Proceedings* 105: 491–495.

[4] Rojas-Carulla M, Schölkopf B, Turner R, Peters J (2016) Causal transfer in Machine Learning. *arXiv:150705333*.

[5] Pearl J (2000) Causality: Models, Reasoning and Inference. Cambridge Univ. Press.

[6] Dawid AP (2008) Beware of the dag! *JMLR: Workshop and Conference Proceedings* 6: 59–86.

[7] Peters J, Mooij JM, Janzing D, Schölkopf B (2014) Causal discovery with continuous additive noise models. *Journal of Machine Learning Research* 15: 2009-2053.

[8] Krishnan RG, Shalit U, Sontag D (2015) Deep kalman filters. *NIPS*.

[9] Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL (2015) Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics* 9: 247–274.

[10] Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N (2016) Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* 104: 148-175.

[11] Rasmussen CE (2006) Gaussian processes for machine learning. MIT Press.

[12] Paciorek CJ, Schervish MJ (2004) Nonstationary covariance functions for gaussian process regression. *NIPS*.

[13] Daminaou A, Lawrence A (2013) Deep gaussian processes. *AISTATS*.

[14] McHutchon A, Rasmussen CE (2011) Gaussian process training with input noise. *NIPS*.

[15] Le QV, Smola AJ, Canu S (2005) Heteroscedastic gaussian process regression. *ICML*.

[16] Bertsekas D (1995) Nonlinear Programming. Athena Scientific.

[17] Bach F, Jenatton R, Mairal J, Obozinski G (2012) Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* 4: 1-106.

[18] Srinivas N, Krause A, Kakade SM, Seeger M (2010) Gaussian process optimization in the bandit setting: No regret and experimental design. *ICML*.

[19] van der Vaart A, van Zanten H (2011) Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research* 12: 2095-2119.

[20] Shimizu S, Hoyer P, Hyvärinen A, Kerminen AJ (2006) A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7: 2003-2030.

[21] Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, et al. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* 157: 740–752.

[22] K Fernandes PV, Cortez P (2015) A proactive intelligent decision support system for predicting the popularity of online news. *Proceedings of the 17th EPIA Portuguese Conference on Artificial Intelligence*.
Supplementary Material

Contents

S1 Gaussian Process Regression ...................................................................... 10
S2 Algorithms ................................................................................................. 11
  S2.1 Sparse Shift Intervention .................................................................... 11
  S2.2 Sparse Uniform Intervention ............................................................... 12
S3 Simulations .................................................................................................. 13
  S3.1 Linear SEM Analysis ............................................................................ 14
S4 Gene Knockout Interventions ...................................................................... 15
S5 Interventions to Improve Article Popularity ............................................... 15
S6 Proofs and additional Theoretical Results .................................................. 18
  Notation and Definitions ............................................................................. 18
  Proof of Theorem 1 ..................................................................................... 18
  Proof of Theorem 2 ..................................................................................... 19
  Proof of Theorem 3 ..................................................................................... 20
  Proof of Theorem 4 ..................................................................................... 21
  Theorem 6 and Proof ................................................................................... 21

S1 Gaussian Process Regression

Gaussian Process regression [23] adopts a prior under which $f_{p,x_{q}}$, $f_{q}$ follow multivariate Gaussian distribution $N(m_{n}, K_{n,n})$ for any collection $\{x^{(i)}\}_{i=1}^{n}$. The model is specified by a prior mean function $m : R^{d} \rightarrow R$ and positive-definite covariance function $k : R^{d} \times R^{d} \rightarrow R$ which encodes our prior belief regarding properties of the underlying relationship between $X$ and $Y$ (such as smoothness or periodicity). Here, the vector $m_{n} \in R^{n}$ denotes the evaluation of function $m$ at each point $x^{(i)}$ and $K_{n,n}$ denotes the matrix whose $i,j$th component is $k(x^{(i)}, x^{(j)})$.

Given test input points $x^{(i)}$, $x^{(n)} \in R^{d}$ in addition to training data $D_{n}$, we additionally define: $f_{*} := [f(x_{s}^{(1)}), \ldots, f(x_{s}^{(n*))}]]$, $y_{n} = [y^{(1)}, \ldots, y^{(n)}]$, matrix $K_{s,n}$ with $i,j$th entry $k(x^{(i)}, x^{(j)})$ (where $x^{(i)}$ is the $i$th training input), and matrix $K_{s,s}$ which contains pairwise covariances between test inputs.

Assuming the noise $\varepsilon \sim N(0, \sigma^{2})$ is independently sampled for each observation, the posterior for $f$ at the test inputs, $f_{*} \mid D_{n}$, follows $N(\mu_{n*}, \Sigma_{n*})$ distribution with the following mean vector and covariance matrix:

$$\mu_{n*} = m_{*} + (K_{s,n} + \sigma^{2}I)^{-1}(y_{n} - m_{n}), \Sigma_{n*} = K_{s,s} - K_{s,n}(K_{n,n} + \sigma^{2}I)^{-1}K_{n,s}$$

Note that our intervention-optimization framework is not specific to this GP model, but can be combined with any algorithm for learning $f$ that provides reasonable uncertainty estimates. While employing a more powerful model should improve the results of our approach, comparing various
regressors is not our focus. Thus, all practical results of our methodology are presented using only the standard GP regression model, under which the posterior distribution over \( f \) is given by the above expressions. In each application presented here, our GP uses the Automatic-Relevance-Determination (ARD) covariance function, a popular choice for multi-dimensional data [23]:

\[
k(x, x') = \sigma_0^2 \exp \left[ -\frac{1}{2} \sum_{s=1}^{d} \left( \frac{x_s - x_s'}{l_s} \right)^2 \right]
\]  (8)

The ARD kernel relies on length-scale hyperparameters \( l_1, \ldots, l_d \) which determine how much \( f \) can depend on each dimension of the feature-space. All hyperparameters of our GP regression model (covariance-kernel parameters \( l_1, \ldots, l_d \) and \( \sigma_0 \) (the output variance) as well as the variance of the noise \( \sigma^2 \)) are empirically selected via marginal-likelihood maximization [23]. In each application, we employ the 0.05th posterior-quantile (\( \alpha = 0.05 \)) in our method to ensure that with high probability, the intervention it infers to be optimal induces a nonnegative change in expected outcomes.

S2 Algorithms

Note that for personalized interventions, any feature transformation may be re-expressed as a shift intervention (in terms of shift vector \( \Delta_x \) \( \in \mathbb{R}^d \) such that \( T(x) = x + \Delta_x \). For such a \( T \), we write \( \tilde{G}_x(\Delta_x) := G_x(T) \) for notational convenience. As a sparse personalized intervention on \( x \) is equivalent to a sparse \( \Delta_x \), we can employ the same algorithm used to identify sparse population shift interventions (details in §S2.1), only using a slightly different regularized objective function:

\[
J^\text{personal}_{\lambda}(\Delta_x) := \frac{1}{G^{\lambda}_{\alpha}(\Delta_x)}(\alpha) - \lambda \| \Delta_x \|_1
\]  (9)

which is maximized over the feasible set \( C_{\Delta_x} := \{ \Delta_x \in \mathbb{R}^d : x + \Delta_x \in C_x \} \).

To find an optimal transformation of our unregularized objective (\( \lambda = 0 \)), we employ Sequential Least Squares Programming [24]. Under \( \ell_1 \)-regularization (\( \lambda > 0 \)), we instead use the proximal gradient method [25]. At each iterate of this gradient ascent approach, we follow a step in the gradient direction by a soft-thresholding operation [26] as well as a projection back onto the feasible set \( C_{\Delta} \). Note that the \( \ell_1 \) penalty is equivalent to the cost-sensitive regularizer \( \Omega_\lambda \) (defined in §4) with \( \gamma_s = 1 \ \forall s \). By appropriately modifying the soft-thresholding operation, this same algorithm is employed to solve the more general setting with heterogeneous \( \gamma_s \).

However, the intervention objective \( J \) can be nonconvex in certain cases. To deal with local optima in acquisition functions, Bayesian optimization methods employ heuristics like combining the results of many local optimizers or operating over a fine partitioning of the feature space [27, 28]. We instead propose a continuation technique that solves a series of optimization problems, each of which operates on our objective under a smoothed posterior (and the amount of additional smoothing is gradually decreased to zero). Excessive smoothing of the posterior is achieved by simply considering GP models whose kernels are given overly large length-scale parameters. Each time the amount of smoothing is tapered, we initialize our local optimizer using the solution found at the previously greater smoothing level. Intuitively, the highly smoothed GP model is primarily influenced by the global structure in the data, and thus our optimization with respect to the posterior of this model is far less susceptible to low-quality local maxima. Analysis of a similar homotopy strategy under radial basis kernels has been conducted in [29].

S2.1 Sparse Shift Intervention

To find the best \( k \)-sparse population shift intervention, we resort to \( \ell_1 \) relaxation (details provided in the Sparse Shift Algorithm below). As the \( \ell_1 \)-norm provides the closest convex relaxation to the \( \ell_0 \) norm, this is a a commonly adopted strategy to avoid combinatorial search in feature selection [26]. First, we compute the regularization path over different settings of the penalty \( \lambda > 0 \) for the following regularized objective:

\[
J^\text{policy}_{\lambda}(\Delta) := \frac{1}{G^{\lambda}_{\alpha}(\Delta)}(\alpha) - \lambda \| \Delta \|_1
\]  (10)
which is maximized over the feasible set \( C_\Delta := \{ \Delta \in \mathbb{R}^d : x + \Delta \in C_x \text{ for all } x \in \mathbb{R}^d \} \) (we again write \( G_n(\Delta) := G_n(T) \) when \( T(x) = x + \Delta \)).

Subsequently, we identify the regularization penalty which produces a shift of desired cardinality and select our intervention set \( \mathcal{I} \) as the features which receive nonzero shift. Finally, we optimize the original unregularized objective \( (\lambda = 0) \) with respect to only the selected features in \( \mathcal{I} \) to remove bias induced by the regularizer (Step 3 below). Each inner maximization in both the Sparse Shift and Uniform Shift algorithms is performed via the proximal gradient methods combined with our continuation approach discussed above in [S2].

**Sparse Shift Algorithm:** Identifies best \( k \)-sparse shift intervention.

**Input:** Dataset \( D_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n \), Posterior \( f \mid D_n \)

**Parameters:** \( k \in \{1, \ldots, d\} \) specifies the maximal cardinality of the shift vector \( \Delta \in \mathbb{R}^d \), \( C_\Delta \subseteq \mathbb{R}^d \) is the set of feasible shifts (ignoring the sparsity constraint), \( J_\lambda \) is our empirical objective function: either the personalized or population objective in (9) or (10).

1. Perform binary search (over \( \lambda \)) to find:
   \[
   \lambda^* \leftarrow \arg\min \left\{ \lambda \geq 0 \text{ s.t. } \Delta^* := \arg\max_{\Delta \in C_\Delta} J_\lambda(\Delta) \text{ has } \leq k \text{ nonzero entries } \right\}
   \]

2. Define \( \mathcal{I} \leftarrow \text{support}(\Delta^*_{\lambda^*}) \subseteq \{1, \ldots, d\} \) where \( \Delta^*_{\lambda^*} := \arg\max_{\Delta \in C_\Delta} J_{\lambda^*}(\Delta) \)

3. **Return:** \( \Delta^* \in \mathbb{R}^d \leftarrow \arg\max_{\Delta \in B} J_0(\Delta) \) where \( B := C_\Delta \cap \{ \Delta \in \mathbb{R}^d : \Delta_i = 0 \text{ if } i \notin \mathcal{I} \} \)

### S2.2 Sparse Uniform Intervention

Here, our goal is to identify a uniform intervention which sets \( k \) of the features to particular fixed constants uniformly across all individuals from the population. We employ the forward step-wise selection algorithm described below, as the form of the optimization in this case is not amenable to \( \ell_1 \)-relaxation. Recall that \( \mathcal{I} \subseteq \{1, \ldots, d\} \) denotes the subset of features which are intervened upon, and the uniform intervention produces vector \( T_{\mathcal{I}-z}(x) \in \mathbb{R}^d \) such that \( T_{\mathcal{I}-z}(x)_s = x_s \) if \( s \notin \mathcal{I} \), otherwise \( T_{\mathcal{I}-z}(x)_s = z_s \) which is a constant chosen by the policy-maker. This same transformation is applied to each individual in the population, creating a more homogeneous group which share the same value for the features in \( \mathcal{I} \). For a given \( \mathcal{I} \), the objective function to find the best constants is:

\[
J_{\mathcal{I}}^{\text{unif}} \left( \{z_s\}_{s \in \mathcal{I}} \right) := F\left( G_{\alpha}^{-1}(T_{\mathcal{I}-z}) \right) \tag{11}
\]

with

\[
G_{\alpha}(T_{\mathcal{I}-z}) = \frac{1}{n} \sum_{i=1}^{n} \left[ f(z^{(i)}) - f(x^{(i)}) \right] \mid D_n \text{ where } z^{(i)} = \begin{cases} x^{(i)} &, \text{ if } s \notin \mathcal{I} \\ z_s &, \text{ otherwise} \end{cases}
\]

which is maximized over the constraints: \( z_s \in C_s \subseteq \mathbb{R} \) for \( s \in \mathcal{I} \).

**Sparse Uniform Intervention Algorithm:** Identifies best \( k \)-sparse uniform intervention.

**Input:** Dataset \( D_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n \), Posterior \( f \mid D_n \)

**Parameters:** \( k \in \{1, \ldots, d\} \) specifies the maximal number of features which may be set by the uniform intervention, \( C_1, \ldots, C_d \subseteq \mathbb{R} \) are sets of feasible settings for each feature.

1. Initialize \( \mathcal{I} \leftarrow \emptyset \), \( \mathcal{U} \leftarrow \{1, \ldots, d\} \), \( J^* \leftarrow 0 \)

2. While \( |\mathcal{I}| < k \):
   3. Set \( J^* \leftarrow \max_{C_\mathcal{I}, e \in \mathcal{U} \cup \{s\}} J_{\mathcal{I} \cup \{s\}}^{\text{unif}} \left( \{z_r\}_{r \in \mathcal{I} \cup \{s\}} \right) \) for each \( s \in \mathcal{U} \)
   4. Find \( s^* \leftarrow \arg\max_{s \in \mathcal{U}} \{J_s^*\} \)
   5. If \( J_{s^*}^* > J^* \):
      \( J^* \leftarrow J_{s^*}^* \), \( \mathcal{I} \leftarrow \mathcal{I} \cup \{s^*\} \), \( \mathcal{U} \leftarrow \mathcal{U} \setminus s^* \)
   6. Else: break

7. **Return:** \( \{z_s\}_{s \in \mathcal{I}} \leftarrow \arg\max_{C_\mathcal{I}, e \in \mathcal{I}} J_{\mathcal{I}}^{\text{unif}} \left( \{z_s\}_{s \in \mathcal{I}} \right) \)
**S3 Simulations**

Figure S1: The mean (solid) and 0.05th quantile (dashed) expected outcome change produced under personalized interventions suggested by various methods, over 100 datasets of each sample size. Each dataset contains 10-dimensional features, with $X_i \sim \text{Unif}[-1, 1]$, and $Y$ is determined by the indicated relationships and additive Gaussian noise ($\sigma = 0.2$). The black lines indicate the best possible expected outcome change (when the best change depends on which individual received the intervention, the black solid/dashed lines indicates the mean and 0.05th quantile over our 100 trials).

(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$

(B) Quadratic: $f(X) = 1 - X_1^2 - X_2^2$

(C) Product: $f(X) = X_1 \cdot X_2$

Figure S2: The mean (solid) and 0.05th quantile (dashed) expected outcome change produced by our population intervention method, over 100 datasets for each sample size (same setting as in Figure S1). The black line indicates the best possible expected outcome improvement.

(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$

(B) Quadratic: $Y = 1 - X_1^2 - X_2^2$
Over the simulated data summarized in Figure S1, we apply our basic personalized intervention method \((\alpha = 0.05)\) with purely local optimization (standard) and our continuation technique (smoothed), which significantly improves results. For each of the 100 datasets, we randomly sampled a new point (from the same underlying distribution) to receive a personalized intervention. The magnitude of each intervention is bounded by 1, except for in data from the quadratic relationship. We also infer sparse interventions (with a cardinality constraint of 2 for the linear and quadratic relationships, 1 for the product relationship). When \(Y = X_1 \cdot X_2 + \epsilon\), the optimal (constrained) intervention may drastically vary depending upon the individual’s feature-values, and our algorithm is able to correctly infer this behavior (Simulation C). Finally, we also apply a variant of our method which entirely ignores uncertainty \((\alpha = 0.5)\). While this approach is on average better for larger sample sizes, highly harmful interventions are occasionally proposed, whereas our uncertainty-adverse method \((\alpha = 0.05)\) is much less prone to producing damaging interventions (preferring to abstain by returning \(T(x) = x\) instead). This is an invaluable characteristic since interventions generally require effort and are only worth conducting when they are likely to produce a substantial benefit.

Figure S2 displays the behavior of both the population shift intervention in the linear setting, and the population uniform intervention under the quadratic relationship. The population intervention is notably safer than the individually tailored variants, producing no negative changes in our experiments.

S3.1 Linear SEM Analysis

Suppose that a desired transformation upon variable \(s \in \{1, \ldots, d\}\) cannot be enacted exactly and \(do\)-effects propagate to the descendants of \(s\) in the underlying DAG. This implies that the values of descendant variables are redrawn from the \(do\)-distribution which arises as a result of shifting \(x_s\). Because all relationships are linear in our SEMs, the actual expected outcome change resulting from a particular shift intervention (with \(do\)-effects on descendants of the intervened-upon variable) is easily obtained in closed form.

Our method is applied to the data to infer an optimal \(1\)-sparse shift population intervention (only interventions on a single variable are allowed). The maximal allowed magnitude of the shift is constrained to ensure the optimum is well-defined (to \(\pm 1\) times the standard deviation of each variable in the underlying SEM distribution). An alternative approach to improve outcomes in contrast to our black-box approach is to apply a causal inference method like LinGAM \([30]\) to estimate the SEM from the data, and then identify the optimal single-variable shift in the LinGAM-inferred SEM (since all inferred relationships are also linear, the optimal single-variable shift will be either 0 or the lower/upper allowed shift and we simply search over these possibilities). We compare our approach against LinGAM by evaluating the actual expected outcome change (resulting from a \(do\)-operation in the true SEM) produced by the shift interventions proposed under each method.

In our experiment, two underlying SEM models are considered which were used by the LinGAM authors to demonstrate the utility of their method (albeit with sample size = 10,000) \([30]\). \(SEM_A\) is used to refer to the model depicted in Figure 3 of \([30]\), where we define \(Y\) as \(x_6\) (a sink node in the causal DAG). \(SEM_B\) denotes the underlying model of Figure 4 in the same paper (\(Y\) is defined as sink node \(x_7\)). The remainder of the variables in each SEM are adopted as our observed features \(X\).

This experiment represents an application of our method in a highly misspecified setting. The true data-generating mechanism differs significantly from assumptions of our GP regressor (output noise is now fairly non-Gaussian, the underlying relationships are all linear while we use an ARD kernel). Furthermore, an intervention to transform a single feature incurs the entire magnitude of unintentional effects of a \(do\)-operation on downstream features in the SEM, whereas our method believes transformations are enacted exactly. In contrast, this data exactly follows the special assumptions required by LinGAM, and we properly account for inferred downstream \(do\)-operation effects when identifying the best inferred intervention under LinGAM. Furthermore, since LinGAM only estimates linear relations, the best inferred shift-intervention found by this approach will always be 0 or the minimal/maximal shift allowed for a particular feature. Searching over these three values for each feature ensures the actual optimal shift will be recovered if the LinGAM SEM-estimate were correct. However, under our approach, identifying the optimal population shift-intervention requires solving an optimization problem. Even if the GP regression posterior were to exactly reflect the true data-generating mechanism, our approach might get stuck in a suboptimal local maximum or avoid the minimal/maximal allowed shift due to too much uncertainty about \(f\) in the resulting region of feature-space. In practice, these potential difficulties do not pose much of an issue for our approach.
S4 Gene Knockout Interventions

The data set used for this analysis contains gene expression levels for a set of wild type (ie. ‘observational’) samples, \( D_{\text{obs}} \) (\( n = 161 \)), as well as for a set of ‘interventional’ samples, \( D_{\text{int}} \), in which each individual gene was serially knocked out. In our analysis, we search for potential interventions for affecting the expression of a desired target gene by training our GP regressor on \( D_{\text{obs}} \) and determining which knockout produces the best value of our empirical uniform population intervention objective (for down-regulating the target). Subsequently, we use \( D_{\text{int}} \) to evaluate the actual effectiveness of proposed interventions in the knockout experiments. We only search for interventions present in \( D_{\text{int}} \) (single gene knockouts) rather than optimizing to infer optimal feature transformations.

As candidate genes for this analysis we used only the 700 genes that classified as responsive mutants (at least four transcripts show robust changes in response to the knockout). Furthermore, we omitted genes whose expression over the 161 observational samples had standard deviation < 0.1. Out of the transcription factors present in the remaining set of genes, we defined the top 10 factors as our feature set \( X \), after ranking the transcription factors by the difference between their expression when they were knocked out in the interventional data and their 0.1\textsuperscript{th} quantile expression level in the observational data. This was to ensure that our model would be trained on data that at least resembled the experimental data \( D_{\text{int}} \). The set of genes to down-regulate was simply chosen to be those classified by as small molecule metabolism genes that met the minimum standard deviation requirement in their observational expression marginal distribution. The resulting set was 16 target genes, and the (negative) expression of each of was treated as an outcome \( Y \) in our analyses.

Each method evaluated in this analysis was to propose an intervention (single gene knockout) to down-regulate the expression of each target gene (separately). Once a gene to knock out was proposed, this intervention was evaluated by comparing the resulting expression of the target when the proposed knockout was actually performed in the experimental data \( D_{\text{int}} \). This expression level could then be compared to the ‘optimal’ choice of gene from \( X \) to intervene upon (the gene in \( X \) whose knockout produced the largest down-regulation of the target in \( D_{\text{int}} \)).

We compared our approach against two methods popularly used to draw conclusions about affecting outcomes in the sciences. First, we applied a multivariate regression analysis in which a linear regression model was fit to the observations of \( (X, Y) \) in \( D_{\text{obs}} \). The best gene to knockout was inferred on the basis of the regression coefficients and expression values (if no beneficial regression coefficient was found significant at the 0.05 level under the standard \( t \)-test, then no intervention was proposed). Second, we performed a marginal analysis in which separate univariate linear regression models were fit to \( (X_1, Y), \ldots, (X_d, Y) \), and the best knockout was again inferred on the basis of the regression coefficients and expression values (again, no intervention was recommended if there was no statistically significant beneficial regression coefficient at the 0.05 level, after correcting for multiple testing via the False Discovery Rate).

Figure 3 compares the results produced by these methods to the optimal intervention over \( X \) for down-regulating each \( Y \), as found in the experimental data \( D_{\text{int}} \). Of the 16 small molecule metabolism target genes tested, in three cases our method proposed an intervention which was found to be optimal or near optimal in \( D_{\text{int}} \), while in the remaining cases, the model uncertainty causes the method not to recommend any intervention (except for one very minorly harmful intervention for target \( \text{SAMJ} \)). On the other hand, neither form of linear regression proposed effective interventions for any target other than \( \text{FKS1} \), and in some cases, the linear regressors proposed counterproductive interventions that up-regulated the target. This highlights the importance of a model that properly accounts uncertainty when evaluating potential interventions.

S5 Interventions to Improve Article Popularity

We demonstrate our personalized intervention methodology in a setting with rich nonlinear underlying relationships. The data consist of 39,000 news articles published by Mashable around 2013-15. Each article is annotated with the number of shares it received in social networks (which we use as our outcome variable after log-transform and rescaling). A multitude of features have been extracted from each article (eg. word count, the category such as “tech” or “lifestyle”, keyword properties), many of which were produced by natural language processing algorithms (eg. subjectivity, polarity,
alignment with topics found by Latent Dirichlet Allocation) [32]. After removing many highly redundant features, we center and rescale all variables to unit-variance (see Table S2 for a complete description of the 29 features used in this analysis).

We randomly partition the articles into 3 disjoint groups: a training set (5,000 articles on which scaling-factors are computed and our GP regressor is trained), an improvement set (300 articles we find interventions for), and a held-out set (over 34,000 articles used for evaluation). A large group is left out for validation to ensure there are many near-neighbors for any given article, so we can reasonably estimate the true expected popularity given any setting of the article-features. Subsequently, a basic GP regression model is fitted to the training set. As the predictive power of our GP regressor does not measurably benefit from ARD feature-weighting, we simply use the squared exponential kernel. Over the held-out articles, the Pearson correlation between the observed popularity and the GP (posterior mean) predictions is 0.35. Furthermore, there is a highly significant ($p < 8 \cdot 10^{-41}$) positive correlation of 0.07 between the model’s predictive variance and the actual squared errors of GP predictions over this held-out set. Our model is thus able to make reasonable predictions of popularity based on the available features, and its uncertainty estimates tend to be larger in areas of the feature-space where the posterior mean lies further from actual popularity values.

In this analysis, we compare our personalized intervention methodology which rejects uncertainty (using $\alpha = 0.05$) with a variant of this approach that ignores uncertainty (using the same objective function with $\alpha = 0.5$). Both methods share the same GP regressor, optimization procedure, and set of constraints. For the 300 articles in the intervention set (not part of the training data) we allow intervening upon all features except for the article category which presumably is fixed from an author’s perspective. All feature-transformations are constrained to lie within [-2,2] of the original (rescaled) feature value, and we impose a sparsity constraint that at most 10 features can be intervened upon for a given article.

Unfortunately, no pre-and-post-intervention articles are available for us to ascertain a ground truth evaluation. To crudely measure performance, we estimate the underlying expected popularity of a given feature-setting using benchmark popularity: the (weighted) average observed popularity amongst 100 nearest neighbors (in the feature-space) from the set of held-out articles (with weights based on inverse Euclidean distance). Over our improvement set, the Pearson correlation between articles’ observed popularity and benchmark popularity is 0.28 (highly significant: $p \leq 2 \cdot 10^{-10}$). This approach thus appears to be, on average, a reasonable way to benchmark performance (even though nearest-neighbor held-out articles can individually differ from the text of a particular pre/post-intervention article despite sharing similar values of our 29 measured features).

Figure S3 depicts the results of our personalized intervention for each article in our intervention set. The expected improvement produced by a particular intervention is estimated as the difference between the benchmark popularity of the post-intervention feature-settings and the original feature-settings of the article receiving the personalized intervention. Table S1 summarizes these results. A paired-sample t-test suggests our method is significantly superior on average ($p < 2 \cdot 10^{-6}$).

To provide concrete examples, we present some articles of the Business and Entertainment categories (taken from our improvement set). For this business article: [http://mashable.com/2014/07/30/how-to-beat-the-heat/](http://mashable.com/2014/07/30/how-to-beat-the-heat/) our method proposes shifting the following 10 features (see Table S2 for feature descriptions):
| Method               | Mean   | Median | 0.05th Quantile | Num. Negative |
|---------------------|--------|--------|-----------------|---------------|
| Rejecting Uncertainty | 0.586  | 0.578  | 0.126           | 2             |
| Ignoring Uncertainty  | 0.552  | 0.555  | 0.105           | 4             |

Table S1: Summary statistics for the benchmark popularity change produced by each method over the 300 articles of the intervention set. The last column counts the number of harmful interventions (with change < 0).

num_hrefs: +2, num_self_hrefs: -1.25, average_token_length: -1.771, kw_avg_min: +1.71, kw_avg_avg: +2, self_reference_min_shares: +2, self_reference_max_shares: +1.68, self_reference_avg_shares: +2, global_subjectivity: +1.57, global_sentiment_polarity: -2

For this entertainment article: [http://mashable.com/2014/07/30/how-to-beat-the-heat/](http://mashable.com/2014/07/30/how-to-beat-the-heat/), our method proposes shifting the following 10 features:

average_token_length: -1.55, kw_avg_min: + 1.63, kw_avg_avg: +2, self_reference_min_shares: +2, self_reference_max_shares: +1.85, self_reference_avg_shares: +2.0, LDA_00: +1.63, LDA_01: -2, LDA_04: +0.82, global_subjectivity: +1.62

Indifferent to uncertainty, the method with $\alpha = 0.5$ advocates shifting all these features by the $\pm 2$ maximal allowed amounts, which leads to a 0.04 worse improvement in benchmark popularity compared with the feature changes specified above for this article.

| Feature                          | Description                                                                 |
|----------------------------------|-----------------------------------------------------------------------------|
| n_tokens_title                   | Number of words in the title                                               |
| n_tokens_content                 | Number of words in the content                                             |
| n_unique_tokens                  | Rate of unique words in the content                                       |
| n_non_stop_words                 | Rate of non-stop words in the content                                      |
| num_hrefs                        | Number of links                                                            |
| num_self_hrefs                   | Number of links to other articles published by Mashable                    |
| average_token_length             | Average length of the words in the content                                 |
| num_keywords                     | Number of keywords in the metadata                                        |
| data_channel_is_lifestyle        | Is the article category “Lifestyle”?                                       |
| data_channel_is_entertainment    | Is the article category “Entertainment”?                                   |
| data_channel_is_bus              | Is the article category “Business”?                                        |
| data_channel_is_socmed           | Is the article category “Social Media”?                                    |
| data_channel_is_tech             | Is the article category “Tech”?                                            |
| data_channel_is_world            | Is the article category “World”?                                           |
| kw_avg_min                       | Avg. shares of articles with the least popular keyword used for this article|
| kw_avg_max                       | Avg. shares of articles with the most popular keyword used for this article|
| kw_avg_avg                       | Avg. shares of the average-popularity keywords used for this article       |
| self_reference_min_shares        | Min. shares of referenced articles in Mashable                              |
| self_reference_max_shares        | Max. shares of referenced articles in Mashable                              |
| self_reference_avg_shares        | Avg. shares of referenced articles in Mashable                              |
| LDA_00                           | Closeness to first LDA topic                                               |
| LDA_01                           | Closeness to second LDA topic                                              |
| LDA_02                           | Closeness to third LDA topic                                               |
| LDA_03                           | Closeness to fourth LDA topic                                              |
| LDA_04                           | Closeness to fifth LDA topic                                               |
| global_subjectivity              | Subjectivity score of the text                                             |
| global_sentiment_polarity       | Sentiment polarity of the text                                             |
| title_subjectivity               | Subjectivity score of title                                                |
| title_sentiment_polarity         | Sentiment polarity of title                                                |

Table S2: The 29 features of each article (dimensions of $X$ in this analysis). Features involving the share-counts of other articles and LDA were based only on data known before the publication date.
S6 proofs and additional theoretical results

Notation and Definitions

All points $x \in \mathbb{R}^d$ lie in convex and compact domain $C \subset \mathbb{R}^d$.

C denotes constants whose value may change from line to line.

All occurrences of $f$ are implicitly referring to $f \mid D_n$.

$\mu_n(\cdot)$, $\sigma_2^2(\cdot)$, and $\sigma_n(\cdot, \cdot)$ respectively denote the mean, variance, and covariance function of our posterior for $f \mid D_n$ under the GP$(0, k(x, x'))$ prior.

$F_Z^{-1}(\alpha)$ denotes the $\alpha$th quantile of random variable $Z$.

$\Phi^{-1}(\cdot)$ denotes the $N(0, 1)$ quantile function.

$|| \cdot ||_k$ denotes the norm of reproducing kernel Hilbert space $\mathcal{H}_k$.

$B_\delta(x) \subset \mathbb{R}^d$ denotes the ball of radius $\delta$ centered at $x \in C$.

$I \subseteq \{1, \ldots, d\}$ represents the set of variables which are intervened upon in sparse settings.

$pa(Y)$ denotes the set of variables which are parents of variable $Y$ in a causal directed acyclic graph (DAG) (corresponding to a structural equation model) [33].

desc$(I)$ is the set of variables which are descendants of at least one variable in $I$ according to the causal DAG.

$A^C$ denotes the complement of set $A$.

The squared exponential kernel (with length-scale parameter $l > 0$) is defined:

$$k(x, x') = \exp \left( -\frac{1}{2l^2} ||x - x'||^2 \right)$$

The Matérn kernel (with another parameter $\nu > 0$ controlling smoothness of sample paths) is defined:

$$k(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} r^\nu B_\nu(r) \quad \text{where} \quad r = \frac{\sqrt{2\nu}}{l} ||x - x'||, B_\nu \text{ is a modified Bessel function}$$

Random variables $\varepsilon^{(1)}, \ldots, \varepsilon^{(n)}$ form a martingale difference sequence which is uniformly bounded by $\sigma$ if $E[\varepsilon^{(i)} \mid \varepsilon^{(i-1)}, \ldots, \varepsilon^{(1)}] = 0$ and $\varepsilon^{(i)} \leq \sigma \quad \forall i \in \mathbb{N}$.

A function $f$ is Lipshitz continuous with constant $L$ if: $|f(x) - f(x')| \leq L|x - x'|$ for every $x, x' \in C$.

Suppose $\rho > 0$ is expressed as $\rho = m + \eta$ for nonnegative integer $m$ and $0 < \eta \leq 1$.

The Hölder space $C^\alpha[0, 1]^d$ is the space of functions with existing partial derivatives of orders $(k_1, \ldots, k_d)$ for all integers $k_1, \ldots, k_d \geq 0$ satisfying $k_1 + \cdots + k_d \leq m$. Additionally, each function’s highest order partial derivative must form a function $h$ that satisfies: $|h(x) - h(y)| \leq C|x - y|^\eta$ for any $x, y$.

**Theorem 5** (van der Vaart & van Zanten [34]). Under the assumptions of Theorem 7

$$E_{D_n} \int_C [f(x) - f^*(x)]^2 p_X(x) \, dx \, d\Pi_n(f \mid D_n) \leq C \cdot \Psi_{f^*}(n)$$

where $\Psi_{f^*}(n)$ is defined as in Eq. (7). See [34] for a detailed discussion about this function.

**Proof of Theorem 5**

**Proof.** Recall $G_T(x) := f(T(x)) - f(x) \mid D_n$ depends on $f$. We fix $x_0, T(x_0) \in C$ and adapt the bound provided by Theorem 5 to show our result. Let $B_\delta(x) \subset C$ denote the ball of radius $0 < \delta < \frac{1}{2}$ centered at $x \in C$. We first establish the bound:

$$\int_C |f(x) - f^*(x)| p_X(x) \, dx$$
\[
\begin{align*}
&\geq \int_{B_2(x_0)} |f(x) - f^*(x)|p_X(x) \, dx + \int_{B_2(T(x_0))} |f(x) - f^*(x)|p_X(x) \, dx \\
&\geq a \cdot \text{Vol}(B_\delta) \left[ \min_{x \in B_\delta(x_0)} |f(x) - f^*(x)| + \min_{x \in B_\delta(T(x_0))} |f(x) - f^*(x)| \right] \\
&\geq a \cdot \text{Vol}(B_\delta) \cdot \left[ |f(T(x_0)) - f(x_0) - f^*(T(x_0)) - f^*(x_0)| - 8\delta L \right] \\
&\geq a \cdot \text{Vol}(B_\delta) \cdot \left[ G_{x_0}(T) - G^*_{x_0}(T) - 8\delta L \right]
\end{align*}
\]

where Vol(B_\delta) = O(\delta^d). Theorem 5 implies the following inequality (ignoring constant factors):

\[
[C \cdot \Psi_f(n)]^{1/2} \geq \left[ \mathbb{E}_{D_n} \int_{\mathcal{C}} [f(x) - f^*(x)]^2 \pi_X(x) \, dx \, d\Pi_n(f \mid D_n) \right]^{1/2}
\]

\[
\geq \mathbb{E}_{D_n} \int_{\mathcal{C}} [f(x) - f^*(x)] \pi_X(x) \, dx \, d\Pi_n(f \mid D_n) \quad \text{by Jensen’s inequality}
\]

\[
\geq a\delta^d \cdot \mathbb{E}_{D_n} \int G_{x_0}(T) - G^*_{x_0}(T) - \delta L \, d\Pi_n(f \mid D_n) \quad \text{via the bound from (12)}
\]

\[
= -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{D_n} \int_0^{\infty} \Pr (|G_{x_0}(T) - G^*_{x_0}(T)| \geq r) \, dr 
\]

\[
= -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{D_n} \int_0^{1} F^{-1}_{[G_{x_0}(T) - G^*_{x_0}(T)]}(\tilde{\alpha}) \, d\tilde{\alpha} 
\]

\[
\geq -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{D_n} \int_0^{1} F^{-1}_{G_{x_0}(T)}(\tilde{\alpha}) - G^*_{x_0}(T) \, d\tilde{\alpha} 
\]

\[
\geq -aL\delta^{d+1} + a(1 - \alpha)\delta^d \cdot \mathbb{E}_{D_n} \left[ F^{-1}_{G_{x_0}(T)}(\alpha) - G^*_{x_0}(T) \right]
\]

We can similarly bound \(G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\alpha)\):

\[
- aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{D_n} \int_0^{1} F^{-1}_{[G^*_{x_0}(T) - G_{x_0}(T)]}(\tilde{\alpha}) \, d\tilde{\alpha} 
\]

\[
\geq -aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{D_n} \int_0^{\alpha} G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\tilde{\alpha}) \, d\tilde{\alpha} 
\]

\[
\geq -aL\delta^{d+1} + a\alpha\delta^d \cdot \mathbb{E}_{D_n} \left[ G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\alpha) \right]
\]

Choosing \(\delta := [\Psi_f(n)]^{\frac{1}{d+1}}\) and combining (13) and (14) produces the desired result, since assuming \(\alpha < 0.5\) implies \(\alpha < 1 - \alpha\).

\[\square\]

**Proof of Theorem 2**

*Proof.* Combining the results of Lemmas 1 and 2 below, we obtain the desired upper bound through a straightforward application of the triangle inequality. Note that we’ve simplified the bound using the identity \(-\log(1 - \alpha) < 1/\alpha\) for \(\alpha < 0.5\).

\[\square\]

**Lemma 1.** Under the assumptions of Theorem 2 for any \(x, T(x) \in \mathcal{C}\):

\[
\mathbb{E}_{D_n} \left| F^{-1}_{G_{x}(T)}(\alpha) - F^{-1}_{G_{x}(T)}(\alpha) \right| \leq C \cdot \left[ \frac{L^2d}{n} \log(1 - \alpha) \right]^{1/2}
\]
Proof of Lemma 2. Define random variables $Z_i := f(T(x^{(i)})) - f(x^{(i)}) \mid D_n$ for $i = 1, \ldots, n$. Note that these variables all share the same expectation: $E_X[Z] := E_X[Z_i] = G_X(T)$ and $G_n(T) = \frac{1}{n} \sum_{i=1}^{n} Z_i$. The Lipschitz continuity of $f$ combined with the fact that $C = [0, 1]^d$ implies: $Z_i \in [-L \sqrt{d}, L \sqrt{d}]$ for all $i$. Thus, Hoeffding’s inequality ensures:

$$
Pr \left( \left| G_n(T) - G_X(T) \right| \geq t \right) \leq 2 \exp \left( \frac{-nt^2}{2L^2d} \right)
$$

$$
\Rightarrow F_{G_n(T) - G_X(T)}^{-1}(\alpha) \leq C \cdot \left[ \frac{-L^2d}{n} \log(1 - \alpha) \right]^{1/2}
$$

Because posteriors $G_n(T), G_X(T)$ follow a Gaussian distribution:

$$
F_{G_n(T)}^{-1}(\alpha) - F_{G_X(T)}^{-1}(\alpha) \leq F_{G_n(T) - G_X(T)}^{-1}(\alpha)
$$

and

$$
F_{G_X(T)}^{-1}(\alpha) - F_{G_n(T)}^{-1}(\alpha) \leq F_{G_n(T) - G_X(T)}^{-1}(\alpha)
$$

Lemma 2. Under the assumptions of Theorem 2 for any $x, T(x) \in C$:

$$
E_{D_n} \left| G_{X}(T) - G_{X}^{*}(T) \right| \leq \frac{C}{\alpha} \cdot \left( L + \frac{1}{a} \right) \cdot \left[ \Psi_{f*}(n) \right]^{1/2(d+1)}
$$

Proof of Lemma 2. A similar argument as the proof of Theorem 1 applies here. We again first bound:

$$
\int_C \left| f(x) - f^*(x) \right| p_X(x) \, dx
$$

$$
\geq a \cdot Vol(B_{\delta}) \cdot \left[ \int_C \left| f(x) - f^*(x) \right| p_X(x) \, dx + \int_C \left| f(T(x)) - f^*(T(x)) \right| p_X(x) \, dx - 8\delta L \right]
$$

$$
\geq a \cdot Vol(B_{\delta}) \cdot \left[ E_X[|f(x) - f^*(x)|] + E_X[f(T(x)) - f^*(T(x))] - 8\delta L \right]
$$

Following the same reasoning as in the proof of Theorem 1 we obtain (up to constant factors):

$$
-aL\delta^{d+1} + a\alpha \delta^d \cdot E_{D_n} \left[ G^*_X(T) - F_{G_X(T)}^{-1}(\alpha) \right] \leq \left[ C \cdot \Psi_{f*}(n) \right]^{1/2}
$$

and we can use the same argument to similarly bound

$$
E_{D_n} \left[ F_{G_X(T)}^{-1}(\alpha) - G^*_X(T) \right]
$$

Proof of Theorem 3

Here, we employ subscripts to index particular features of $X$. The notation $[a_R, a_S] = a \in \mathbb{R}^d$ is used to denote a vector assembled from disjoint subsets of dimensions $R, S \subseteq \{1, \ldots, d\}$. Regardless of the ordering of these partitions in our notation, we assume they are correctly arranged in the assembled vector based on their subscript-indices (ie. $a = [a_R, a_S] = [a_S, a_R]$).

Proof:

$$
E_{do(X_T = z)}[f^*(x)]
$$

$$
= \int f^*((x_{TC}, z_T)) \cdot p(x_{TC} \mid do(X_T = z_T)) \, dx_{TC}
$$
Theorem 6. Suppose we adopt a GP assumption that no variable in \( \text{pa}(Y) \) belongs the complement of desc assumption that no variable in \( \text{pa}(Y) \) since the marginal distribution over \( X \) equals the do-distribution by assumption (A10) does not improve this quantity. Thus, either \( \text{pa}(Y) \subseteq \mathcal{I}^* \), or \( \mathcal{I}^* \subset \text{pa}(Y) \). The first case immediately implies (A10). When \( \mathcal{I}^* \subset \text{pa}(Y) \): our assumption that no variable in \( \text{pa}(Y) \) is a descendant of other parents implies the other parents must belong the complement of desc(\( \mathcal{I}^* \)), since this is a subset of desc(\( \text{pa}(Y) \)).

Proof of Theorem 4

Recall we defined:

\[
\mathcal{I}^* := \arg\min \left\{ |\mathcal{I}| \mid \exists T_{\mathcal{I} \rightarrow Z} \in \arg\max_{T_{\mathcal{I} \rightarrow Z}} \mathbb{E}_{\mathcal{X}}\left[ f^*(T_{\mathcal{I} \rightarrow Z}(x)) - f^*(x) \right] \right\}
\]

as the intervention set corresponding to the optimal sparse uniform transformation (taken to be the set of minimal cardinality in cases with multiple maxima).

Proof. Since \( \mathbb{E}_{\mathcal{X}}\left[ f^*(T_{\mathcal{I} \rightarrow Z}(x)) \right] \) does not change when \( z_j := [T_{\mathcal{I} \rightarrow Z}(x)]_j \) is altered for any \( j \notin \text{pa}(Y) \), including variables outside of the parent set in \( \mathcal{I} \) does not improve this quantity. Thus, either \( \text{pa}(Y) \subseteq \mathcal{I}^* \), or \( \mathcal{I}^* \subset \text{pa}(Y) \). The first case immediately implies (A10). When \( \mathcal{I}^* \subset \text{pa}(Y) \): our assumption that no variable in \( \text{pa}(Y) \) is a descendant of other parents implies the other parents must belong the complement of desc(\( \mathcal{I}^* \)), since this is a subset of desc(\( \text{pa}(Y) \)).

Theorem 6 and Proof

Theorem 6. Suppose we adopt a GP \((0, k(x', x'))\) prior and, in addition to the assumptions outlined in 4 the following conditions hold: (A12) \( f^* \in \mathcal{H}_k(C) \) which is the RKHS induced by our covariance function \( k \) with norm \( \| \cdot \|_k \) (cf. [23] §6.1), (A13) noise variables \( \varepsilon^{(i)} \) form a uniformly bounded martingale difference sequence \( \varepsilon^{(i)} \leq \sigma \) for \( i = 1, \ldots, n \).

Then, for any \( x, T(x) \in C \) : \( F_{-1}^{-1}(\alpha) \leq G_x^*(T) \)

with probability (over the noise) greater than \( 1 - C(n + 1) \cdot \exp\left(-\frac{[\Phi^{-1}(\alpha)]^2 - 2\|f^*\|^2_k}{\gamma_n}\right) \)

In Theorem 6 \( \gamma_n := \max_{A \subseteq C, |A| = n} \frac{1}{2} \log \left| I + \sigma^{-2} K_A \right| \) is a kernel-dependent quantity \((K_A := [k(x, x')]_{x, x' \in A})\) which, in the Gaussian setting, is the mutual information between \( f \) and observations of \( Y \) at the most informative choice of \( n \) points. When the kernel satisfies \( k(x, x') \leq 1 \), the following bounds are known [35]: \( \gamma_n = O(d \log n) \) for the linear kernel, \( \gamma_n = O((\log n)^{d+1}) \) for the squared exponential kernel, and \( \gamma_n = O(n^d((d+1)/(2\nu+d+1))/((\log n))) \) for the Matérn kernel with smoothness parameter \( \nu \).

Note that while \( f^* \) is not required to be drawn from our prior and \( \varepsilon \) may be non-Gaussian, this result assumes the kernel \( k \) and noise-level \( \sigma \) are correctly set. Our proof relies on the following statement:
Theorem 7 (Srinivas et al. [35]). Assume conditions (A12) - (A13), fix $\delta \in (0, 1)$, and define:

$$\beta_n := 2||f^*||_2^2 + 300\gamma_n[\log(n/\delta)]^3$$

(see §5 for definition of $\gamma_n$)

Then:

$$\Pr \left[ \forall x \in C : |\mu_n(x) - f^*(x)| \leq \sqrt{\beta_{n+1}\sigma_n(x)} \right] \geq 1 - \delta$$

\begin{proof}
Fix $x, T(x) \in C$, and define $\delta := (n + 1) \cdot \exp\left(-\frac{[\Phi^{-1}(\alpha)]^2 - 2||f^*||_2^2}{300\gamma_n}\right)$.

In this case, $-\sqrt{\beta_{n+1}} = \Phi^{-1}(\alpha)$ (see definition in previous theorem).

Theorem 7 implies that with probability $\geq 1 - \delta$:

$$|\mu_n(x) - f^*(x)| \leq -\Phi^{-1}(\alpha) \cdot \sigma_n(x) \quad \text{and} \quad |\mu_n(T(x)) - f^*(T(x))| \leq -\Phi^{-1}(\alpha) \cdot \sigma_n(T(x))$$

Since our posterior is Gaussian:

$$F_{\tilde{G}_{\tau}(T)}(\alpha) = \mu_n(T(x)) - \mu_n(\alpha) + \Phi^{-1}(\alpha)\left[\sigma_n^2(T(x)) + \sigma_n^2(x) - 2\sigma_n(x, T(x))\right]^{1/2}$$

Therefore:

$$f^*(T(x)) - f^*(x) = F_{\tilde{G}_{\tau}(T)}^{-1}(\alpha)$$

$$= f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\left[\sigma_n^2(T(x)) + \sigma_n^2(x) - 2\sigma_n(x, T(x))\right]^{1/2}$$

$$\leq f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\left[\sigma_n^2(T(x)) + \sigma_n^2(x) + 2\sqrt{\sigma_n^2(x)\sigma_n^2(T(x))}\right]^{1/2}$$

since we assume $\alpha < 0.5 \Rightarrow \Phi^{-1}(\alpha) < 0$, and $|\sigma_n(x, T(x))| \leq \sqrt{\sigma_n^2(x)\sigma_n^2(T(x))}$

$$= f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\left[\sigma_n(T(x)) + \sigma_n(x)\right]$$

$$= [f^*(T(x)) - \mu_n(T(x)) - \Phi^{-1}(\alpha)\sigma_n(T(x))] + [\mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\sigma_n(x)]$$

which is less than 0 with probability at most $\delta$.
\end{proof}
Additional References for the Supplementary Information

[23] Rasmussen CE (2006) Gaussian processes for machine learning. MIT Press.
[24] Kraft D (1988) A software package for sequential quadratic programming. DLR German Aerospace Center - Institute for Flight Mechanics, Koln, Germany.
[25] Bertsekas D (1995) Nonlinear Programming. Athena Scientific.
[26] Bach F, Jenatton R, Mairal J, Obozinski G (2012) Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 4: 1-106.
[27] Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N (2016) Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE 104: 148-175.
[28] Lizotte DJ (2008) Practical Bayesian Optimization. Ph.D. thesis, University of Alberta.
[29] Mobahi H, L ZC, Ma Y (2012) Seeing through the blur. CVPR.
[30] Shimizu S, Hoyer P, Hyvärinen A, Kerminen AJ (2006) A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research 7: 2003-2030.
[31] Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, et al. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell 157: 740–752.
[32] K Fernandes PV, Cortez P (2015) A proactive intelligent decision support system for predicting the popularity of online news. Proceedings of the 17th EPIA Portuguese Conference on Artificial Intelligence.
[33] Pearl J (2000) Causality: Models, Reasoning and Inference. Cambridge Univ. Press.
[34] van der Vaart A, van Zanten H (2011) Information rates of nonparametric gaussian process methods. Journal of Machine Learning Research 12: 2095-2119.
[35] Srinivas N, Krause A, Kakade SM, Seeger M (2010) Gaussian process optimization in the bandit setting: No regret and experimental design. ICML.