Supplementary Material

Contents

S1 Gaussian Process Regression

Gaussian Process regression (Rasmussen 2006) 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 GP model, but can be combined with any algorithm that learns a reasonable posterior for f. While employing a more powerful model should improve the results of our approach, comparing various regressors is not our focus. Thus, all practical results of our methodology are presented using only the standard GP regression model, under which the posterior distribution over f is given by the above expressions. In each application presented here, our GP uses the Automatic-Relevance-Determination (ARD) covariance function, a popular choice for multi-dimensional data (Rasmussen 2006):

$$k(x, x') = \sigma_0^2 \cdot \exp\left[-\frac{1}{2} \sum_{s=1}^{d} \left(\frac{x_s - x'_s}{l_s} \right)^2 \right] \tag{12}$$

The ARD kernel relies on length-scale hyperparameters $l_1, \ldots, l_d$ which determine how much f can depend on each dimension of the feature-space. All hyperparameters of our GP regression model (covariance-kernel parameters $l_1 \ldots, l_d$ and σ_0 (the output variance) as well as the variance of the noise σ^2) are empirically selected via marginal-likelihood maximization (Rasmussen 2006). In each application, we employ the 0.05^{th} posterior-quantile ($\alpha = 0.05$) in our method to ensure that with high probability, the intervention it infers to be optimal induces a nonnegative change in expected outcomes.

S2 Algorithmic Details

To find an optimal transformation of our regularized objective J_λ in (13), we employ the proximal gradient method described in §4. When $\lambda = 0$ and there is no penalty, we instead use Sequential Least Squares Programming (Kraft 1988). 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 feature space (Shahriari et al. 2016, Lizotte 2008). 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 Mobahi et al. (2012).

S2.1 Sparse Shift Intervention

Here, we provide an explanatory description of the Sparse Shift Algorithm from §4. To find the best k-sparse population shift intervention, we resort to 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 (Bach et al. 2012). First, we compute the regularization path over different settings of the penalty $\lambda > 0$ for the following regularized objective:

$$J_\lambda(\Delta) := F^{-1}_{G_n(\Delta)}(\alpha) - \lambda ||\Delta||_1 \tag{13}$$

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 covariates which receive nonzero shift. Finally, we optimize the original unregularized objective ($\lambda = 0$) with respect to only the selected covariates in $\mathcal{I}$ to remove bias induced by the regularizer. Each inner maximization in both the Sparse Shift/Covariate-fixing algorithms is performed via the proximal gradient methods combined with our continuation approach introduced in §S2.

S2.2 Sparse Covariate-fixing Intervention

Another goal is to identify the optimal covariate-fixing intervention which sets k of the covariates to particular fixed constants uniformly across all individuals from the population. We employ the forward step-wise selection algorithm described below, as the form of the optimization in this case is not amenable to ℓ_1-relaxation. Recall

$\mathcal{I} \subseteq \{1, \ldots, d\}$ denotes the subset of covariates which are intervened upon, and the covariate-fixing 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:

$$J_{\mathcal{I}}^{\text{unif}}\big(\{z_s\}_{s \in \mathcal{I}}\big) := F^{-1}_{G_n(T_{\mathcal{I} \to z})}(\alpha) \tag{14}$$

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

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

Sparse Covariate-fixing Algorithm: Identifies best k-sparse covariate-fixing intervention.

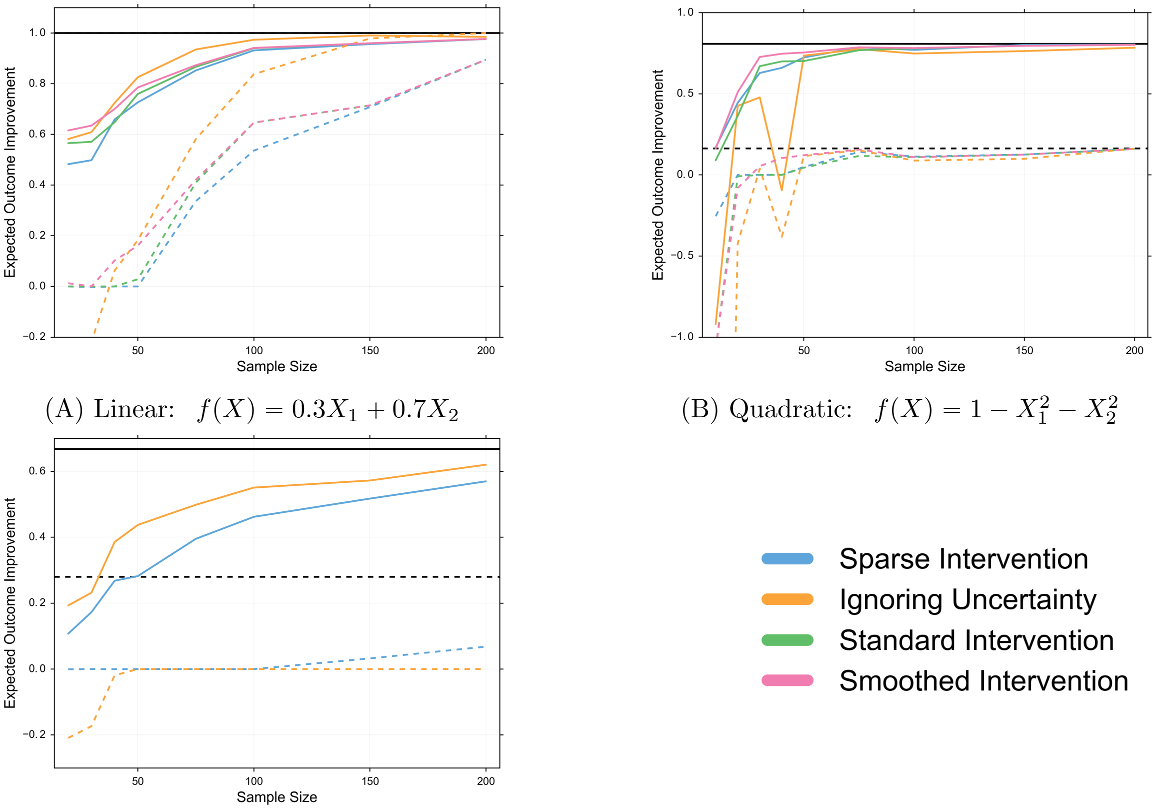
Input: Dataset $\mathcal{D}_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, Posterior $f \mid \mathcal{D}_n$
Parameters: $k \in \{1, \ldots, d\}$ specifies the maximal number of covariates which may be set by the covariate-fixing intervention, $\mathcal{C}_1, \ldots, \mathcal{C}_d \subseteq \mathbb{R}$ are sets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

```
Initialize $\mathcal{I} \leftarrow \varnothing$, $\mathcal{U} \leftarrow \{1, \ldots, d\}$, $J^* \leftarrow 0$
While $|\mathcal{I}| < k$:
    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 $s \in \mathcal{U}$
    Find $s^* \leftarrow \text{argmax}_{s \in \mathcal{U}} \{J_s^*\}$
    If $J^*_{s^*} > J^*$:    $J^* \leftarrow J^*_{s^*}$, $\mathcal{I} \leftarrow \mathcal{I} \cup \{s^*\}$, $\mathcal{U} \leftarrow \mathcal{U} \backslash s^*$
    Else:    break
Return: $\{z_s^*\}_{s \in \mathcal{I}} \leftarrow \text{argmax}_{\mathcal{C}_s : s \in \mathcal{I}} J^{\text{unif}}_{\mathcal{I}}\big(\{z_s\}_{s \in \mathcal{I}}\big)$
```

S3 Simulations

Over the simulated data summarized in Figure S1, we apply our basic personalized intervention method ($\alpha = 0.05$) with purely local optimization (standard) and our continuation technique (smoothed), which significantly improves results. For each of the 100 datasets, we randomly sampled a new point (from the same underlying distribution) to receive a personalized intervention. The magnitude of each intervention is bounded by 1, except for in data from the quadratic relationship. We also infer sparse interventions (with a cardinality constraint of 2 for the linear and quadratic relationships, 1 for the product relationship). When $Y = X_1 \cdot X_2 + \varepsilon$, 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 S2 displays the behavior of both the population shift intervention in the linear setting, and the population covariate-fixing intervention under the quadratic relationship. The population intervention is notably safer than the individually tailored variants, producing no negative changes in our experiments.

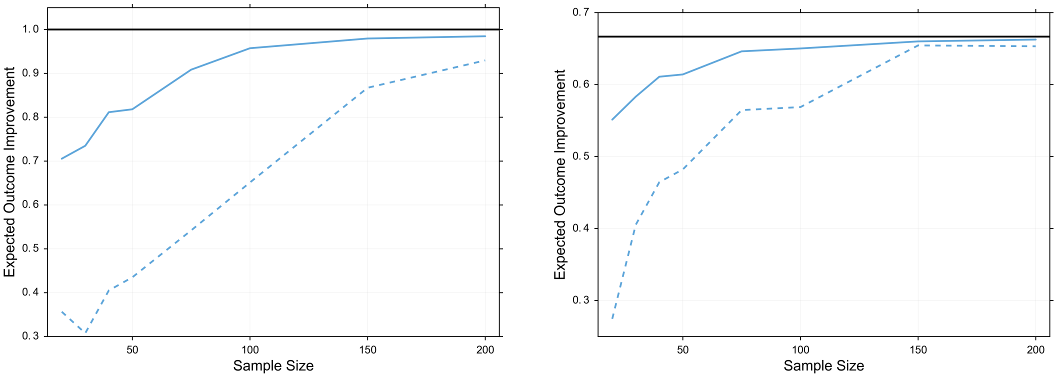


(A) Linear: $f(X) = 0.3X_1 + 0.7X_2$

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

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

Figure S1: The mean (solid) and 0.05th quantile (dashed) expected outcome change produced under personalized interventions suggested by various methods, over 100 datasets of each sample size. Each dataset contains 10-dimensional 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 0.05th 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 S2: The mean (solid) and 0.05th quantile (dashed) expected outcome change produced by our population intervention method, over 100 datasets for each sample size (same setting as in Figure §S1). The black line indicates the best possible expected outcome improvement.

S3.1 Linear SEM Analysis

Here, 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 (Shimizu et al. 2006) 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 Shimizu et al. (2006) 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 Shimizu et al. (2006), 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 feature-space. In practice, these potential difficulties do not pose much of an issue for our approach.

S4 Gene Knockout Interventions

The data set used for this analysis contains gene expression levels for a set of wild type (ie. ‘observational’) samples, $\mathcal{D}_{obs}$ ($n = 161$), 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 covariate-fixing 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 Kemmeren et al. (2014) classified as responsive mutants (at least four transcripts show robust changes in response to the knockout). Furthermore, we omitted genes whose expression over the 161 observational samples had standard deviation < 0.1. Out of the transcription factors present in the remaining set of genes, we defined the top 10 factors as our feature set X, after ranking the transcription factors by the difference between their expression when they were knocked out in the interventional data and their 0.1^{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 Kemmeren et al. (2014) 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 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 2 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.

S5 Interventions to Improve Article Popularity

We demonstrate our personalized intervention methodology in a setting with rich nonlinear underlying relationships. The data consist of 39,000 news articles published by Mashable around 2013-15 (Fernandes et al. 2015). Each article is annotated with the number of shares it received in social networks (which we use as our outcome variable after log-transform and rescaling). A multitude of features have been extracted from each article (eg. word count, the category such as "tech" or "lifestyle", keyword properties), many of which Fernandes et al. (2015) 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 did not measurably benefit from ARD feature-weighting, we simply use the squared exponential kernel. Over the held-out articles, the Pearson correlation between the observed popularity and the GP (posterior mean) predictions is 0.35. Furthermore, there is a highly significant ($p < 8 \cdot 10^{-41}$) positive correlation of 0.07 between the model's predictive variance and the actual squared errors of GP predictions over this held-out set. Our model is thus able to make reasonable predictions of popularity based on the available covariates, and its uncertainty estimates tend to be larger in areas of the feature-space where the posterior mean lies further from actual popularity values.

In this analysis, we compare our personalized intervention methodology which *rejects* uncertainty (using $\alpha = 0.05$) with a variant of 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 feature-space) from the set of held-out articles (with weights based on inverse Euclidean distance). Over our improvement set, the Pearson correlation between articles' observed popularity and benchmark popularity is 0.28 (highly significant: $p \leqslant 2 \cdot 10^{-10}$). 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 S3 depicts the results of our personalized intervention for each article in our intervention set. The expected improvement produced by a particular intervention is estimated as the difference between the benchmark popularity of the post-intervention 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 ($p < 2 \cdot 10^{-6}$).

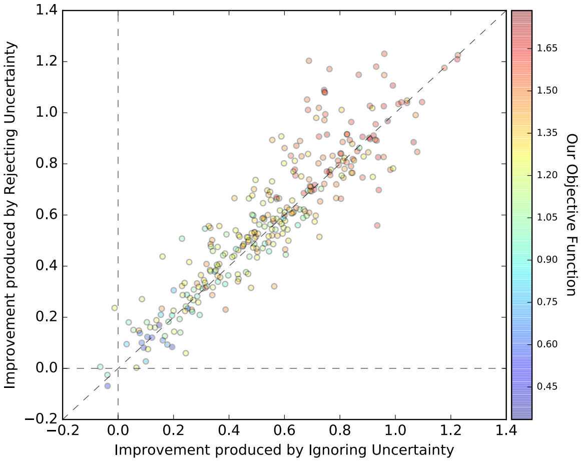


Figure S3: Benchmark popularity changes produced by the personalized interventions for 300 articles suggested by our method with $\alpha = 0.05$ (Rejecting Uncertainty) vs. $\alpha = 0.5$ (Ignoring Uncertainty). The points (ie. articles) are colored according to the value of our personalized intervention 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 feature 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.

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

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

S6 Proofs and additional Theoretical Results

Notation and Definitions

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

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 $\mathrm{GP}\big(0, k(x,x')\big)$ 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 $x \in \mathcal{C}$.

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

$\mathrm{pa}(Y)$ denotes the set of variables which are parents of Y in a causal *directed acyclic graph* (DAG) (Pearl 2000)

$\mathrm{desc}(\mathcal{I})$ 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:

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

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

A function f is *Lipshitz 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 $(k_1, \dots, k_d)$ for all integers $k_1, \dots, k_d \geqslant 0$ satisfying $k_1 + \dots + k_d \leqslant m$. 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 (2011))**.** *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 $\Psi_{f*}^{-1}(n)$ is defined as in §5. See van der Vaart & van Zanten (2011) 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:

$$\begin{aligned}
&\int_{\mathcal{C}} \big|f(x) - f^*(x)\big| p_X(x) \ \mathrm{d}x \\
\geqslant &\int_{\mathcal{B}_\delta(x_0)} \big|f(x) - f^*(x)\big| p_X(x) \ \mathrm{d}x + \int_{\mathcal{B}_\delta(T(x_0))} \big|f(x) - f^*(x)\big| p_X(x) \ \mathrm{d}x \\
\geqslant & a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \Big[\min_{x \in \mathcal{B}_\delta(x_0)} \big|f(x) - f^*(x)\big| + \min_{x \in \mathcal{B}_\delta(T(x_0))} \big|f(x) - f^*(x)\big| \Big] \\
\geqslant & a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \Big[\Big| f(T(x_0)) - f(x_0) - \big[f^*(T(x_0)) - f^*(x_0) \big] \Big| - 8\delta L \Big] \\
\geqslant & a \cdot \mathrm{Vol}(\mathcal{B}_\delta) \cdot \Big[\Big| G_{x_0}(T) - G^*_{x_0}(T) \Big| - 8\delta L \Big]
\end{aligned} \tag{15}$$

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

$$\begin{aligned}
&[C \cdot \Psi_{f*}(n)]^{1/2} \\
\geqslant & \left[\mathbb{E}_{\mathcal{D}_n} \int \int_{\mathcal{C}} [f(x) - f^*(x)]^2 p_X(x) \ \mathrm{d}x \ \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) \right]^{1/2} \\
\geqslant & \mathbb{E}_{\mathcal{D}_n} \int \int_{\mathcal{C}} \big|f(x) - f^*(x)\big| p_X(x) \ \mathrm{d}x \ \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) && \text{by Jensen's inequality} \\
\geqslant & a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int \big| G_{x_0}(T) - G^*_{x_0}(T) \big| - \delta L \ \ \mathrm{d}\Pi_n(f \mid \mathcal{D}_n) && \text{via the bound from (15)} \\
= & - aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^\infty \Pr\Big(\big| G_{x_0}(T) - G^*_{x_0}(T) \big| \geqslant r \Big) \ \mathrm{d}r \\
= & - aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^1 F^{-1}_{|G_{x_0}(T) - G^*_{x_0}(T)|}(\widetilde{\alpha}) \ \ \mathrm{d}\widetilde{\alpha} \\
\geqslant & - aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_\alpha^1 F^{-1}_{G_{x_0}(T)}(\widetilde{\alpha}) - G^*_{x_0}(T) \ \ \mathrm{d}\widetilde{\alpha} \\
\geqslant & - aL\delta^{d+1} + a(1-\alpha)\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \Big[F^{-1}_{G_{x_0}(T)}(\alpha) - G^*_{x_0}(T) \Big]
\end{aligned} \tag{16}$$

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

$$\begin{aligned}
& - aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^1 F^{-1}_{|G^*_{x_0}(T) - G_{x_0}(T)|}(\widetilde{\alpha}) \ \mathrm{d}\widetilde{\alpha} \\
\geqslant & - aL\delta^{d+1} + a\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \int_0^\alpha G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\widetilde{\alpha}) \ \ \mathrm{d}\widetilde{\alpha} \\
\geqslant & - aL\delta^{d+1} + a\alpha\delta^d \cdot \mathbb{E}_{\mathcal{D}_n} \Big[G^*_{x_0}(T) - F^{-1}_{G_{x_0}(T)}(\alpha) \Big]
\end{aligned} \tag{17}$$

Choosing $\delta := [\Psi_{f*}(n)]^{\frac{1}{2(d+1)}}$ and combining (16) and (17) 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 $x, T(x) \in \mathcal{C}$:*

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

Proof of Lemma 1. Define random variables $Z_i := f(T(x^{(i)}) - f(x^{(i)}) \mid \mathcal{D}_n$ for $i = 1, \dots, n$. 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:

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

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

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

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

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

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

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

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

and we can use the same argument to similarly bound

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

□

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 $R, S \subseteq \{1, \dots, d\}$. 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.

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{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_{\mathrm{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\mathrm{pa}(Y)}, a_{\mathcal{I}^C\setminus\mathrm{pa}(Y)}]\big)\cdot p\big(x_{\mathcal{I}^C\setminus\mathrm{pa}(Y)} \mid x_{\mathrm{pa}(Y)\setminus\mathcal{I}}, do(X_{\mathcal{I}} = z_{\mathcal{I}})\big) \\
&\qquad\qquad \cdot p\big(x_{\mathrm{pa}(Y)\setminus\mathcal{I}} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\big)\ \mathrm{d}x_{\mathcal{I}^C\setminus\mathrm{pa}(Y)}\ \mathrm{d}x_{\mathrm{pa}(Y)\setminus\mathcal{I}} \\
&\text{where covariate-subset } a_{\mathcal{I}^C\setminus\mathrm{pa}(Y)} \text{ can take arbitrary values since } f^* \text{ is constant along covariates} \notin \mathrm{pa}(Y) \\
=&\int f^*\big([x_{\mathrm{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\mathrm{pa}(Y)}, a_{\mathcal{I}^C\setminus\mathrm{pa}(Y)}]\big)\ p\big(x_{\mathrm{pa}(Y)\setminus\mathcal{I}} \mid do(X_{\mathcal{I}} = z_{\mathcal{I}})\big)\ \mathrm{d}x_{\mathrm{pa}(Y)\setminus\mathcal{I}} \\
=&\int f^*\big([x_{\mathrm{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\mathrm{pa}(Y)}, a_{\mathcal{I}^C\setminus\mathrm{pa}(Y)}]\big)\ p\big(x_{\mathrm{pa}(Y)\setminus\mathcal{I}}\big)\ \mathrm{d}x_{\mathrm{pa}(Y)\setminus\mathcal{I}} \\
&\text{since the marginal distribution over } X_{\mathrm{pa}(Y)\setminus\mathcal{I}} \text{ equals the } do\text{-distribution by assumption (A7)} \\
=&\int\int f^*\big([x_{\mathrm{pa}(Y)\setminus\mathcal{I}}, z_{\mathcal{I}\cap\mathrm{pa}(Y)}, x_{\mathcal{I}^C\setminus\mathrm{pa}(Y)}]\big)\ p\big(x_{\mathcal{I}^C\setminus\mathrm{pa}(Y)} \mid x_{\mathrm{pa}(Y)\setminus\mathcal{I}}\big)\ p\big(x_{\mathrm{pa}(Y)\setminus\mathcal{I}}\big)\ \mathrm{d}x_{\mathcal{I}^C\setminus\mathrm{pa}(Y)}\ \mathrm{d}x_{\mathrm{pa}(Y)\setminus\mathcal{I}} \\
=&\mathbb{E}_X\Big[f^*(T_{\mathcal{I}\to z}(x))\Big]
\end{aligned}
$$

□

Proof of Theorem 4

Recall we defined:

$$
\mathcal{I}^* := \operatorname{argmin}\Big\{|\mathcal{I}'| \ \text{ s.t. } \ \exists\ T_{\mathcal{I}'\to z} \in \underset{T_{\mathcal{I}\to z}:|\mathcal{I}|\leqslant k}{\operatorname{argmax}}\ \mathbb{E}_X\big[f^*(T_{\mathcal{I}\to z}(x)) - f^*(x)\big]\Big\} \tag{18}
$$

as the intervention set corresponding to the optimal sparse covariate-fixing 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 $j \notin \mathrm{pa}(Y)$, including variables outside of the parent set in $\mathcal{I}$ does not improve this quantity. Thus, either $\mathrm{pa}(Y) \subseteq \mathcal{I}^*$, or $\mathcal{I}^* \subset \mathrm{pa}(Y)$. The first case immediately implies (A7). When $\mathcal{I}^* \subset \mathrm{pa}(Y)$: our assumption that no variable in $\mathrm{pa}(Y)$ is a descendant of other parents implies the other parents must belong the complement of $\mathrm{desc}(\mathcal{I}^*)$, since this is a subset of $\mathrm{desc}\big(\mathrm{pa}(Y)\big)$. □

Theorem 6 and Proof

Theorem 6. *Suppose we adopt a $GP\big(0, k(x,x')\big)$ prior and, in addition to the assumptions outlined in §5, the following conditions hold: (A9) $f^* \in \mathcal{H}_k(\mathcal{C})$ which is the RKHS induced by our covariance function k with norm $||\cdot||_k$ (cf. Rasmussen (2006) §6.1), (A10) noise variables $\varepsilon^{(i)}$ form a uniformly bounded martingale difference sequence $\varepsilon^{(i)} \leqslant \sigma$ for $i = 1,\ldots,n$.*

Then, for any $x, T(x) \in \mathcal{C}$: $\quad F^{-1}_{G_x(T)}(\alpha) \leqslant G^*_x(T)$

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

In Theorem 6, $\gamma_n := \max\limits_{A\subset\mathcal{C}:|A|=n} \dfrac{1}{2}\log\big|\mathbf{I} + \sigma^{-2}\mathbf{K}_A\big|$ is a kernel-dependent quantity ($\mathbf{K}_A := [k(x,x')]_{x,x'\in A}$) which, in the Gaussian setting, is the mutual information between f and observations of Y at the most informative choice of n points. When the kernel satisfies $k(x,x') \leqslant 1$, the following bounds are known (Srinivas et al.

2010): $\gamma_n = \mathcal{O}(d \log n)$ for the linear kernel, $\gamma_n = \mathcal{O}((\log n)^{d+1})$ for the squared exponential kernel, and $\gamma_n = \mathcal{O}(n^{d(d+1)/(2\nu+d(d+1))}(\log n))$ for the Matérn kernel with smoothness parameter ν.

Note that while f^* is not required to be drawn from our prior and ε may be non-Gaussian, this result assumes the kernel k and noise-level σ are correctly set. Our proof relies on the following statement:

Theorem 7 (Srinivas et al. (2010))**.** *Assume conditions (A9) - (A10), fix $\delta \in (0,1)$, and define:*

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

Then:

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

Proof of Theorem 6. Fix $x, T(x) \in \mathcal{C}$, and define $\delta := (n+1) \cdot \exp\left(-\frac{[\Phi^{-1}(\alpha)]^2 - 2||f^*||_k^2}{300\gamma_n}\right)$.
In this case, $-\sqrt{\beta_{n+1}} = \Phi^{-1}(\alpha)$ (see definition in previous theorem).

Theorem 7 implies that with probability $\geqslant 1 - \delta$:
$|\mu_n(x) - f^*(x)| \leqslant -\Phi^{-1}(\alpha) \cdot \sigma_n(x)$ and $|\mu_n(T(x)) - f^*(T(x))| \leqslant -\Phi^{-1}(\alpha) \cdot \sigma_n(T(x))$

Since our posterior is Gaussian:

$$F^{-1}_{G_x(T)}(\alpha) = \mu_n(T(x)) - \mu_n(x) + \Phi^{-1}(\alpha)\Big[\sigma_n^2(T(x)) + \sigma_n^2(x) - 2\sigma_n(x, T(x))\Big]^{1/2}$$

Therefore:

$$\begin{aligned}
&f^*(T(x)) - f^*(x) - F^{-1}_{G_x(T)}(\alpha) \\
=&f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\Big[\sigma_n^2(T(x)) + \sigma_n^2(x) - 2\sigma_n(x, T(x))\Big]^{1/2} \\
\leqslant&f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\Big[\sigma_n^2(T(x)) + \sigma_n^2(x) + 2\sqrt{\sigma_n^2(x)\sigma_n^2(T(x))}\Big]^{1/2} \\
&\qquad \text{since we assume } \alpha \leqslant 0.5 \Rightarrow \Phi^{-1}(\alpha) \leqslant 0, \text{ and } \big|\sigma_n(x, T(x))\big| \leqslant \sqrt{\sigma_n^2(x)\sigma_n^2(T(x))} \\
=&f^*(T(x)) - \mu_n(T(x)) + \mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\Big[\sigma_n(T(x)) + \sigma_n(x)\Big] \\
=&\big[f^*(T(x)) - \mu_n(T(x)) - \Phi^{-1}(\alpha)\sigma_n(T(x))\big] + \big[\mu_n(x) - f^*(x) - \Phi^{-1}(\alpha)\sigma_n(x)\big]
\end{aligned}$$

which is less than 0 with probability at most δ. □

Additional References for the Supplementary Material

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

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* .

Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, 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.

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

Lizotte, D. J. (2008), Practical Bayesian Optimization, PhD thesis, University of Alberta.

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

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

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

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.