Learning Optimal Interventions

Jonas Mueller **David Reshef** **George Du** **Tommi Jaakkola**
MIT Computer Science and Artificial Intelligence Laboratory

Abstract

Our goal is to identify beneficial interventions from observational data. We consider interventions that are narrowly focused (impacting few covariates) and may be tailored to each individual or globally enacted over a population. For applications where harmful intervention is drastically worse than proposing no change, we propose a conservative definition of the optimal intervention. Assuming the underlying relationship remains invariant under intervention, we develop efficient algorithms to identify the optimal intervention policy from limited data and provide theoretical guarantees for our approach in a Gaussian Process setting. Although our methods assume covariates can be precisely adjusted, they remain capable of improving outcomes in misspecified settings where interventions incur unintentional downstream effects. Empirically, our approach identifies good interventions in two practical applications: gene perturbation and writing improvement.

1 Introduction

In many data-driven applications, including medicine, the primary interest is identifying interventions that produce a desired change in some associated outcome. Due to experimental limitations, learning in such domains is commonly restricted to an observational dataset $\mathcal{D}_n := \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^n$ which consists of IID samples from a population with joint distribution $\mathbb{P}_{XY}$ over covariates (features) $X \in \mathbb{R}^d$ and outcomes $Y \in \mathbb{R}$. Typically, such data is analyzed using models which facilitate understanding of the relations between variables (eg. assuming linearity/additivity). Based on conclusions drawn from this analysis, the analyst decides how to intervene in a manner they confidently believe will improve outcomes.

Formalizing such beliefs via Bayesian inference, we develop a framework that identifies beneficial interventions directly from the data. In our setup, an intervention on an individual with pre-treatment covariates X produces post-treatment covariate values $\widetilde{X}$ that determine the resulting outcome Y (depicted as the graphical model: $X \to \widetilde{X} \to Y$). Each possible intervention results in a diffferent $\widetilde{X}$. More concretely, we make the following simplifying assumption:

$$Y = f(\widetilde{X}) + \varepsilon \quad \text{with } \mathbb{E}[\varepsilon] = 0, \varepsilon \perp\!\!\!\perp \widetilde{X}, X \tag{1}$$

for some underlying function f that encodes the effects of causal mechanisms (ie. $\widetilde{X}$ represents a fair description of the system state, and some covariates in $\widetilde{X}$ causally affect Y, not vice-versa). The observed data is comprised of naturally occurring covariate values where we presume $\widetilde{x}^{(i)} = x^{(i)}$ for $i = 1, \ldots, n$ (ie. the system state remains static without intervention, so the observed covariate values directly influence the observed outcomes). Moreover, we assume the relationship between these covariate values and the outcomes remains invariant, following the same (unknown) function f for any $\widetilde{X}$ arising from one of our feasible interventions (or no intervention at all). Note that this assumption precludes the presence of hidden confounding. Peters et al. (2016) have also relied on this invariance assumption, verifying it as a reasonable property of causal mechanisms in nature.

Given this data, we aim to learn an intervention policy defined by a covariate transformation $T : \mathbb{R}^d \to \mathbb{R}^d$, applied to each individual in the population. Here, $T(x)$ presents a desired setting of the covariates that should be reflected by subsequent intervention to actually influence outcomes. When T only specifies changes to a subset of the covariates, an intervention seeking to realize T may have unintended side-effects on covariates outside of this subset. We ignore such "fat hand" settings (Duvenaud et al. 2010) until §7. Instead, our methods assume interventions can always

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Flordia, USA. JMLR: W&CP volume 54.

be carried out with great precision to ensure the desired transformation T is exactly reflected in the post-treatment values: $\widetilde{x} = T(x)$. Our goal is to identify the transformation T which produces the largest corresponding post-treatment improvement with high certainty. $T(x)$ can either represent a single mapping to be performed on all individuals (global policy) or encode a personalized policy where the intervened upon variables and their values may change with x.

Our strong assumptions are made to ensure that statistical modeling alone suffices to identify beneficial interventions. While many real-world tasks violate these conditions, there exist important domains in which violations are sufficiently minor that our methods can discover effective interventions (cf. Rojas-Carulla et al. (2016), Peters et al. (2016)). We use two applications to illustrate our framework. One is a writing improvement task where the data consists of documents labeled with associated outcomes (eg. grades or popularity) and the goal is to suggest beneficial changes to the author. Our second example is a gene perturbation task where the expression of some regulatory genes can be up/down-regulated in a population (eg. cells or bacteria) with the goal of inducing a particular phenotype or activation/repression of a downstream gene. In these examples, covariates are known to cause outcomes and our other assumptions may hold to some degree, depending on the type of external intervention used to alter covariate values.

The contributions of this work include: (1) a formal definition of the optimal intervention that exhibits desirable characteristics under uncertainty due to limited data, (2) widely applicable types of (sparse) intervention policy that are easily enacted across a whole population, (3) algorithms to find the optimal intervention under practical constraints, (4) theoretical insight regarding our methods' properties in Gaussian Process settings as well as certain misspecified applications.

2 Related Work

The same invariance assumption has been exploited by Peters et al. (2016) and Rojas-Carulla et al. (2016) for causal variable selection in regression models. Recently, researchers such as Duvenaud et al. (2010) and Kleinberg et al. (2015) have supported a greater role for predictive modeling in various decision-making settings. Zeevi et al. (2015) use gradient boosting to predict glycemic response based on diet (and personal/microbiome covariates), and found they can naively leverage their regressor to select personalized diets which result in superior glucose levels than the meals proposed by a clinical dietitian. As treatment-selection in high-impact applications (eg. healthcare) grows increasingly reliant on supervised learning methods, it is imperative to properly handle uncertainty.

Nonlinear Bayesian predictive models have been employed by Hill (2011), Brodersen et al. (2015), and Krishnan et al. (2015) for quantifying the effects of a given treatment from observations of individuals who have been treated and those who have not. Rather than considering a single given intervention, we introduce the notion of an optimal intervention under various practical constraints, and how to identify such a policy from a limited dataset (in which no individuals have necessarily received any interventions).

Although our goals appear similar to Bayesian optimization and bandit problems (Shahriari et al. 2016, Agarwal et al. 2013), additional data is not collected in our setup. Since we consider settings where interventions are proposed based on all available data, acquisition functions for sequential exploration of the response-surface are not appropriate. As most existing data is not generated through sequential experimentation, our methods are more broadly applicable than iterative approaches like Bayesian optimization.

A greater distinction is our work's focus on the pre vs. post-intervention change in outcome for each particular individual, whereas Bayesian optimization seeks a single globally optimal configuration of covariates. In practice, feasible covariate transformations are constrained based on an individual's naturally occurring covariate-values, which stem from some underlying population beyond our control. For example in the writing improvement task, the goal is not to identify a globally optimal configuration of covariates that all texts should strive to achieve, but rather to inform a particular author of simple modifications likely to improve the outcome of his/her existing article. Appropriately treating such constraints is particularly important when we wish to prescribe a global policy corresponding to a single intervention applied to all individuals from the population (there is no notion of an underlying population in Bayesian optimization).

3 Methods

Our strategy is to first fit a Bayesian model for $Y \mid X$ whose posterior encodes our beliefs about the underlying function f given the observed data. Subsequently, the posterior for $f \mid \mathcal{D}_n$ is used to identify a transformation of the covariates $T : \mathbb{R}^d \to \mathbb{R}^d$ which is likely to improve expected post-intervention outcomes according to our current beliefs. The posterior for $f \mid \mathcal{D}_n$ may be summarized at any points $x, x' \in \mathbb{R}^d$ by mean function $\mathbb{E}[f(x) \mid \mathcal{D}_n]$ and covariance function

$\mathrm{Cov}(f(x), f(x') \mid \mathcal{D}_n)$.

3.1 Intervening at the Individual Level

For $x \in \mathbb{R}^d$ representing the covariate-measurements from an individual, we are given a set $\mathcal{C}_x \subset \mathbb{R}^d$ that denotes constraints of possible transformations of x. Let $T(x) = \widetilde{x} \in \mathcal{C}_x$ denote the new covariate-measurements of this individual after a particular intervention on x which alters covariates as specified by transformation $T : \mathbb{R}^d \to \mathbb{R}^d$. Recall that we assume an intervention can be conducted to produce post-treatment covariate-values that exactly match any feasible transformation: $\widetilde{x} = T(x)$, and we thus write $f(T(x))$ in place of $\mathbb{E}_\varepsilon[Y \mid \widetilde{X} = T(x)]$.

We first consider *personalized interventions* in which T may be tailored to a particular x. Under the Bayesian perspective, $f \mid \mathcal{D}_n$ is randomly distributed according to our posterior beliefs, and we define the *individual expected gain* function:

$$G_x(T) := f(T(x)) - f(x) \mid \mathcal{D}_n \tag{2}$$

Since $f(x) = \mathbb{E}_\varepsilon[Y \mid \widetilde{X} = x]$, random function G_x evaluates the expected outcome-difference at the post vs. pre-intervention setting of the covariates (this expectation is over the noise ε, not our posterior). To infer the best personalized intervention (assuming higher outcomes are desired), we use optimization over vectors $T(x) \in \mathbb{R}^d$ to find:

$$T^*(x) = \underset{T(x) \in \mathcal{C}_x}{\operatorname{argmax}} \ F^{-1}_{G_x(T)}(\alpha) \tag{3}$$

where $F^{-1}_{G(\cdot)}(\alpha)$ denotes the α^{th} quantile of our posterior distribution over $G(\cdot)$. We choose $0 < \alpha < 0.5$, which implies the intervention that produces $T^*(x)$ should improve the expected outcome with probability $\geqslant 1 - \alpha$ under our posterior beliefs.

Defined based on known constraints of feasible interventions, the set $\mathcal{C}_x \subset \mathbb{R}^d$ enumerates possible transformations that can be applied to an individual with covariate values x. If the set of possible interventions is independent of x (ie. $\mathcal{C}_x = \mathcal{C} \ \forall x$), then our goal is similar to the optimal covariate-configuration problem studied in Bayesian optimization. However, in many practical applications, x-independent transformations are not realizable through intervention. Consider gene perturbation, a scenario where it is impractical to simultaneously target more than a few genes due to technological limitations. If alternatively 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 $\mathcal{C}_x := \{z \in \mathbb{R}^d : ||x - z||_0 \leqslant k\}$ reflects the constraint that at most k covariates can be intervened upon. We can denote limits on the amount that the s^{th} covariate may be altered by $\mathcal{C}_x := \{z \in \mathbb{R}^d : |x_s - z_s| \leqslant \gamma_s\}$ for $s \in \{1, \ldots, d\}$. In realistic settings, $\mathcal{C}_x$ may be the intersection of many such sets reflecting other possible constraints such as boundedness, impossible joint configurations of multiple covariates, etc.

For any $x, T(x) \in \mathbb{R}^d$: the posterior distribution for $G_x(T)$ has:

$$\text{mean} = \mathbb{E}[f(T(x) \mid \mathcal{D}_n] - \mathbb{E}[f(x) \mid \mathcal{D}_n \tag{4}$$

$$\begin{aligned} \text{variance} = \mathrm{Var}(f(T(x)) \mid \mathcal{D}_n) + \mathrm{Var}(f(x) \mid \mathcal{D}_n) \\ - 2\mathrm{Cov}(f(T(x)), f(x) \mid \mathcal{D}_n) \end{aligned} \tag{5}$$

which is easily computed using the corresponding mean/covariance functions of the posterior $f \mid \mathcal{D}_n$. When $T(x) = x$, the objective in (3) takes value 0, so any superior optimum corresponds to an intervention we are confident will lead to expected improvement. If there is no good intervention in $\mathcal{C}_x$ (corresponding to a large increase in the posterior mean) or too much uncertainty about $f(x)$ given limited data, then our method simply returns $T^*(x) = x$ indicating no intervention should be performed.

Our objective exhibits these desirable characteristics because it relies on the posterior beliefs regarding both $f(T(x))$ and $f(x)$, which are tied via the covariance function. In contrast, a similarly-conservative lower confidence bound objective (ie. the UCB acquisition function with lower rather than upper quantiles) would only consider $f(T(x))$, and could propose unsatisfactory transformations where $\mathbb{E}[f(x) \mid \mathcal{D}_n] > \mathbb{E}[f(T(x)) \mid \mathcal{D}_n]$.

3.2 Intervening on Entire Populations

The above discussion focused on personalized interventions tailored on an individual basis. In certain 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. Relying on such a *global policy* is the only option in cases where we no longer observe covariate-measurements of new individuals outside the data. In our gene perturbation example, gene expression may no longer be individually profiled in future specimens that receive the decided-upon intervention to save costs/labor.

Here, the covariates X are assumed distributed according to some underlying (pre-intervention) population, and we define the *population expected gain* function:

$$G_X(T) := \mathbb{E}_X[G_x(T)] = \mathbb{E}_X\big[f(T(x)) - f(x) \mid \mathcal{D}_n\big]$$

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

$$T^* = \underset{T \in \mathcal{T}}{\text{argmax}} \ F^{-1}_{G_X(T)}(\alpha) \tag{6}$$

Here, the family of possible transformations $\mathcal{T}$ is constrained such that $T(x) \in \mathcal{C}_x$ for all $T \in \mathcal{T}, x \in \mathbb{R}^d$. As a good model of our multivariate features may be unknown, we instead work with the empirical estimate:

$$T^* = \underset{T \in \mathcal{T}}{\text{argmax}} \ F^{-1}_{G_n(T)}(\alpha) \tag{7}$$

$$\text{where} \quad G_n(T) := \frac{1}{n}\sum_{i=1}^{n}\left[f(T(x^{(i)})) - f(x^{(i)})\right] \ \mid \mathcal{D}_n$$

is the *empirical* population expected gain, whose posterior distribution has:

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(T(x^{(i)})) \mid \mathcal{D}_n] - \mathbb{E}[f(x^{(i)}) \mid \mathcal{D}_n] \tag{8}$$

$$\begin{aligned}\text{variance} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[&\text{Cov}\left(f(x^{(i)}), f(x^{(j)}) \mid \mathcal{D}_n\right) \\ &- \text{Cov}\big(f(T(x^{(i)})), f(x^{(j)}) \mid \mathcal{D}_n\big) \\ &- \text{Cov}\big(f(x^{(i)}), f(T(x^{(j)})) \mid \mathcal{D}_n\big) \\ &+ \text{Cov}\big(f(T(x^{(i)})), f(T(x^{(j)})) \mid \mathcal{D}_n\big)\Big]\end{aligned} \tag{9}$$

The population intervention objective in (7) is again 0 for the identity mapping $T(x) = x$. Under excessive uncertainty or a dearth of beneficial transformations in $\mathcal{T}$, the policy produced by this method will again simply be to perform no intervention. In this population intervention setting, T is designed assuming future individuals will stem from the same underlying distribution as the samples in $\mathcal{D}_n$. 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 covariates).

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' covariates (eg. always adding 3 to the value of the second covariate corresponds to $T(x) = [x_1, x_2 + 3, \ldots, x_d]$). *Covariate-fixing* interventions are policies which set certain covariates to a constant value for all individuals, and involve transformations $T_{\mathcal{I}\to z}(x) = [z_1, \ldots, z_d]$ such that for some covariate-subset $\mathcal{I} \subseteq \{1, \ldots, d\} : z_j = x_j \ \forall j \notin \mathcal{I}$ and for $j \in \mathcal{I}$: $z_j \in \mathbb{R}$ is fixed across all x (eg. always setting the first covariate to 0, for example in gene knockout, corresponds to $T(x) = [0, x_2, \ldots, x_d] \ \forall x$). 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 covariate-fixing intervention (here, setting X_1 to a large value is superior to intervening on X_2).

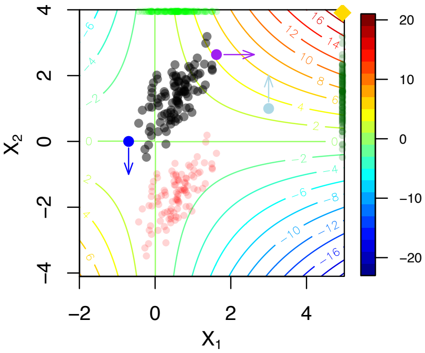


Figure 1: Contour plot of expected outcomes over feature space $[X_1, X_2]$ for relationship $Y = X_1 \cdot X_2 + \ \varepsilon$. Black points: the underlying population. Gold diamond: optimal covariate-setting if any transformation in the box were feasible. Red points: same population after shift intervention $\Delta = [-3, 0]$. Light (or dark) green points (along border): best covariate-fixing intervention which can only set X_2 (or only X_1) to a fixed value. Blue, purple, light blue points: individuals who receive a single-variable personalized intervention (arrows indicate direction of optimal transformation).

4 Algorithms

Throughout this work, we use Gaussian Process (GP) regression (Rasmussen 2006) to model $Y \mid X$ as described in §S1 ('S' indicates references 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 (Rasmussen 2006), non-stationary relationships (Paciorek & Schervish 2004), deep representations (Daminaou & Lawrence 2013), measurement error (McHutchon & Rasmussen 2011), and heteroscedastic noise (Le et al. 2005). 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 a useful posterior for $f \mid \mathcal{D}_n$).

Under the standard GP model, $G_x(T)$ follows a Gaussian distribution and the α^{th} quantile of our personalized gain is simply given by:

$$F^{-1}_{G_x(T)} = \mathbb{E}[G_x(T)] + \Phi^{-1}(\alpha) \cdot \mathrm{Var}[G_x(T)] \qquad (10)$$

where Φ^{-1} denotes the $N(0,1)$ quantile function. The quantiles of the empirical population gain may be similarly obtained. When a smooth smooth covariance kernel $k(\cdot,\cdot)$ is adopted in the GP prior, derivatives of our intervention-objectives are easily computed with respect to T.

In many practical settings, an intervention that only affects a small subset of variables is desired. Software to improve text, for example, should not overwhelm authors with a multitude of desired changes, but rather present a concise list of the most beneficial revisions in order to retain underlying semantics. Note that identifying a sparse transformation of the covariates 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 $\mathcal{I} \subset \{1,\dots,d\}$ along which one of our feasible covariate-transformations can produce the largest high-probability increase in f, assuming the other covariates remain fixed at their initial pre-treatment values (in the case of personalized intervention) or follow the same distribution as the pre-intervention population (in the case of a global policy).

For a shift intervention $T(x) = x + \Delta$, we introduce the convenient notation $G_n(\Delta) := G_n(T)$. In applications where shifting x_s (the s^{th} covariate for $s \in \{1,\dots,d\}$) by one unit incurs cost γ_s, we account for these costs by considering the following regularized intervention-objective:

$$J_\lambda(\Delta) := F^{-1}_{G_n(\Delta)}(\alpha) - \lambda \sum_{s=1}^{d} \gamma_s |\Delta_s| \qquad (11)$$

By maximizing this objective over feasible set $\mathcal{C}_\Delta := \{\Delta \in \mathbb{R}^d : x + \Delta \in \mathcal{C}_x \text{ for all } x \in \mathbb{R}^d\}$, policy-makers can decide which variables to intervene upon (and how much to shift them), depending on the relative value of outcome-improvements (specified by λ).

This optimization is performed using the proximal gradient method (Bertsekas 1995), where at each iterate: a step in the gradient direction is followed by a soft-thresholding operation (Bach et al. 2012) as well as a projection back onto the feasible set $\mathcal{C}_\Delta$. However, a simple gradient method may suffer from local optima. To avoid severely suboptimal solutions, we develop a continuation technique (Mobahi et al. 2012) that performs a series of gradient-based optimizations over variants of this objective with tapering levels of added smoothness (details in §S2).

In some settings, one may want to ensure at most $k < d$ covariates are intervened upon. We identify the optimal k-sparse shift intervention via the Sparse Shift Algorithm below, which relies on ℓ_1-relaxation (Bach et al. 2012) and the regularization path of our penalized objective in (13).

Sparse Shift Algorithm: Finds best k-sparse shift intervention.

```
Set $\gamma_s = 1$ for $s = 1, \dots, d$
Perform binary search over $\lambda$ to find:
    $\lambda^* \leftarrow \mathrm{argmin}\Big\{\lambda \geqslant 0 \text{ s.t. } \Delta^* := \underset{\Delta \in \mathcal{C}_\Delta}{\mathrm{argmax}}\, J_\lambda(\Delta)$
        $\text{has} \leqslant k \text{ nonzero entries}\Big\}$
Define $\mathcal{I} \leftarrow \mathrm{support}(\Delta^*_{\lambda^*}) \subseteq \{1, \dots, d\}$
    where $\Delta^*_{\lambda^*} := \underset{\Delta \in \mathcal{C}_\Delta}{\mathrm{argmax}}\, J_{\lambda^*}(\Delta)$
**Return:**    $\Delta^* \in \mathbb{R}^d \leftarrow \underset{\Delta \in B}{\mathrm{argmax}}\, J_0(\Delta)$
    where $B := \mathcal{C}_\Delta \bigcap \big\{\Delta \in \mathbb{R}^d : \Delta_s = 0 \text{ if } s \notin \mathcal{I}\big\}$
```

Recall that in the case of personalized intervention, we simply optimize over vectors $T(x) \in \mathcal{C}_x$. Any personalized transformation can therefore be equivalently expressed as a shift in terms of $\Delta_x \in \mathbb{R}^d$ such that $T(x) = x + \Delta_x$. After substituting the individual gain $G_x(\Delta_x)$ in place of the population gain $G_n(\Delta)$ within our definition of J_λ in (13), we can thus employ the same algorithms to identify sparse/cost-sensitive personalized interventions. To find a covariate-fixing intervention which sets k of the covariates to particular fixed constants across all individuals from the population, we instead employ a forward step-wise selection algorithm (detailed in §S2.2), as the form of the optimization is not amenable to ℓ_1-relaxation in this case.

5 Theoretical Results

Consider the following basic conditions: (A1) all data lies in $\mathcal{C} := [0,1]^d$, (A2) $0 < \alpha \leqslant 0.5$. Throughout this section, we assume (A1), (A2), and the conditions laid out in §1 hold. For clarity, we rewrite the true underlying relationship as f^*, letting f now denote arbitrary functions. Our results are with respect to the *true improvement* of an intervention $G^*_x(T) := f^*(T(x)) - f^*(x)$, $G^*_X(T) := \mathbb{E}_X[G^*_x(T)]$ (note that G^*_x, G^*_X are no longer random). Our theory relies on Gaussian Process results derived by Srinivas et al. (2010), van der Vaart & van Zanten (2011), and we relegate proofs and technical definitions to §S6.

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

(A3) noise variables $\varepsilon^{(i)} \overset{iid}{\sim} N(0,\sigma^2)$ *(A4) there exist*

$\rho > 0$ *such that the Hölder space* $C^{\rho}[0,1]^d$ *has probability one under our prior (see van der Vaart & van Zanten (2011)). (A5)* f^* *and any* f *supported by the prior are Lipschitz continuous over* $\mathcal{C}$ *with constant* L *(A6) the density of our input covariates* $p_X \in [a,b]$ *is bounded above and below over domain* $\mathcal{C}$*.*

Then, for all $x, T(x) \in \mathcal{C}$*:*

$$\mathbb{E}_{\mathcal{D}_n}\left|F^{-1}_{G_x(T)}(\alpha) - G^*_x(T)\right| \leqslant \frac{C}{\alpha}\left(L + \frac{1}{a}\right) \cdot \Psi_{f*}(n)^{1/[2(d+1)]}$$

where constant C depends on the prior and density p_X and we define:

$$\Psi_f(n) := \begin{cases} \left[\psi_f^{-1}(n)\right]^2 & \text{if } \psi_f^{-1}(n) \leqslant n^{-d/(4\rho+2d)} \\ n \cdot [\psi_{f*}^{-1}(n)]^{(4\rho+4d)/d} & \text{otherwise} \end{cases}$$

$\psi_{f*}^{-1}(n)$ is the (generalized) inverse of $\psi_{f*}(\epsilon) := \frac{\phi_{f*}(\epsilon)}{\epsilon^2}$ which depends on the concentration function $\phi_{f*}(\epsilon) = \inf_{h \in \mathcal{H}_k : ||h - f^*||_\infty < \epsilon} ||h||_k^2 - \log \Pi\big(f : ||f||_\infty < \epsilon\big)$. ϕ_{f*} measures how well the RKHS of our GP prior $\mathcal{H}_k$ approximates f^* (see van der Vaart & van Zanten (2011) 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 a ρ-smooth function, a typical bound is given by $\psi_{f*}^{-1}(n) = \mathcal{O}(n^{-\min\{\nu,\rho\}/(2\nu+d)})$ if k is the Matérn kernel with smoothness parameter ν. When k is the squared exponential kernel and f^* is β-regular (in Sobolev sense), $\psi_{f*}^{-1}(n) = \mathcal{O}((1/\log n)^{\beta/2 - d/4})$ (van der Vaart & van Zanten 2011).

Theorem 2. *Under the assumptions of Theorem 1, for any* T *such that* $\Pr(T(X) \in \mathcal{C}) = 1$*:*

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_n}\left|F^{-1}_{G_n(T)}(\alpha) - G^*_X(T)\right| \\ &\leqslant \frac{C}{\alpha}\left[L\sqrt{\frac{d}{n}} + \left(L + \frac{1}{a}\right)\Psi_{f*}(n)^{\frac{1}{2(d+1)}}\right] \end{aligned}$$

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 for any n: optimizing our personalized intervention objective corresponds to improving a lower bound on the true improvement with high probability, when α 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.

6 Results

§S3 contains an analysis of our approach on simulated data from simple covariate-outcome 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 those suggested by optimizing the posterior mean function (which ignores uncertainty).

6.1 Gene Perturbation

Next, we applied our method to search for population interventions in observational yeast gene expression data from Kemmeren et al. (2014). 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 2).

Insets (a) and (b) in Figure 2 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.

6.2 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 from Fernandes et al. (2015) containing various features about individual Mashable articles along with their subsequent popularity in social networks (detailed description/results for

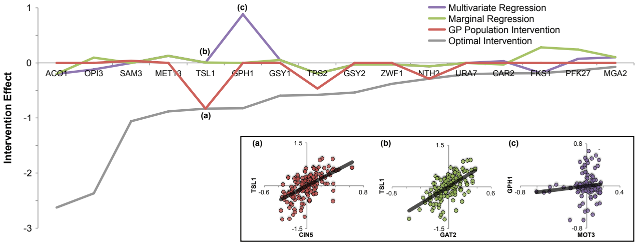


Figure 2: Actual effects of proposed interventions (single gene knockout) over a set transcription factors on down-regulation of each of a set of 16 small molecule metabolism target genes.

this analysis 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* (defined 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 and keywords that were historically popular.

7 Misspecified Interventions

Our methodology heavily relies on the assumption that the outcome-determining covariate values $\widetilde{x}$ produced through intervention exactly match the desired covariate transformation $T(x)$. When transformations are only allowed to alter at most $k < d$ covariates, this requires that we can intervene to alter only this subset without affecting the values of other covariates. If T specifies a sparse change affecting only a subset of the covariates $\mathcal{I} \subset \{1, \dots, d\}$, our methods assume the post-treatment value of any non-intervened-upon covariate remains at its initial value (ie. $\widetilde{x}_s = x_s \ \forall s \notin \mathcal{I}$).

In some domains, the covariate-transformation induced via sparse external intervention can only be roughly controlled (eg. our gene perturbation example when the profiled genes belong to a common regulatory network). Let $T_{\mathcal{I} \to z}$ denote a covariate-fixing transformation which sets a subset of covariates in $\mathcal{I} \subset \{1, \dots, d\}$ to constant values $z_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ across all individuals in the population. In this section, we consider an alternative assumption under which the intervention applied in hopes of achieving $T_{\mathcal{I} \to z}$ propagates downstream to affect other covariates outside $\mathcal{I}$ (so there may exist $s \notin \mathcal{I}$: $\widetilde{x}_s \neq x_s$), which we formalize as the *do*-operation in the causal calculus of Pearl (2000). Here, we suppose the underlying population of X, Y follows a *structural equation model* (SEM) (Pearl 2000). The outcome Y is restricted to be a sink node of the causal DAG, so we can still write $Y = f^*(\widetilde{X}) + \varepsilon$ and maintain the other conditions from §1. Rather than exhibiting covariate-distribution $T_{\mathcal{I} \to z}(X)$ with $Y = f^*(T_{\mathcal{I} \to z}(X)) + \varepsilon$ (as presumed in our methods), the post-treatment population which arises from an intervention seeking to enact transformation $T_{\mathcal{I} \to z}$ is now assumed to follow the distribution specified by $p(X, Y \mid do(X_{\mathcal{I}} = z_{\mathcal{I}}))$. Note that the *do*-operation here is only applied to some nodes in the DAG (variables in subset $\mathcal{I}$) as discussed by Peters et al. (2014), but its effects can alter the distributions of non-intervened-upon covariates outside of $\mathcal{I}$ which lie downstream in the DAG.

Theorem 3. *For some $\mathcal{I} \subseteq \{1, \dots, d\}$, suppose the condition: (A7) $pa(Y) \subseteq \mathcal{I} \bigcup desc(\mathcal{I})^C$ holds. Then, for any covariate-fixing transformation $T_{\mathcal{I} \to z}$: $\mathbb{E}_X\big[f^*(T_{\mathcal{I} \to z}(x)) - f^*(x)\big]$ and $\mathbb{E}_{\widetilde{x} \sim do(X_{\mathcal{I}} = z_{\mathcal{I}})}\big[f^*(\widetilde{x})\big] - \mathbb{E}_X\big[f^*(x)\big]$ are equal.*

Here, $\text{pa}(Y)$ denotes the variables which are parents of outcome Y in the underlying causal DAG, and $\text{desc}(\mathcal{I})^C$ is the set of variables which are *not* descendants of variables in subset $\mathcal{I}$. For the next result, we define: $\mathcal{I}^* := \text{argmin} \Big\{ |\mathcal{I}'| \text{ s.t. } \exists \ T_{\mathcal{I}' \to z} \in \underset{T_{\mathcal{I} \to z} : |\mathcal{I}| \leqslant k}{\text{argmax}} \ \mathbb{E}_X\big[f^*(T_{\mathcal{I} \to z}(x)) - f^*(x)\big] \Big\}$ as the intervention set corresponding to the optimal k-sparse covariate-fixing transformation (where in the case of ties, the set of smallest cardinality is chosen), if transformations were exactly realized by our interventions (which is not necessarily the case in this section).

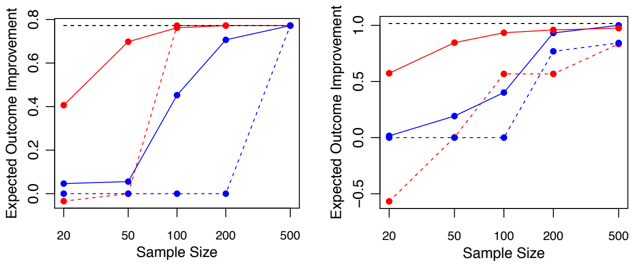
Theorem 4. *Suppose the underlying DAG satisfies: (A8) No variable in* $pa(Y)$ *is a descendant of other parents, ie.* $\nexists\ j \in pa(Y)$ *s.t.* $j \in desc(pa(Y)\backslash\{j\})$*. Then,* $\mathcal{I}^*$ *satisfies (A7).*

In the absence of extremely strong interactions between variables in $\text{pa}(Y)$, the equality of Theorem 3 will also hold for $\mathcal{I}^*$ if $|\text{pa}(Y)| \leqslant k$. For settings where sparse interventions elicit unintentional *do*-effects and the causal DAG meets condition (A8), Theorems 3 and 4 imply that, under complete certainty about f^*, the (minimum cardinality) maximizer of our covariate-fixing intervention objective corresponds to an transformation that produces an equally good outcome change when the corresponding intervention is actually realized as a *do*-operation in the underlying population. Combined with Theorem 2, our results ensure that, even in this misspecified setting, the empirical maximizer of our sparse covariate-fixing intervention objective (7) produces (in expectation as $n \to \infty$) beneficial interventions for populations whose underlying causal relationships satisfy certain conditions.

Next, we empirically investigate how effective our methods are in this misspecified SEM setting, where a proposed sparse population transformation is actually realized as a *do*-operation and can therefore unintentionally affect other covariates in the post-intervention population. We generate data from an underlying linear *non*-Gaussian SEM, and where Y is 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 the LinGAM estimator of Shimizu et al. (2006) and subsequently identifying the optimal single-node *do*-operation 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 3A and 3B demonstrate that the inferred best single-variable shift population intervention (under constraints on the magnitude of the shift) matches the performance the interventions suggested by LinGAM (except for in rare cases with tiny sample size) when the proposed interventions are evaluated as *do*-operations in the true underlying SEM. 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 estimation of the causal structure becomes difficult (Peters et al. 2014)).

The assumption of sparse interventions realized as a *do*-operation (as defined by Peters et al. (2014)) may also be an inappropriate in many domains, particularly if off-target effects of interventions are explicitly mitigated via external controls. To appreciate the intricate nature of assumptions regarding non-intervened-upon variables, consider our example of modeling text documents represented using two features: polarity and word count. A desired transformation to increase the text's polarity can be accomplished by inserting additional positive adjectives, but such an intervention also increases articles' word count. Alternatively, polarity may be identically increased by replacing words with more positive alternatives, an external intervention which would not affect the word count (and thus follows the assumptions of our framework).



(A) SEM$_A$ (B) SEM$_B$

Figure 3: The average (solid) and 0.05th quantile (dashed) expected outcome change produced by our method (red) vs LinGAM (blue) over 100 datasets drawn from two underlying SEMs chosen by Shimizu et al. (2006). The black dashed line indicates the best possible improvement in each case.

8 Discussion

This work introduces methods for directly learning beneficial interventions from purely observational data without treatments. While this objective is, strictly speaking, only possible under stringent assumptions, our approach performs well in both intentionally-misspecified and complex real-world settings. As supervised learning algorithms grow ever more popular, we expect intervention-decisions in many domains will increasingly rely on predictive models. Our conservative definition of the optimal intervention provides a principled approach to handle the inherent uncertainty in these settings due to finite data. Able to employ any Bayesian regressor, our ideas are widely applicable, considering practical types of interventions that can either be personalized or enacted uniformly over a population.

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

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L. & Schapire, R. (2013), 'Taming the monster: A fast and simple algorithm for contextual bandits', *30th International Conference on Machine Learning (ICML)* .

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

Bertsekas, D. (1995), *Nonlinear Programming*, Athena Scientific.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N. & Scott, S. L. (2015), 'Inferring causal impact using bayesian structural time-series models', *Annals of Applied Statistics* **9**, 247–274.

Daminaou, A. & Lawrence, A. (2013), 'Deep Gaussian processes', *16th International Conference on Artificial Intelligence and Statistics (AISTATS)* .

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

Fernandes, K., Vinagre, P. & Cortez, P. (2015), 'A proactive intelligent decision support system for predicting the popularity of online news', *17th EPIA Portuguese Conference on Artificial Intelligence* .

Hill, J. L. (2011), 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., ODuibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W. et al. (2014), 'Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors', *Cell* **157**(3), 740–752.

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review: Papers & Proceedings* **105**(5), 491–495.

Krishnan, R. G., Shalit, U. & Sontag, D. (2015), 'Deep kalman filters', *Advances in Neural Information Processing Systems (NIPS)* **28**.

Le, Q. V., Smola, A. J. & Canu, S. (2005), 'Heteroscedastic Gaussian process regression', *22nd International Conference on Machine Learning (ICML)* .

McHutchon, A. & Rasmussen, C. E. (2011), 'Gaussian process training with input noise', *Advances in Neural Information Processing Systems (NIPS)* **24**.

Mobahi, H., L, Z. C. & Ma, Y. (2012), 'Seeing through the blur', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .

Paciorek, C. J. & Schervish, M. J. (2004), 'Nonstationary covariance functions for Gaussian process regression', *Advances in Neural Information Processing Systems (NIPS)* **17**.

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

Peters, J., Bühlmann, P. & Meinshausen, N. (2016), 'Causal inference using invariant prediction: identification and confidence intervals', *Journal of the Royal Statistical Society: Series B* **78**, 1–42.

Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. (2014), 'Causal discovery with continuous additive noise models', *Journal of Machine Learning Research* **15**, 2009–2053.

Rasmussen, C. E. (2006), *Gaussian processes for machine learning*, MIT Press.

Rojas-Carulla, M., Schölkopf, B., Turner, R. & Peters, J. (2016), 'Causal transfer in machine learning', *arXiv:1507.05333* .

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. (2016), 'Taking the human out of the loop: A review of Bayesian optimization', *Proceedings of the IEEE* **104**(1), 148–175.

Shimizu, S., Hoyer, P., Hyvärinen, A. & Kerminen, A. J. (2006), 'A linear non-Gaussian acyclic model for causal discovery', *Journal of Machine Learning Research* **7**, 2003–2030.

Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. (2010), 'Gaussian process optimization in the bandit setting: No regret and experimental design', *27th International Conference on Machine Learning (ICML)* .

van der Vaart, A. & van Zanten, H. (2011), 'Information rates of nonparametric Gaussian process methods', *Journal of Machine Learning Research* **12**, 2095–2119.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M. et al. (2015), 'Personalized nutrition by prediction of glycemic responses', *Cell* **163**(5), 1079–1094.