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

We adapt principles of probabilistic black-box optimization for applications where learning is restricted to a single dataset rather than sequential experimentation. Our goal is to identify beneficial interventions given limited observations of an underlying population. We consider interventions that are narrowly focused (impacting few covariates) and may be tailored to each individual or globally enacted over the population. Proposing a conservative definition of the optimal intervention, we develop efficient algorithms for the optimization, and provide theoretical guarantees for our approach in a Gaussian Process setting.

1 Introduction

In many data-driven applications, including medicine, the primary interest is identifying interventions that produce a desired change in some associated outcome. Most existing data is, however, not generated through sequential experimentation, limiting applications of Bayesian optimization and bandit algorithms. 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 $X \in \mathbb{R}^d$ and outcomes $Y \in \mathbb{R}$. Rather than relying on interpretable models to summarize the underlying relationships (via simplifications such as linearity), we introduce a framework to identify the most 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 graphical model: $X \rightarrow \widetilde{X} \rightarrow Y$). We make the following simplifying assumption:

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

for some (unknown) function f that encodes the effects of causal mechanisms. The relationship between outcomes and covariate values is assumed to remain invariant, following the same f, ϵ whether the covariate values resulted from any of the interventions under consideration, or no intervention at all. This invariance property has been previously adopted as a reasonable assumption for mechanistic systems in the absence of serious confounding [1, 2]. Additionally, we suppose $\mathcal{D}_n$ is comprised of naturally occurring covariate values where all $\widetilde{x}^{(i)} = x^{(i)}$ (ie. covariates remain static without intervention, so the observed covariate values directly influence the observed outcomes).

Given this data, we aim to learn an intervention policy defined by a covariate transformation $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$, applied to each individual in the population. $T(x)$ presents a desired setting of the covariates that should be reflected by subsequent external intervention to actually influence outcomes. We assume precise intervention is possible to ensure any feasible T is exactly reflected in the post-treatment values: $\widetilde{x} = T(x)$. Our strong assumptions are made to ensure that statistical modeling alone suffices to identify beneficial interventions. While many real-world tasks severely violate these conditions, there exist important domains where violations are sufficiently minor and our methods

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

effectively identify a transformation to produce substantial post-treatment improvement in outcomes (§S8 studies a misspecified setting where our assumptions are wrong).

Related work. Various nonlinear Bayesian predictive models have been proposed for estimating treatment effects [3, 4, 5], but this previous work does not consider identifying the optimal intervention from limited data. Our goals appear similar to Bayesian optimization [6], but sequential exploration of the response-surface is not appropriate in our context. Whereas Bayesian optimization seeks a single globally optimal configuration of covariates, we focus on the pre vs. post-intervention change in outcome for each individual. 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, a system to help improve writing (eg. to increase the online popularity of an article [7]) should not simply output the single globally optimal text (with eg. higher expected popularity than any other writing), but should rather inform an author of simple modifications likely to improve the outcome of his/her existing article. Treating such constraints becomes particularly important when we wish to prescribe a single intervention applied to the entire population (there is no underlying population in Bayesian optimization).

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

Personalized Intervention. For $x \in \mathbb{R}^d$, we are given a set $\mathcal{C}_x \subset \mathbb{R}^d$ that denotes constraints of possible transformations of x. Let $\widetilde{x} = T(x) \in \mathcal{C}_x$ denote the new covariate-measurements of this individual after a particular intervention on x which alters covariates as specified by transformation T. Assuming $\widetilde{x} = T(x)$, we write $f(T(x)) = \mathbb{E}_\varepsilon[Y \mid \widetilde{X} = T(x)]$. We first consider *personalized interventions* in which T may be tailored to a particular x, and define the *individual expected gain*:

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

Under the Bayesian perspective, G_x is a random function which evaluates the expected outcome-difference at the post vs. pre-intervention setting of the covariates (where expectation $\mathbb{E}_\varepsilon$ is over the noise ε, not our posterior). To infer the best personalized intervention (supposing higher outcomes are desired), we use optimization over vectors $T(x) \in \mathbb{R}^d$ to find:

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

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

Defined based on known constraints of feasible interventions, the set $\mathcal{C}_x \subset \mathbb{R}^d$ enumerates possible interventions that can be applied to an individual with covariate values x. 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 $s \in \{1, \ldots, d\}$.

For any $x, T(x) \in \mathbb{R}^d$: the posterior for $G_x(T)$ has mean = $\mathbb{E}[f(T(x) \mid \mathcal{D}_n] - \mathbb{E}[f(x) \mid \mathcal{D}_n]$,

$$\text{variance} = \text{Var}(f(T(x)) \mid \mathcal{D}_n) + \text{Var}(f(x) \mid \mathcal{D}_n) - 2\text{Cov}(f(T(x)), f(x) \mid \mathcal{D}_n)$$

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. Note that our objective exhibits the aforementioned 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 (akin to acquisition functions from Bayesian Optimization) would only consider $f(T(x))$, and can propose unsatisfactory transformations where $\mathbb{E}[f(x) \mid \mathcal{D}_n] > \mathbb{E}[f(T(x)) \mid \mathcal{D}_n]$.

Population Intervention. 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.

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

$$G_X(T) := \mathbb{E}_X[G_x(T)] = \mathbb{E}_X\big[f(T(x)) - f(x) \mid \mathcal{D}_n\big] \tag{4}$$

which is also randomly distributed based on our posterior ($\mathbb{E}_X$ is expectation with respect to 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. However, the multivariate distribution of X is unknown and difficult to model in practice, so we find the optimal population intervention using an empirical estimate:

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmax}} \; F^{-1}_{G_n(T)}(\alpha) \tag{5}$$

where $G_n(T) := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\big[f(T(x^{(i)})) - f(x^{(i)})\big] \;\mid \mathcal{D}_n$ is the *empirical* population expected gain,

whose posterior distribution has mean $= \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbb{E}[f(T(x^{(i)})) \mid \mathcal{D}_n] - \mathbb{E}[f(x^{(i)}) \mid \mathcal{D}_n]$

$$\begin{aligned}\text{variance} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[&\text{Cov}\left(f(x^{(i)}), f(x^{(j)}) \mid \mathcal{D}_n\right) - \text{Cov}\left(f(T(x^{(i)})), f(x^{(j)}) \mid \mathcal{D}_n\right) \\ &- \text{Cov}\left(f(x^{(i)}), f(T(x^{(j)})) \mid \mathcal{D}_n\right) + \text{Cov}\left(f(T(x^{(i)})), f(T(x^{(j)})) \mid \mathcal{D}_n\right)\Big]\end{aligned} \tag{6}$$

$\mathcal{T}$, the family of possible transformations, is constrained such that $T(x) \in \mathcal{C}_x$ for all $T \in \mathcal{T}, x \in \mathbb{R}^d$. Again, the population intervention objective in (5) is 0 if $T(x) = x$, so the resulting policy is to perform no intervention under excessive uncertainty or a dearth of favorable transformations in $\mathcal{T}$. 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 global policies 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 $T(x) = [x_1, x_2 + 3, \ldots, x_d]\ \forall x$). *Uniform*-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 $\mathcal{I} \subset \{1, \ldots, d\} : z_j = x_j\ \forall j \notin \mathcal{I}$ (eg. always setting the first covariate to 0, for example in a gene knockout, corresponds to $T(x) = [0, x_2, \ldots, x_d]\ \forall x$).

3 Algorithms

Able to utilize any Bayesian predictive model, our framework is demonstrated using Gaussian Process (GP) regression [8] to model $Y \mid X$ in this work (see §S2). 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:

$$F^{-1}_{G_x(T)} = \mathbb{E}[G_x(T)] + \Phi^{-1}(\alpha) \cdot \text{Var}[G_x(T)] \tag{7}$$

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

In many practical settings, an intervention that only affects a small subset of variables is desired. 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 $s \in \{1, \ldots, d\}$) by one unit incurs cost γ_s, we account for these costs by considering the following regularized intervention-objective:

$$J_\lambda(\Delta) := F^{-1}_{G_n(\Delta)}(\alpha) - \lambda \sum_{s=1}^{d} \gamma_s |\Delta_s| \tag{8}$$

By maximizing J_λ over $\mathcal{C}_\Delta := \{\Delta \in \mathbb{R}^d : x + \Delta \in \mathcal{C}_x \text{ for all } x \in \mathbb{R}^d\}$, 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 [9], where at each iterate: a step in the gradient direction is followed by a soft-thresholding operation [10] as well as a projection back onto the feasible set $\mathcal{C}_\Delta$. To avoid poor local optima, we employ a continuation technique [11] that performs a series of gradient-based optimizations over variants of this objective with tapering levels of added smoothness (details in §S3).

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 in §S3.1, which relies on ℓ_1-relaxation [10] and the regularization path of our penalized objective in (8). To find a uniform intervention that sets k of the covariates to particular fixed constants across the entire population, we instead employ a forward step-wise selection algorithm (detailed in §S3.2). 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 (8), we can thus employ the same algorithms to identify sparse/cost-sensitive personalized interventions.

4 Results

Theoretical. Theorems 1 and 2 in §S4 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 [12]). These theorems imply the maximizer of our intervention-objectives will converge to the true optimal transformation as $n \to \infty$ (under a reasonable prior).

Empirical. In §S5, we apply our approach to simulated data from simple outcome-covariate relationships. The average improvement produced by our chosen interventions rapidly approaches the best possible value as n grows. We find that sparse-interventions consistently alter the correct covariate subset, and proposed transformations under our conservative choice $\alpha = 0.05$ are much more rarely harmful than those produced by optimizing the posterior mean (ignoring uncertainty).

In §S6, we consider the task of proposing a transcription factor gene to knockdown in order to down-regulate a target gene. Using expression data from [13], our sparse population intervention methodology is able to identify better knockdown candidates than frequentist linear models.

In §S7, we apply our sparse personalized intervention methodology to a writing improvement task [7], where it appears to be highly effective. Here, the choice of quantile $\alpha = 0.05$ produces much better results than $\alpha = 0.5$ (which is equivalent to simply optimizing predictive mean).

Misspecified Setting. In §S8, we suppose (X, Y) follow a structural equation model in which sparse interventions are realized as a *do*-operation [14]. Thus, an intervention on one covariate can affect the values of the other covariates (which is a violation of our assumptions: $\widetilde{x} \neq T(x)$). Nonetheless, we show theoretically and empirically that our methods can still work effectively in this regime.

Discussion. This work introduces methods for directly learning beneficial interventions from observational data rather than sequential experimentation. While this objective is, strictly speaking, only possible under stringent assumptions, our approach performs well in both intentionally-misspecified and complex real-world settings. Adopting a similar philosophy, [15] recently used 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 are superior to those proposed by a clinical dietitian. However, as treatment-selection in high-impact applications (eg. healthcare) grows increasingly reliant on supervised learning, it is imperative to properly handle uncertainty/constraints, and this work provides a principled approach.

References

[1] 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.

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

[3] Hill JL (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20: 217-240.

[4] 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.

[5] Krishnan RG, Shalit U, Sontag D (2015) Deep kalman filters. *Advances in Neural Information Processing Systems (NIPS)* 28.

[6] 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.

[7] 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* .

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

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

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

[11] Mobahi H, L ZC, Ma Y (2012) Seeing through the blur. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .

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

[13] 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.

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

[15] Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, et al. (2015) Personalized nutrition by prediction of glycemic responses. *Cell* 163: 1079–1094.

Supplementary Material

Contents

S1 Types of Intervention

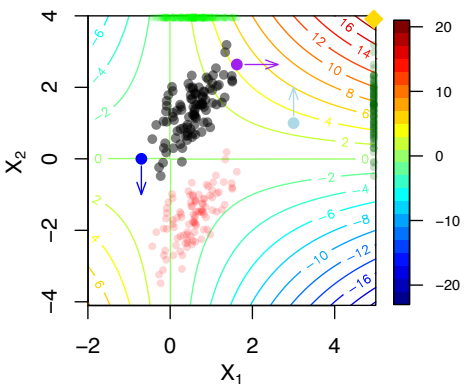


Figure S1: Contour plot of relationship $Y = X_1 \cdot X_2 + \varepsilon$ depicting outcomes Y expected across covariate-space $[X_1, X_2]$. Black points: the underlying population. Red points: same population after global shift intervention $T(X) = X + [-3, 0]$. Gold diamond: optimal feature configuration 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 S1 depicts examples of the different interventions introduced in this work. 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.

S2 Gaussian Process Regression

Gaussian Process regression [16] adopts a prior under which $f(x^{(1)}), \ldots, f(x^{(n)})$ follow multivariate Gaussian distribution $\mathrm{N}(\mathbf{m}_n, \mathbf{K}_{n,n})$ for any collection $\{x^{(i)}\}_{i=1}^n$. The model is specified by a prior mean function $m : \mathbb{R}^d \to \mathbb{R}$ and positive-definite covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which encodes our prior belief regarding properties of the underlying relationship between X and Y

(such as smoothness or periodicity). Here, the vector $\mathbf{m}_n \in \mathbb{R}^n$ denotes the evaluation of function m at each point $\{x^{(i)}\}_{i=1}^n$, and $\mathbf{K}_{n,n}$ denotes the matrix whose i,j^{th} component is $k(x^{(i)}, x^{(j)})$. Given test input points $x_*^{(1)}, \ldots, x_*^{(n_*)} \in \mathbb{R}^d$ in addition to training data $\mathcal{D}_n$, we additionally define: $\mathbf{f}_* := [f(x_*^{(1)}), \ldots, f(x_*^{(n_*)})]$, $\mathbf{y}_n = [y^{(1)}, \ldots, y^{(n)}]$, matrix $\mathbf{K}_{n,*}$ with i,j^{th} entry $k(x^{(i)}, x_*^{(j)})$ (where $x^{(i)}$ is the i^{th} training input), and matrix $\mathbf{K}_{*,*}$ which contains pairwise covariances between test inputs.

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

$$\mu_{\mathbf{n}*} = \mathbf{m}_* + (\mathbf{K}_{n,n} + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_n - \mathbf{m}_n),\ \mathbf{\Sigma}_{\mathbf{n}*} = \mathbf{K}_{*,*} - \mathbf{K}_{*,n}(\mathbf{K}_{n,n} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{n,*}$$

Note that our intervention-optimization framework is not specific to this basic GP model, but can be combined with any algorithm that learns a reasonable posterior for f, such as the superior GP variants designed for for nonstandard settings including: non-Gaussian response variables [16], non-stationary relationships [17], deep architectures [18], measurement error [19], and heteroscedastic noise [20]. As comparing various regressors is not our focus, our methodology is evaluated using only the standard GP regression model, under which the posterior distribution over f is given by the above expressions. Our GP uses the Automatic-Relevance-Determination (ARD) covariance function, a popular choice for multivariate data [16]. All hyperparameters of the GP are empirically selected via marginal-likelihood maximization [16].

S3 Algorithmic Details

To find an optimal transformation of our regularized objective J_λ in (8), we employ the proximal gradient method described in §3. When $\lambda = 0$ and there is no penalty, we instead use Sequential Least Squares Programming [21]. However, the intervention objective J_λ may be highly nonconcave. 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 covariate space [22, 23]. 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 by [24].

S3.1 Sparse Shift Intervention

To find the best k-sparse population shift intervention, we use the Sparse Shift Algorithm which relies on ℓ_1 relaxation. As the ℓ_1-norm provides the closest convex relaxation to the ℓ_0 norm, this is a a commonly adopted strategy to avoid combinatorial search in feature selection [25]. First, we compute the regularization path over different settings of the penalty $\lambda > 0$ for the following regularized objective (which is (8) with all $\gamma_s = 1.$):

$$J_\lambda(\Delta) := F^{-1}_{G_n(\Delta)}(\alpha) - \lambda||\Delta||_1 \tag{9}$$

which is maximized over the feasible set $\mathcal{C}_\Delta := \{\Delta \in \mathbb{R}^d : x + \Delta \in \mathcal{C}_x \text{ for all } x \in \mathbb{R}^d\}$
(recall we 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

Sparse Uniform Intervention algorithms is performed via the proximal gradient methods combined with our continuation approach discussed previously.

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

Input: Dataset $\mathcal{D}_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, Posterior $f \mid \mathcal{D}_n$
Parameters: $k \in \{1, \ldots, d\}$ specifies the maximal cardinality of the shift vector $\Delta \in \mathbb{R}^d$, $\mathcal{C}_\Delta \subseteq \mathbb{R}^d$ is the set of feasible shifts (ignoring the sparsity constraint), J_λ is our objective function in (8).

1: Set $\gamma_s = 1$ for $s = 1, \ldots, d$

2: Perform binary search (over λ) to find:

$$\lambda^* \leftarrow \operatorname{argmin} \left\{ \lambda \geqslant 0 \text{ s.t. } \Delta^* := \operatorname*{argmax}_{\Delta \in \mathcal{C}_\Delta} J_\lambda(\Delta) \text{ has } \leqslant k \text{ nonzero entries} \right\}$$

3: Define $\mathcal{I} \leftarrow \text{support}(\Delta^*_{\lambda*}) \subseteq \{1, \ldots, d\}$ where $\Delta^*_{\lambda*} := \operatorname*{argmax}_{\Delta \in \mathcal{C}_\Delta} J_{\lambda*}(\Delta)$

4: **Return:** $\Delta^* \in \mathbb{R}^d \leftarrow \operatorname{argmax}_{\Delta \in B} J_0(\Delta)$ where $B := \mathcal{C}_\Delta \bigcap \{\Delta \in \mathbb{R}^d : \Delta_i = 0 \text{ if } i \notin \mathcal{I}\}$

Recall that a (sparse) personalized intervention can be equivalently re-expressed as a (sparse) shift intervention. Thus, we identify the optimal sparse personalized intervention using the same objective in (8) with G_x in place of G_n, and can also apply the Sparse Shift Algorithm to find the best personalized intervention under a hard cardinality constraint. 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 $\mathcal{I} \subset \{1, \ldots, d\}$ 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).

S3.2 Sparse Uniform Intervention

Another goal is to identify the optimal uniform intervention which sets k of the covariates to particular fixed constants uniformly across all individuals from the population. We employ the forward stepwise selection algorithm described below, as the form of the optimization in this case is not amenable to ℓ_1-relaxation. Recall $\mathcal{I} \subseteq \{1, \ldots, d\}$ denotes the subset of covariates which are intervened upon, and the uniform intervention produces vector $T_{\mathcal{I}\to z}(x) \in \mathbb{R}^d$ such that $T_{\mathcal{I}\to z}(x)_s = x_s$ if $s \notin \mathcal{I}$, otherwise $T_{\mathcal{I}\to z}(x)_s = z_s$ 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 covariates in $\mathcal{I}$. For a given $\mathcal{I}$, the objective function to find the best constants is:

$$J^{\text{unif}}_{\mathcal{I}}\big(\{z_s\}_{s\in\mathcal{I}}\big) := F^{-1}_{G_n(T_{\mathcal{I}\to z})}(\alpha) \tag{10}$$

$$\text{with } G_n(T_{\mathcal{I}\to z}) = \frac{1}{n}\sum_{i=1}^n \big[f(z^{(i)}) - f(x^{(i)})\big] \mid \mathcal{D}_n \text{ where } z_s^{(i)} = \begin{cases} x^{(i)} & \text{if } s \notin \mathcal{I} \\ z_s & \text{otherwise} \end{cases}$$

which is maximized over the constraints: $z_s \in \mathcal{C}_s \subseteq \mathbb{R}$ for $s \in \mathcal{I}$.

Sparse Uniform Intervention Algorithm: Identifies best k-sparse uniform intervention.

Input: Dataset $\mathcal{D}_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, Posterior $f \mid \mathcal{D}_n$
Parameters: $k \in \{1, \dots, d\}$ specifies the maximal number of covariates which may be set by the uniform intervention, $\mathcal{C}_1, \dots, \mathcal{C}_d \subseteq \mathbb{R}$ are sets of feasible settings for each covariate.

1: Initialize $\mathcal{I} \leftarrow \varnothing,\ \mathcal{U} \leftarrow \{1, \dots, d\},\ J^* \leftarrow 0$
2: **While** $|\mathcal{I}| < k$:
3: Set $J_s^* \leftarrow \max_{\mathcal{C}_r : r \in \mathcal{I} \cup \{s\}} J^{\text{unif}}_{\mathcal{I} \cup \{s\}}\big(\{z_r\}_{r \in \mathcal{I} \cup \{s\}}\big)$ **for** each $s \in \mathcal{U}$
4: Find $s^* \leftarrow \text{argmax}_{s \in \mathcal{U}} \big\{J_s^*\big\}$
5: **If** $J_{s^*}^* > J^*$: $J^* \leftarrow J_{s^*}^*,\ \mathcal{I} \leftarrow \mathcal{I} \cup \{s^*\},\ \mathcal{U} \leftarrow \mathcal{U} \backslash s^*$
6: **Else:** **break**
7: **Return:** $\{z_s^*\}_{s \in \mathcal{I}} \leftarrow \text{argmax}_{\mathcal{C}_s : s \in \mathcal{I}}\ J^{\text{unif}}_{\mathcal{I}}\big(\{z_s\}_{s \in \mathcal{I}}\big)$

S4 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 [26], and we relegate proofs and technical definitions to §S9.

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 [26]). (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 $x, T(x) \in \mathcal{C}$: $\mathbb{E}_{\mathcal{D}_n}\Big|F^{-1}_{G_x(T)}(\alpha) - G_x^*(T)\Big| \leqslant \dfrac{C}{\alpha}\Big(L + \dfrac{1}{a}\Big) \cdot \Psi_{f^*}(n)^{1/[2(d+1)]}$

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

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

$\psi_{f^*}^{-1}(n)$ 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 [26] 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), $\psi_{f^*}^{-1}(n) = \mathcal{O}((1/\log n)^{\beta/2 - d/4})$ [26].

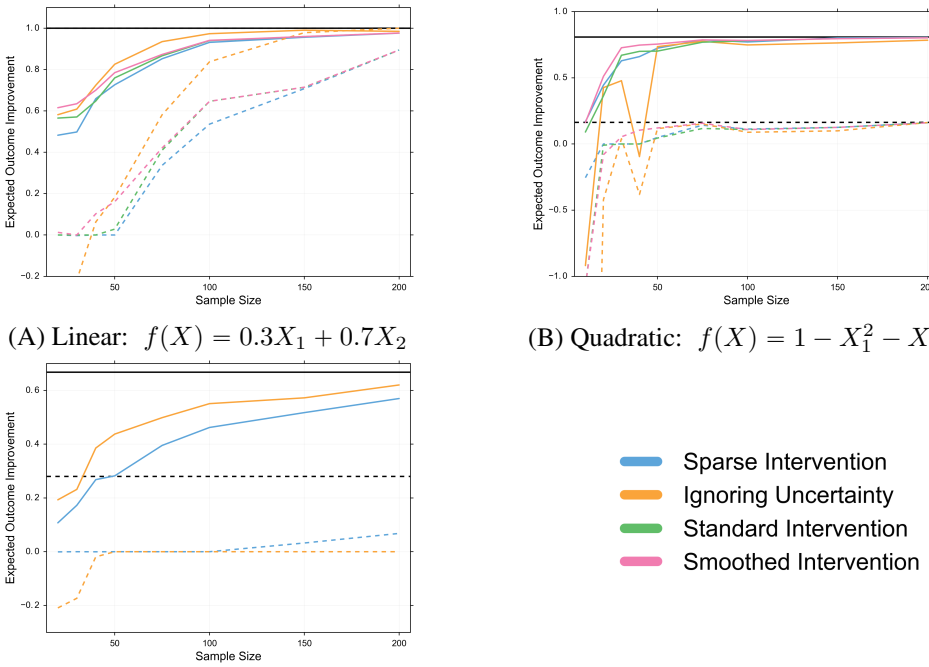
Theorem 2. *Under the assumptions of Theorem 1, for any T such that $\Pr(T(X) \in \mathcal{C}) = 1$:*

$$\mathbb{E}_{\mathcal{D}_n}\Big|F^{-1}_{G_n(T)}(\alpha) - G_X^*(T)\Big| \leqslant \frac{C}{\alpha}\Big[L\sqrt{\frac{d}{n}} + \Big(L + \frac{1}{a}\Big)\Psi_{f^*}(n)^{\frac{1}{2(d+1)}}\Big]$$

S5 Simulation Study

Over the simulated data summarized in Figure S2, 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 $Y = X_1 \cdot X_2 + \varepsilon$, the optimal (constrained) intervention may drastically vary depending upon the individual's covariate 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 S3 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.

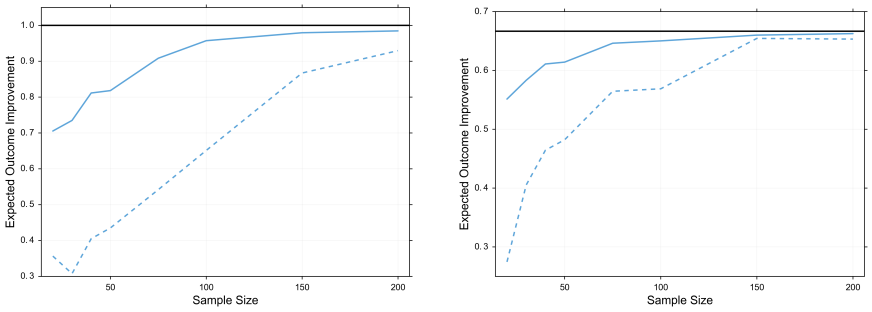


(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$

(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 0.05^{th} 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 covariates, 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 0.05^{th} quantile over our 100 trials).



(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$ (B) Quadratic: $Y = 1 - X_1^2 - X_2^2$

Figure S3: The mean (solid) and 0.05^{th} quantile (dashed) expected outcome change produced by our population intervention method, over 100 datasets for each sample size (same setting as in Figure §S2). The black line indicates the best possible expected outcome improvement.

S6 Population Intervention for Gene Perturbation

Next, we applied our method to search for population interventions in observational yeast gene expression data from [27]. 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 S4).

Insets (a) and (b) in Figure S4 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.

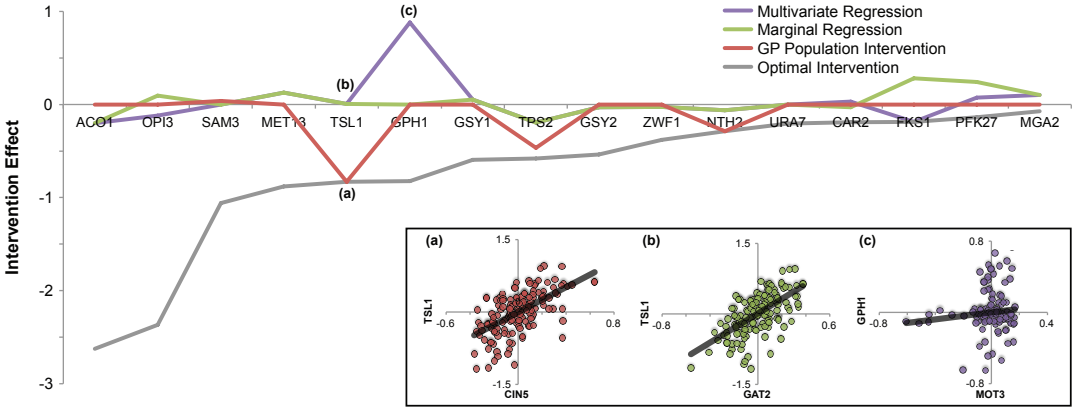


Figure S4: Actual effects of proposed interventions (single gene knockout) over a set transcription factors on down-regulation of each of a set of 16 small molecule metabolism target genes.

S6.1 Details

The data set used for this analysis contains gene expression levels for a set of wild type (ie. 'observational') samples, $\mathcal{D}_{obs}$ ($n = 161$), as well as for a set of 'interventional' samples, $\mathcal{D}_{int}$, 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 $\mathcal{D}_{obs}$ and determining which knockout produces the best value of our empirical uniform population intervention objective (for down-regulating the target). Subsequently, we use $\mathcal{D}_{int}$ to evaluate the actual effectiveness of proposed interventions in the knockout experiments. We only search for interventions present in $\mathcal{D}_{int}$ (single gene knockouts) rather than optimizing to infer optimal covariate transformations.

As candidate genes for this analysis we used only the 700 genes that [27] 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 covariate 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^{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 $\mathcal{D}_{int}$. The set of genes to down-regulate was simply chosen to be those classified by [27] 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 $\mathcal{D}_{int}$. 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 $\mathcal{D}_{int}$).

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 $\mathcal{D}_{obs}$. The best gene to knockout was inferred on the basis of the regression coefficients and expression values (if no beneficial regression coefficient was found statistically 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 S4 compares the results produced by these methods to the optimal intervention over X for down-regulating each Y, as found in the experimental data $\mathcal{D}_{int}$. 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 $\mathcal{D}_{int}$, 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 *SAM3*). On the other hand, neither form of linear regression proposed effective interventions for any target other than *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.

S7 Personalized Intervention for Writing Improvement

We demonstrate our personalized intervention methodology in a setting with rich nonlinear underlying relationships. Here, it is applied to the task of transforming a given news article into one which will be more widely-shared on social media. The observed data contain various covariates about individual Mashable articles along with their subsequent popularity in social networks [28]. 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 below). 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 S5.

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.

S7.1 Details

The data consist of 39,000 news articles published by Mashable around 2013-15 [28]. 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 covariates have been extracted from each article (eg. word count, the category such as "tech" or "lifestyle", keyword properties), many of which [28] produced using natural language processing algorithms (eg. subjectivity, polarity, alignment with topics found by Latent Dirichlet Allocation). After removing many highly redundant covariates, we center and rescale all variables to unit-variance (see Table S2 for a complete description of the 29 covariates 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-covariates. 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 covariate-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 covariates, and its uncertainty estimates tend to be larger in areas of the covariate-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 the 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 covariates except for the article category which presumably is fixed from an author's perspective. All covariate-transformations are constrained to lie within [-2,2] of the original (rescaled) covariate value, and we impose a sparsity constraint that at most 10 covariates 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 covariate-setting using *benchmark popularity*: the (weighted) average observed popularity amongst 100 nearest neighbors (in the covariate-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: $p \leqslant 2 \cdot 10^{-10}$). 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 covariates).

Figure S5 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 covariate-settings and the original covariate-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 ($p < 2 \cdot 10^{-6}$).

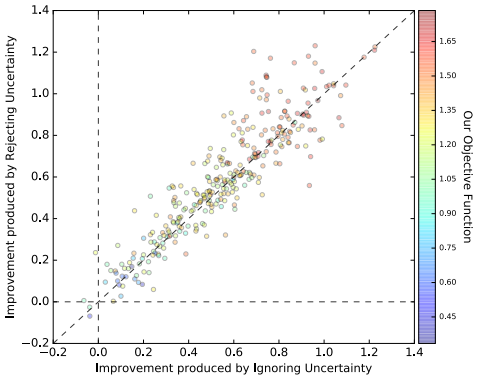


Figure S5: Benchmark popularity changes produced by the personalized interventions for 300 articles suggested by our method with $\alpha = 0.05$ (Rejecting Uncertainty) vs. $\alpha = 0.5$ (Ignoring Uncertainty). The points (ie. articles) are colored according to the value of our personalized intervention objective with $\alpha = 0.05$. Using $\alpha = 0.05$ outperforms $\alpha = 0.5$ in this analysis in 177/300 articles in the improvement set.

Method	Mean	Median	0.05^{th} 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).

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/`, our method proposes shifting the following 10 covariates (see Table S2 for covariate descriptions):

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_sharess: +2, global_subjectivity: +1.57, global_sentiment_polarity: -2

For this entertainment article: `http://mashable.com/2014/07/30/how-to-beat-the-heat/`, our method proposes shifting the following 10 covariates:

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 covariates by the ± 2 maximal allowed amounts, which leads to a 0.04 worse improvement in benchmark popularity compared with the covariate changes specified above for this article.

Covariate	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 covariates of each article (dimensions of X in this analysis). Covariates involving the share-counts of other articles and LDA were based only on data known before the publication date.

S8 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 $\mathcal{I} \subset \{1, \ldots, d\}$, our methods assume the post-treatment value of any non-intervened-upon covariate remains at its initial value (ie. $\widetilde{x}_s = x_s \; \forall s \notin \mathcal{I}$).

However, in many domains (such as our gene perturbation example when the profiled genes belong to a common regulatory network), the covariate-transformation produced by a sparse external intervention can only be roughly controlled. Let $T_{\mathcal{I} \to z}$ denote a uniform transformation which sets a subset of covariates in $\mathcal{I} \subset \{1, \ldots, d\}$ 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 [29]. Here, we suppose the underlying population of X, Y follows a *structural equation model* (SEM) [29]. 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 [30], 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 $\mathcal{I} \subseteq \{1, \ldots, d\}$, suppose the condition: (A7) $pa(Y) \subseteq \mathcal{I} \bigcup desc(\mathcal{I})^C$ holds. Then, for any uniform 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: $\mathcal{I}^* := \text{argmin}\left\{|\mathcal{I}'| \text{ s.t. } \exists\, T_{\mathcal{I}'\to z} \in \underset{T_{\mathcal{I}\to z}:|\mathcal{I}|\leqslant k}{\text{argmax}}\ \mathbb{E}_X\big[f^*(T_{\mathcal{I}\to z}(x)) - f^*(x)\big]\right\}$ 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 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).*

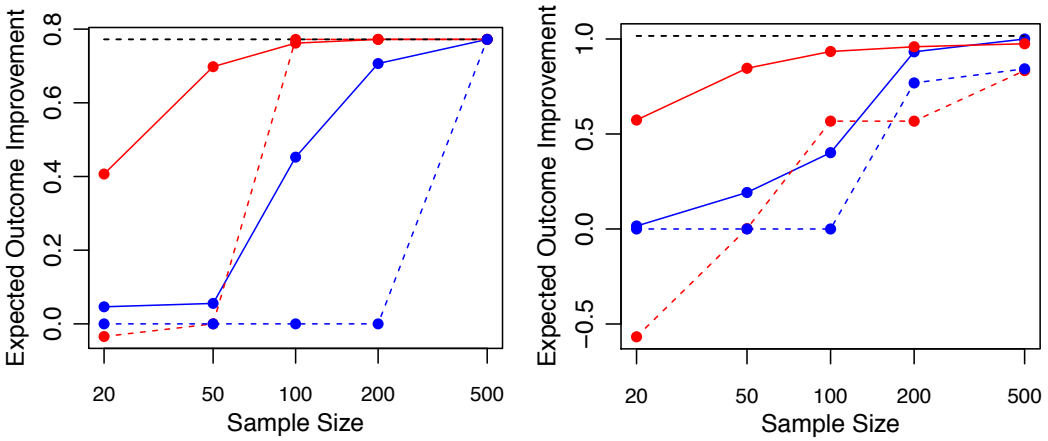
In the absence of extremely strong interactions between variables in $\text{pa}(Y)$, the equality of Theorem 3 will also hold for $\mathcal{I}^*$ if $|\text{pa}(Y)| \leqslant k$. For settings where sparse interventions elicit unintentional do-effects and the causal DAG meets condition (A8), Theorems 3 and 4 imply that, under complete certainty about f^*, the (minimum cardinality) maximizer of our uniform 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 uniform intervention objective (5) produces (in expectation as $n \to \infty$) beneficial interventions for populations whose underlying causal relationships satisfy certain conditions.

S8.1 Empirical Results

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. Our approach to identify a beneficial sparse population intervention is compared with inferring the complete SEM using the LinGAM estimator of [31] 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 S6A and S6B 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 [30]).

The assumption of sparse interventions realized as a do-operation (as defined by [30]) 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 covariates: 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 S6: 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 Shimizu et al. [31]. The black dashed line indicates the best possible improvement in each case.

S8.2 Additional Details

In our analysis, we suppose that a desired transformation upon variable $s \in \{1, \ldots, d\}$ cannot be enacted exactly and the Y which arises post-treatment is distributed according to $do(X_s = \mathbb{E}[X_s] + \Delta)$, where $\mathbb{E}[X_s]$ is the mean of the pre-treatment marginal distribution of the sth covariate. In this case, *do*-effects can propagate to other covariates which are descendants of s in the DAG because the values of descendant variables are redrawn from the *do*-distribution which arises as a result of shifting $\mathbb{E}[X_s]$. Because all relationships are linear in our SEMs, the actual expected outcome change resulting from a particular shift (resulting from the corresponding *do*-operation) is easily obtained in closed form.

Our GP framework 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 ± 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 [31] to estimate the SEM from the data, and then identify the optimal single-variable shift Δ_s^* 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 produced by the shift Δ_s^* proposed by each method (where the actual expected outcome change is found by analytically performing the $do(X_s = x_s + \Delta_s^*)$ operation in the true underlying SEM) .

In our experiment, two underlying SEM models are considered which were used by [31] to demonstrate the utility of their LinGAM method (albeit with impractically large sample size = 10,000). SEM$_A$ is used to refer to the model depicted in Figure 3 of [31], where we define Y as x6 (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 x7). The remainder of the variables in each SEM are adopted as our observed covariates 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 covariate incurs a multitude of unintentional off-target effects resulting from the *do*-effects propagating to downstream covariates in the SEM, whereas our method believes only the chosen covariate is changed. 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. The only disadvantage of the LinGAM method is that it does not know the direction of the causal relationship

$X \to Y$ (although we found it always estimated this direction correctly except on rare occasions with tiny sample sizes of $n = 20$).

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 covariate. Searching over these three values for each covariate 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 covariate-space. In practice, these potential difficulties do not pose much of an issue for our approach.

S9 Proofs

Notation and Definitions

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

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

All occurrences of f are implicitly referring to $f \mid \mathcal{D}_n$.

$\mu_n(\cdot)$, $\sigma_n^2(\cdot)$, and $\sigma_n(\cdot, \cdot)$ respectively denote the mean, variance, and covariance function of our posterior for $f \mid \mathcal{D}_n$ under the $\text{GP}\big(0, k(x, x')\big)$ prior.

$F_Z^{-1}(\alpha)$ denotes the α^{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 $\mathcal{H}_k$.

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

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

pa(Y) denotes the set of variables which are parents of Y in a causal *directed acyclic graph* (DAG) [29]

desc$(\mathcal{I})$ is the set of variables which are descendants of at least one variable in $\mathcal{I}$ 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:

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

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

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

A function f is *Lipschitz continuous* with constant L if: $|f(x) - f(x')| \leqslant L|x - x'|$ for every $x, x' \in \mathcal{C}$.

Suppose $\rho > 0$ is expressed as $\rho = m + \eta$ for nonnegative integer m and $0 < \eta \leqslant 1$.
The *Hölder space* $C^\rho[0, 1]^d$ is the space of functions with existing partial derivatives of orders $(k_1, \ldots, k_d)$ for all integers $k_1, \ldots, k_d \geqslant 0$ satisfying $k_1 + \cdots + k_d \leqslant m$. Additionally, each function's highest order partial derivative must form a function h that satisfies: $|h(x) - h(y)| \leqslant C|x - y|^\eta$ for any x, y.

Theorem 5 (van der Vaart & van Zanten [26])**.** *Under the assumptions of Theorem 1:*

$$\mathbb{E}_{\mathcal{D}_n} \int\int_{\mathcal{C}} [f(x) - f^*(x)]^2 p_X(x) \mathrm{d}x \; \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \;\leqslant\; C \cdot \Psi_{f^*}(n)$$

where $\Psi^{-1}_{f^*}(n)$ is defined as in §S4. See [26] for a detailed discussion about this function.

Proof of Theorem 1

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

$$\begin{aligned}
&\int_{\mathcal{C}} \big|f(x) - f^*(x)\big| p_X(x) \, \mathrm{d}x \\
\geqslant &\int_{\mathcal{B}_\delta(x_0)} \big|f(x) - f^*(x)\big| p_X(x) \, \mathrm{d}x + \int_{\mathcal{B}_\delta(T(x_0))} \big|f(x) - f^*(x)\big| p_X(x) \, \mathrm{d}x \\
\geqslant &a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \Big[\min_{x \in \mathcal{B}_\delta(x_0)} \big|f(x) - f^*(x)\big| + \min_{x \in \mathcal{B}_\delta(T(x_0))} \big|f(x) - f^*(x)\big| \Big] \\
\geqslant &a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \Big[\Big| f(T(x_0)) - f(x_0) - \big[f^*(T(x_0)) - f^*(x_0)\big] \Big| - 8\delta L \Big] \\
\geqslant &a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \Big[\Big| G_{x_0}(T) - G^*_{x_0}(T) \Big| - 8\delta L \Big] \qquad (11)
\end{aligned}$$

where $\mathrm{Vol}(\mathcal{B}_\delta) = \mathcal{O}(\delta^d)$. Theorem 5 implies the following inequality (ignoring constant factors):

$$\begin{aligned}
&[C \cdot \Psi_{f^*}(n)]^{1/2} \\
\geqslant &\left[\mathbb{E}_{\mathcal{D}_n} \int\int_{\mathcal{C}} [f(x) - f^*(x)]^2 p_X(x) \, \mathrm{d}x \, \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \right]^{1/2} \\
\geqslant &\mathbb{E}_{\mathcal{D}_n} \int\int_{\mathcal{C}} \big|f(x) - f^*(x)\big| p_X(x) \, \mathrm{d}x \, \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) && \text{by Jensen's inequality} \\
\geqslant &a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int \big|G_{x_0}(T) - G^*_{x_0}(T)\big| - \delta L \; \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) && \text{via the bound from (11)} \\
= &- aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^\infty \Pr\Big(\big|G_{x_0}(T) - G^*_{x_0}(T)\big| \geqslant r \Big) \, \mathrm{d}r \\
= &- aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^1 F^{-1}_{|G_{x_0}(T) - G^*_{x_0}(T)|}(\widetilde{\alpha}) \; \mathrm{d}\widetilde{\alpha} \\
\geqslant &- aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_\alpha^1 F^{-1}_{G_{x_0}(T)}(\widetilde{\alpha}) - G^*_{x_0}(T) \; \mathrm{d}\widetilde{\alpha} \\
\geqslant &- aL\delta^{d+1} + a(1-\alpha)\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \Big[F^{-1}_{G_{x_0}(T)}(\alpha) - G^*_{x_0}(T) \Big] \qquad (12)
\end{aligned}$$

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

$$\begin{aligned}
&- aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^1 F^{-1}_{|G^*_{x_0}(T) - G_{x_0}(T)|}(\widetilde{\alpha}) \, \mathrm{d}\widetilde{\alpha} \\
\geqslant &- aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^\alpha G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\widetilde{\alpha}) \; \mathrm{d}\widetilde{\alpha} \\
\geqslant &- aL\delta^{d+1} + a\alpha\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \Big[G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\alpha) \Big] \qquad (13)
\end{aligned}$$

Choosing $\delta := \big[\Psi_{f^*}(n)\big]^{\frac{1}{2(d+1)}}$ and combining (12) and (13) produces the desired result, since assuming $\alpha < 0.5$ implies $\alpha < 1 - \alpha$. □

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 $\alpha < 0.5$. □

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

$$\mathbb{E}_{\mathcal{D}_n}\left|F^{-1}_{G_n(T)}(\alpha) - F^{-1}_{G_X(T)}(\alpha)\right| \leqslant C \cdot \Big[\frac{-L^2 d}{n}\log(1-\alpha)\Big]^{1/2}$$

Proof of Lemma 1. Define random variables $Z_i := f(T(x^{(i)}) - f(x^{(i)}) \mid \mathcal{D}_n$ for $i = 1, \ldots, n$. Note that these variables all share the same expectation: $\mathbb{E}_X[Z] := \mathbb{E}_X[Z_i] = G_X(T)$ and $G_n(T) = \frac{1}{n}\sum_{i=1}^n Z_i$. The Lipschitz continuity of f combined with the fact that $\mathcal{C} = [0,1]^d$ implies: $Z_i \in [-L\sqrt{d}, L\sqrt{d}]$ for all i. Thus, Hoeffding's inequality ensures:

$$\Pr\left(\left|G_n(T) - G_X(T)\right| \geqslant t\right) \leqslant 2\exp\left(\frac{-nt^2}{2L^2 d}\right)$$
$$\Rightarrow F^{-1}_{\left|G_n(T) - G_X(T)\right|}(\alpha) \leqslant C \cdot \Big[\frac{-L^2 d}{n}\log(1-\alpha)\Big]^{1/2}$$

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

$$F^{-1}_{G_n(T)}(\alpha) - F^{-1}_{G_X(T)}(\alpha) \leqslant F^{-1}_{\left|G_n(T) - G_X(T)\right|}(\alpha)$$
$$\text{and } F^{-1}_{G_X(T)}(\alpha) - F^{-1}_{G_n(T)}(\alpha) \leqslant F^{-1}_{\left|G_n(T) - G_X(T)\right|}(\alpha)$$

□

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

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

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

$$\begin{aligned}
&\int_{\mathcal{C}} |f(x) - f^*(x)| p_X(x)\,\mathrm{d}x \\
\geqslant& a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \left[\int_{\mathcal{C}} |f(x) - f^*(x)| p_X(x)\,\mathrm{d}x + \int_{\mathcal{C}} |f(T(x)) - f^*(T(x))| p_X(x)\,\mathrm{d}x - 8\delta L\right] \\
\geqslant& a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \left[\Big|\mathbb{E}_X[f(x) - f^*(x)] + \mathbb{E}_X[f(T(x)) - f^*(T(x))]\Big| - 8\delta L\right]
\end{aligned}$$

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

$$-aL\delta^{d+1} + a\alpha\delta^d \cdot \mathbb{E}_{\mathcal{D}_n}\Big[G^*_X(T) - F^{-1}_{G_X(T)}(\alpha)\Big] \leqslant [C \cdot \Psi_{f*}(n)]^{1/2}$$

and we can use the same argument to similarly bound

$$\mathbb{E}_{\mathcal{D}_n}\Big[F^{-1}_{G_X(T)}(\alpha) - G^*_X(T)\Big]$$

□

Proof of Theorem 3

Here, we employ subscripts to index particular covariates 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.

$$\mathbb{E}_{\text{do}(X_{\mathcal{I}}=z_{\mathcal{I}})}\big[f^*(x)\big]$$

$$= \int f^*\big([x_{\mathcal{I}^C}, z_{\mathcal{I}}]\big)\, p\big(x_{\mathcal{I}^C} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\big)\, \mathrm{d}x_{\mathcal{I}^C}$$

$$= \int\int f^*\big([x_{\text{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\text{pa}(Y)}, a_{\mathcal{I}^C\setminus\text{pa}(Y)}]\big) \cdot p\big(x_{\mathcal{I}^C\setminus\text{pa}(Y)} \mid x_{\text{pa}(Y)\setminus\mathcal{I}}, do(X_{\mathcal{I}} = z_{\mathcal{I}})\big)$$
$$\cdot\, p\big(x_{\text{pa}(Y)\setminus\mathcal{I}} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\big)\, \mathrm{d}x_{\mathcal{I}^C\setminus\text{pa}(Y)}\, \mathrm{d}x_{\text{pa}(Y)\setminus\mathcal{I}}$$

where covariate-subset $a_{\mathcal{I}^C\setminus\text{pa}(Y)}$ can take arbitrary values since f^* is constant along covariates $\notin \text{pa}(Y)$

$$= \int f^*\big([x_{\text{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\text{pa}(Y)}, a_{\mathcal{I}^C\setminus\text{pa}(Y)}]\big)\, p\big(x_{\text{pa}(Y)\setminus\mathcal{I}} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\big)\, \mathrm{d}x_{\text{pa}(Y)\setminus\mathcal{I}}$$

$$= \int f^*\big([x_{\text{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\text{pa}(Y)}, a_{\mathcal{I}^C\setminus\text{pa}(Y)}]\big)\, p\big(x_{\text{pa}(Y)\setminus\mathcal{I}}\big)\, \mathrm{d}x_{\text{pa}(Y)\setminus\mathcal{I}}$$

since the marginal distribution over $X_{\text{pa}(Y)\setminus\mathcal{I}}$ equals the *do*-distribution by assumption (A7)

$$= \int\int f^*\big([x_{\text{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\text{pa}(Y)}, x_{\mathcal{I}^C\setminus\text{pa}(Y)}]\big)\, p\big(x_{\mathcal{I}^C\setminus\text{pa}(Y)} \mid x_{\text{pa}(Y)\setminus\mathcal{I}}\big)\, p\big(x_{\text{pa}(Y)\setminus\mathcal{I}}\big)\, \mathrm{d}x_{\mathcal{I}^C\setminus\text{pa}(Y)}\, \mathrm{d}x_{\text{pa}(Y)\setminus\mathcal{I}}$$

$$= \mathbb{E}_X\Big[f^*(T_{\mathcal{I}\to z}(x))\Big]$$

□

Proof of Theorem 4

Recall we defined:

$$\mathcal{I}^* := \text{argmin}\left\{|\mathcal{I}'| \text{ s.t. } \exists\, T_{\mathcal{I}'\to z} \in \underset{T_{\mathcal{I}\to z}:|\mathcal{I}|\leqslant k}{\text{argmax}}\ \mathbb{E}_X\big[f^*(T_{\mathcal{I}\to z}(x)) - f^*(x)\big]\right\} \tag{14}$$

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 $\mathbb{E}_X[f^*(T_{\mathcal{I}\to z}(x))]$ does not change when $z_j := [T_{\mathcal{I}\to z}(x)]_j$ 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 (A7). 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 $\text{desc}(\mathcal{I}^*)$, since this is a subset of $\text{desc}\big(\text{pa}(Y)\big)$. □

Additional References for the Supplementary Information

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

[17] Paciorek CJ, Schervish MJ (2004) Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems (NIPS)* 17.

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

[19] McHutchon A, Rasmussen CE (2011) Gaussian process training with input noise. *Advances in Neural Information Processing Systems (NIPS)* 24.

[20] Le QV, Smola AJ, Canu S (2005) Heteroscedastic Gaussian process regression. *22nd International Conference on Machine Learning (ICML)* .

[21] Kraft D (1988) A software package for sequential quadratic programming. DLR German Aerospace Center - Institute for Flight Mechanics, Koln, Germany.

[22] 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.

[23] Lizotte DJ (2008) Practical Bayesian Optimization. Ph.D. thesis, University of Alberta.

[24] Mobahi H, L ZC, Ma Y (2012) Seeing through the blur. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .

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

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

[27] 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.

[28] 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* .

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

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

[31] 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.