

# Causal Estimation with Functional Confounders

Aahlad Puli<sup>1</sup>      Adler J. Perotte<sup>2</sup>      Rajesh Ranganath<sup>1,3</sup>  
aahlad@nyu.edu    adler.perotte@columbia.edu    rajeshr@cims.nyu.edu

<sup>1</sup>Computer Science, New York University, New York, NY 10011

<sup>2</sup>Biomedical Informatics, Columbia University, New York, NY 10032

<sup>3</sup>Center for Data Science, New York University, New York, NY 10011

## Abstract

Causal inference relies on two fundamental assumptions: *ignorability* and *positivity*. We study causal inference when the true confounder value can be expressed as a function of the observed data; we call this setting *estimation with functional confounders* (EFC). In this setting ignorability is satisfied, however positivity is violated, and causal inference is impossible in general. We consider two scenarios where causal effects are estimable. First, we discuss interventions on a part of the treatment called *functional interventions* and a sufficient condition for effect estimation of these interventions called *functional positivity*. Second, we develop conditions for nonparametric effect estimation based on the gradient fields of the functional confounder and the true outcome function. To estimate effects under these conditions, we develop Level-set Orthogonal Descent Estimation (LODE). Further, we prove error bounds on LODE's effect estimates, evaluate our methods on simulated and real data, and empirically demonstrate the value of EFC.

## 1 Introduction

Determining the effect of interventions on outcomes using observational data lies at the core of many fields like medicine, economic policy, and genomics. For example, policy makers estimate effects to elect whether to invest in education or job training programs. In medicine, doctors use effects to design optimal treatment strategies for patients. Geneticists perform genome-wide association studies (GWAS) to relate genotypes and phenotypes. In observational data, there could exist unobserved variables that affect both the intervention and the outcome, called confounders. A necessary condition for the causal effect to be identified is that all confounders are observed; called *ignorability*. If ignorability holds, a sufficient condition for causal effect estimation is adequate variation in the intervention after conditioning on the confounders; this condition is called *positivity*.

The data apriori does not differentiate between confounders and interventions. It is the practitioners that select interventions of interest from all pre-outcome variables (variables that occur before the outcome). Then, assuming knowledge of the data generating mechanism, practitioners can label certain variables amongst the remaining pre-outcome variables as confounders. This corresponds to indexing into the set of pre-outcome variables.

In certain problems the confounders are specified as a function of the pre-outcome variables that does not simply index into the set of pre-outcome variables. For a concrete example, consider GWAS. The goal in GWAS is to estimate the influence of genetic variations on phenotypes like disease risk. In GWAS, population and family structures both result in certain genetic variations and affect phenotypes and therefore, are confounders [4]. Practitioners specify these confounders by using the genetic similarity between individuals [15, 19, 31], which is a function of the genetic variations. When the confounders are a function of the same pre-outcome variables that define the interventions, positivity is violated. Then, the class of interventions whose effects are estimable is not well-defined.

We study causal effect estimation in such settings, where a function of the pre-outcome variables provides the confounder and these same pre-outcome variables define the intervention. We call this estimation with functional confounders (EFC). In EFC, one column in the observed data is the outcome and all others are pre-outcome variables. We assume access to a function  $h(\cdot)$  that takes as input the pre-outcome variables and returns the value of the confounder. Further, we assume these confounders give us ignorability. In settings like GWAS, the function  $h$  reflects the practitioner-specified function that captures the genetic variation influenced by the population structure. In traditional observational causal inference (OBS-CI),  $h(\cdot)$  reflects the selection of certain variables in the data and labelling them as confounders. In EFC, two different values of the confounder are never observed for the same setting of the pre-outcome variables. This means that positivity is violated and the effects of only certain interventions may be estimable.

We address this issue in two ways. First, we investigate a class of plausible interventions that are *functions* of the observed pre-outcome variables, called functional interventions. We develop a sufficient condition to estimate the effects of said functional interventions, called functional positivity (F-POSITIVITY). Second, we consider intervening on all pre-outcome variables, called the *full* intervention. We develop a sufficient condition to estimate the effect of the *full* intervention, called causal redundancy (C-REDUNDANCY). For an intervention, given a confounder value, C-REDUNDANCY allows us to compute a surrogate intervention such that the conditional effect of the surrogate is equal to that of the original intervention. We also show that such surrogate interventions exist only under a certain condition that we call Effect Connectivity, that is necessary for nonparametric effect estimation in EFC. This condition is satisfied by default in traditional OBS-CI if ignorability and positivity hold. Then, we develop an algorithm for causal estimation assuming C-REDUNDANCY, called Level-set Orthogonal Descent Estimation (LODE), which estimates effects using surrogate interventions. If the surrogate is not estimated well, LODE's estimates are biased. We establish bounds on this bias that capture the mitigating effect of the smoothness of the true outcome function.

**Related work** The problem of genome-wide association studies (GWAS) is to estimate the effect of genetic variations(also called single nucleotide polymorphisms (SNPs)) on the phenotype [29]. The ancestry of the subjects acts as a confounder in GWAS. In GWAS practice, principle component analysis (PCA) and linear mixed models (LMMS) are used to compute this confounding structure [19, 31]. Lippert et al. [15] suggest estimating the confounders and effects on *separate* subsets of the SNPs. This separation disregards the confounding that is captured in the interaction of the two subsets of SNPs. GWAS is a special case of effects from multiple treatments (MTE) where the confounder value is specified via optimization as a function of the pre-outcome variables [20, 30]. In all these settings, positivity is violated and not all effects are estimable. We provide an avenue for nonparametric effect-estimation of the full intervention under a new condition, C-REDUNDANCY.

**Traditional observational causal inference (OBS-CI) review** We setup causal inference with Structural Causal Models [17] and use  $\text{do}(\mathbf{t} = \mathbf{t}^*)$  to denote making an intervention. Let  $\mathbf{t}$  be a vector of the interventions,  $\mathbf{z}$  be the confounder, and  $\mathbf{y}$  be the outcome. Let  $\boldsymbol{\eta} \sim p(\boldsymbol{\eta})$  ( $\boldsymbol{\eta} \perp\!\!\!\perp (\mathbf{z}, \mathbf{t})$ ) be noise. With  $f$  as the *outcome function*, we define the causal model for traditional OBS-CI as<sup>1</sup>:

$$\mathbf{z} \sim p(\mathbf{z}), \quad \mathbf{t} \sim p(\mathbf{t}|\mathbf{z}), \quad \mathbf{y} = f(\mathbf{t}, \mathbf{z}, \boldsymbol{\eta}).$$

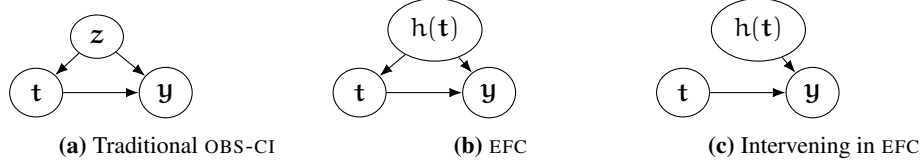
Let  $p(\mathbf{y}, \mathbf{z}, \mathbf{t})$  denote the joint distribution implied by this data generating process. The effects of interest under the full intervention  $\text{do}(\mathbf{t} = \mathbf{t}^*)$  are the *average and conditional effect*

$$(\text{average}) \quad \tau(\mathbf{t}^*) = \mathbb{E}_{\mathbf{z}, \boldsymbol{\eta}} f(\mathbf{t}^*, \mathbf{z}, \boldsymbol{\eta}) \quad (\text{conditional}) \quad \phi(\mathbf{t}^*, \mathbf{z}) = \mathbb{E}_{\boldsymbol{\eta}} [f(\mathbf{t}^*, \mathbf{z}, \boldsymbol{\eta})]. \quad (1)$$

With observed confounders, two assumptions make causal estimation possible: *ignorability* and *positivity*. Ignorability means that *all* confounders  $\mathbf{z}$  are observed in data. Conditioning on all the confounders, the outcome under an intervention is distributed as if conditional on the value of the intervention:  $p(\mathbf{y} = y_1 | \text{do}(\mathbf{t} = \mathbf{t}^*), \mathbf{z} = \mathbf{z}) = p(f(\mathbf{t}^*, \mathbf{z}, \boldsymbol{\eta}) = y_1) = p(\mathbf{y} = y_1 | \mathbf{t} = \mathbf{t}^*, \mathbf{z} = \mathbf{z})$ . This allows the expression of average effect as an expectation over the *observed* outcomes  $\tau(\mathbf{t}^*) = \mathbb{E}_{\mathbf{z}, \boldsymbol{\eta}} [f(\mathbf{t}^*, \mathbf{z}, \boldsymbol{\eta})] = \mathbb{E}_{\mathbf{z}} \mathbb{E}_{\boldsymbol{\eta}} [y | \mathbf{z}, \mathbf{t}^*]$ . The conditional expectation only exists for all  $\mathbf{t}^*$  if  $p(\mathbf{y} | \mathbf{z}, \mathbf{t} = \mathbf{t}^*) = p(\mathbf{y}, \mathbf{z}, \mathbf{t} = \mathbf{t}^*) / p(\mathbf{z})p(\mathbf{t} = \mathbf{t}^* | \mathbf{z})$  exists. *Positivity* guarantees this existence

$$(\text{positivity}) \quad \forall \mathbf{t}^* \in \text{supp}(\mathbf{t}) \quad p(\mathbf{z} = \mathbf{z}) > 0 \implies p(\mathbf{t} = \mathbf{t}^* | \mathbf{z} = \mathbf{z}) > 0. \quad (2)$$

<sup>1</sup>We focus on  $f$  that generates  $\mathbf{y}$  from  $\mathbf{t}, \mathbf{z}$ . SCMs generally specify the function that generates  $\mathbf{t}$  from  $\mathbf{z}$  also.



**Figure 1:** Causal Graphs for Traditional OBS-CI vs. EFC.

## 2 Estimation with functional confounders

In traditional OBS-CI, causal estimation relied on knowing the confounders. In this section, we consider settings where confounders are known via a function of the pre-outcome variables  $h(t) = z$ . We call this setting *estimation with functional confounders* (EFC). An example of this is GWAS, where SNPs (the pre-outcome variables) are used to estimate the confounding population structure through methods like PCA [31]. Assuming the confounders are a function of the pre-outcome variables violates positivity in general. Positivity is violated in this setting because

$$\forall t_1, t_2 \in \text{supp}(t) \text{ s.t. } h(t_2) \neq h(t_1) \implies p(z = h(t_2) | t = t_1) = 0 \neq p(z = h(t_2)) > 0$$

In words, two different confounder values cannot occur for the same  $t$ . **A positivity violation precludes nonparametric effect estimation of the full intervention  $\text{do}(t = t^*)$ .**

**Positivity and Regression Identifiability** Positivity can be viewed as providing identifiability. To see this, let the confounder be  $z = h(t)$  and the outcome be  $y(t, z, \eta) = z + h(t)$ . Now consider regressing  $z$  and  $t$  onto  $y$ . Then, functions  $y = \alpha z + \beta h(t)$  indexed by  $\alpha, \beta$ , such that  $\alpha + \beta = 2$ , are consistent with the observed data. Thus, there exist infinitely many solutions to the conditional expectation of  $y$  on  $(t, z)$ , meaning that the regression is not identifiable. Assuming positivity necessitates sufficient randomness to identify the regression and thus the causal effect. **A violation of positivity means that nonparametric estimation of causal effects needs further assumptions.**

### 2.1 Setup for EFC

In EFC, the confounder is provided as a non-bijective function  $h$  of the pre-outcome variables  $t$ . To reflect this property, we use  $h(t)$  to denote the confounder. As an illustrative example, let  $\mathcal{G}$  be the Gamma distribution and consider  $z \in \{-1, 1\}$ ,  $p(z = 1) = 0.5$  is the confounder and the intervention of interest is  $t = z * \mathcal{G}(1, \exp(z))$ . Note  $\text{sign}(t) = z$  meaning that  $h(t) = \text{sign}(t)$  is the confounder. Figure 1 shows causal graphs connecting our EFC notation to that in traditional OBS-CI. With noise  $\eta \sim p(\eta) (\eta \perp\!\!\!\perp t)$ , our causal model samples, in order, the confounder "part" of pre-outcome variables  $h(t)$ , the pre-outcome variables  $t$ , and the outcome  $y$  via the outcome function  $f$ <sup>2</sup>:

$$h(t) \sim p(h(t)) \quad t \sim p(t | h(t)) \quad y = f(t, h(t), \eta)$$

Similar to traditional OBS-CI, for an intervention  $t^*$  the average effect,  $\tau(\cdot)$ , and the conditional effect,  $\phi(\cdot, \cdot)$  at  $h(t_2^*)$ , respectively, are defined as:

$$\tau(t^*) = \mathbb{E}_{h(t), \eta} [f(t^*, h(t), \eta)], \quad \phi(t^*, h(t_2^*)) = \mathbb{E}_{\eta} [f(t^*, h(t_2^*), \eta)]. \quad (3)$$

**As the pre-outcome variables determine the confounder, positivity is violated. Further, the outcome function  $f(t, h(t), \eta)$  could recover the exact value of  $h(t)$  from  $t$  instead of its second argument. Thus, two different outcome functions could lead to the same observational data distribution, posing a fundamental obstacle to causal effect estimation. This is the central challenge in EFC.**

### 2.2 Causal Questions With Functional Positivity

**Without positivity, we can only estimate the effects of certain functions of  $t$ .** We call such interventions, on some function  $g(t)$ , *functional interventions*. The implied causal model for the outcome for functional intervention value  $g(t^*)$  and confounder value  $h(t_2^*)$  is first  $t \sim p(t | g(t) = g(t^*), h(t) = h(t_2^*))$  and then  $y = f(t, h(t_2^*), \eta)$ <sup>3</sup>. Then, the functional average effect is

$$(\text{average}) \quad \tau(g(t^*)) = \mathbb{E}_{h(t), \eta} \mathbb{E}_{t | g(t)=g(t^*), h(t)} [f(t, h(t), \eta)].$$

**An example of a functional intervention is intervening on the cumulative dosage of a drug. In contrast, traditional interventions would set each individual dose given at different points in time.**

<sup>2</sup>We also assume no interference [10] (also called Stable Unit Treatment Value Assumption [24]) which means that an individual's outcome does not depend on others' treatment. In EFC, when  $t$  and  $\eta$  are sampled IID there is no interference. To see this, note  $\forall i, j \ (t_i, \eta_i) \perp\!\!\!\perp (t_j, \eta_j) \implies (y_i, t_i) \perp\!\!\!\perp (y_j, t_j) \implies y_i \perp\!\!\!\perp t_j$ .

<sup>3</sup>Intervening on  $g(t)$  can be interpreted as making a *soft intervention* [9, 7] of  $t$  to  $p(t | z, g(t) = g(t))$ .

**F-POSITIVITY and Functional Effect Estimation** For the causal model above to be well-defined for all functional interventions  $g(t^*)$ , the conditional  $p(t | g(t) = g(t^*), h(t) = h(t_2^*))$  must exist. To guarantee this existence, we define **functional positivity (F-POSITIVITY)** for any  $g(t^*)$

$$(F-POSITIVITY) \quad p(h(t) = h(t_2^*)) > 0 \implies p(g(t) = g(t^*) | h(t) = h(t_2^*)) > 0. \quad (4)$$

**F-POSITIVITY** says that the function of the pre-outcome variables that is being intervened on needs to have sufficient randomness when the function of the pre-outcome variables that defines the confounders is fixed. Further, under F-POSITIVITY, effect estimation for functional interventions is reduced to traditional OBS-CI on data  $p(y, g(t), h(t))$ . With positivity and ignorability satisfied, traditional causal estimators such as propensity scores [23], matching [21], regression [11], and doubly robust methods [22] can be used to estimate the causal effect. Focusing on regression, let  $f_\theta$  be a flexible function, then  $\min_\theta \mathbb{E}_{y,t}[(y - f_\theta(h(t), g(t)))^2]$  would estimate the conditional expectation of interest:  $\mathbb{E}[y | h(t), g(t^*)]$ . With  $\theta$ , the effect of  $g(t^*)$  can be estimated by averaging the estimate of the conditional expectation over the marginal distribution  $p(h(t))$ :

$$\tau(g(t^*)) = \mathbb{E}_t[f_\theta(h(t), g(t^*))]. \quad (5)$$

### 3 Identification of effects of the full intervention

When positivity is violated, causal effects cannot be estimated as conditional expectations over the observed data in general. We give a functional condition, called **causal redundancy (C-REDUNDANCY)**, that allows us to estimate the effect of the full intervention  $do(t = t^*)$ , even when positivity is violated. Specifically, C-REDUNDANCY allows us to construct a *surrogate intervention*  $t'(t^*, h(t_2^*))$  whose conditional effect at  $h(t')$  matches the conditional effect of interest,  $\phi(t^*, h(t_2^*))$ . Let  $\tilde{t}$  be a fixed value of the full intervention, then C-REDUNDANCY is

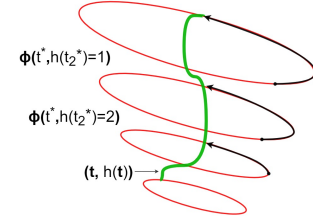
**Assumption.** Recall the outcome  $y = f(\tilde{t}, h(\tilde{t}), \eta)$ . With  $\nabla_{\tilde{t}}$  as gradient w.r.t. to argument  $\tilde{t}$ :

$$\forall \tilde{t}, h(\tilde{t}_2), \eta, \quad \nabla_{\tilde{t}} f(\tilde{t}, h(\tilde{t}_2), \eta)^\top \nabla_{\tilde{t}} h(\tilde{t}) = 0.$$

In words, C-REDUNDANCY is the condition that the outcome function  $f$  uses the value of the confounder from its second argument instead of computing  $h(t)$  from the first argument<sup>4</sup>. To compute the conditional effect  $\phi(t^*, h(t_2^*))$ , we develop Level-set Orthogonal Descent Estimation (LODE). LODE's key step is to construct a surrogate intervention  $t'(t^*, h(t_2^*))$  such that

$$\phi(t^*, h(t_2^*)) = \phi(t'(t^*, h(t_2^*)), h(t_2^*)), \quad h(t_2^*) = h(t'(t^*, h(t_2^*))).$$

By definition, a surrogate intervention lives in the conditional effect level-set:  $\{\tilde{t} : \phi(\tilde{t}, h(t_2^*)) = \phi(t^*, h(t_2^*))\}$ . So LODE searches this level-set for  $t'(t^*, h(t_2^*))$ . See fig. 2 which plots the conditional effect level-sets with the value of  $h(t)$  fixed (red) in  $(\text{supp}(t), \text{supp}(h(t)))$ -space. Green corresponds to the observed data,  $\text{supp}(t, h(t))$ . LODE finds  $t'(t^*, h(t_2^*))$  by traversing the level-sets (black) to account for the confounder part mismatch  $h(t^*) \neq h(t_2^*)$ . C-REDUNDANCY ensures LODE can traverse these level-sets as it implies  $\nabla_{\tilde{t}} \phi(\tilde{t}, h(\tilde{t}_2)) \nabla_{\tilde{t}} h(\tilde{t}) = 0$  under the regularity conditions in theorem 1. Thus, under C-REDUNDANCY, surrogate interventions can be constructed by solving a gradient flow equation which guarantees identification as follows:



**Figure 2:** LODE's traversal.

**Theorem 1.** Assume C-REDUNDANCY holds. Assuming the following:

1. Let  $t'(t^*, h(t_2^*))$  be the limiting solution to the gradient flow equation  $\frac{d\tilde{t}(s)}{ds} = -\nabla_{\tilde{t}}(h(\tilde{t}(s)) - h(t_2^*))^2$ , initialized at  $\tilde{t}(0) = t^*$ ; i.e.  $t'(t^*, h(t_2^*)) = \lim_{s \rightarrow \infty} \tilde{t}(s)$ . Further, let  $h(t'(t^*, h(t_2^*))) = h(t_2^*)$  and  $t'(t^*, h(t_2^*)) \in \text{supp}(t)$ .
2.  $f(\tilde{t}, h(\tilde{t}), \eta)$  and  $h(\tilde{t})$  as functions of  $\tilde{t}, h(\tilde{t})$  are continuous and differentiable and the derivatives exist for all  $\tilde{t}, \eta$ . Let  $\nabla_{\tilde{t}} f(\tilde{t}, h(\tilde{t}), \eta)$  exist and be bounded and integrable w.r.t. the probability measure corresponding to  $p(\eta)$ , for all values of  $\tilde{t}$  and  $h(\tilde{t})$ .

<sup>4</sup>If  $f$  transforms its first argument  $\tilde{t}$  into  $h(\tilde{t})$  as one amongst many different computations, the chain rule implies  $\nabla_{\tilde{t}} f(\tilde{t}, h(t_2^*))^\top \nabla_{\tilde{t}} h(\tilde{t})$  has a term  $\|\nabla_{\tilde{t}} h(\tilde{t})\|^2$  which is non-zero in general.

Then the conditional effect (and therefore the average effect) is identified:

$$\phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) = \phi(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)), h(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)))) = \mathbb{E}[\mathbf{y} | \mathbf{t} = \mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))] \quad (6)$$

In words, the key idea is that starting at  $\tilde{\mathbf{t}}(0) = \mathbf{t}^*$  and following  $\nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}})$  means  $\tilde{\mathbf{t}}(s)$  always lies in the level-set  $\{\tilde{\mathbf{t}} : \phi(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*)) = \phi(\mathbf{t}^*, h(\mathbf{t}_2^*))\}$ . See [appendix A.2](#) for the proof. While C-REDUNDANCY is stated in terms of the gradient of the outcome function, it suffices for [theorem 1](#) to assume a weaker condition about the gradient of the conditional effect:  $\nabla_{\tilde{\mathbf{t}}} \mathbb{E}_{\boldsymbol{\eta}} f(\tilde{\mathbf{t}}, \tilde{\mathbf{t}}_2, \boldsymbol{\eta})^\top \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}) = 0$ .

**Surrogate Positivity** In [theorem 1](#), we assumed that the surrogate  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)) \in \text{supp}(\mathbf{t})$ . This condition, which we call surrogate positivity (analogous to positivity), states that for any intervention and confounder, surrogate interventions that are limiting solutions to the gradient flow equation have nonzero density conditional on the confounder value. Formally, for any intervention  $\mathbf{t} = \mathbf{t}^*$

$$p(h(\mathbf{t}) = h(\mathbf{t}_2^*)) > 0 \implies p(\mathbf{t} = \mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)) | h(\mathbf{t}) = h(\mathbf{t}_2^*)) > 0, \quad (7)$$

and  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  satisfies assumption 1 in [theorem 1](#). Surrogate positivity along with C-REDUNDANCY, is sufficient for full effect estimation under EFC. Next, we show that the positivity assumption in traditional causal inference is a special case of surrogate positivity.

**Traditional observational causal inference (OBS-CI) and LODE** Let the confounder and intervention of interest in traditional OBS-CI be  $\mathbf{z}$  and  $\mathbf{a}$  respectively. Assume both are scalars and ignorability and positivity hold. This setup can be embedded in EFC by defining the vector of pre-outcome variables as:  $\mathbf{t} = [\mathbf{a}; \mathbf{z}]$ . In this setting, C-REDUNDANCY and surrogate positivity([eq. \(7\)](#)) hold by default. Let the outcome be  $\mathbf{y} = f(\mathbf{t}, h(\mathbf{t})) = f(\mathbf{a}, \mathbf{z})$ , where  $f$  only depends on the first element of  $\mathbf{t}$ , i.e. [a](#)<sup>5</sup>. Let  $\mathbf{e}_1 = [1, 0]$  and  $\mathbf{e}_2 = [0, 1]$ . In traditional OBS-CI as EFC,  $\nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*)) \propto \mathbf{e}_1$  and  $\nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}) \propto \mathbf{e}_2$  meaning that  $\nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*))^\top \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}) = 0$ . Thus, C-REDUNDANCY holds by default. Moreover, under positivity of  $\mathbf{a}$  w.r.t.  $\mathbf{z}$ , we also have surrogate positivity for traditional OBS-CI as an EFC problem. In this setting, LODE computes  $\mathbf{t}' = [\mathbf{a}^*, h(\mathbf{t}_2^*)]$  by following  $-\nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}) = [0, -1]$ , which only changes the value of  $h(\tilde{\mathbf{t}}_2)$ , not the value of  $\mathbf{a}$ . Thus,  $\mathbf{t}^*$  and  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  will have the same first element and  $\mathbf{t}'$ 's second element will be  $h(\mathbf{t}_2^*)$ . As  $\mathbf{a}$  has positivity w.r.t.  $\mathbf{z}$ , we have  $p(\mathbf{a} = \mathbf{a}^*, \mathbf{z} = h(\mathbf{t}_2^*)) > 0$  which means  $\mathbf{t}' \in \text{supp}(\mathbf{t})$ . The estimated conditional effect is  $\mathbb{E}[\mathbf{y} | \mathbf{t} = \mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))] = f([\mathbf{a}^*, \mathbf{z}^*], h(\mathbf{t}_2^*)) = \mathbb{E}[\mathbf{y} | \mathbf{a} = \mathbf{a}^*, \mathbf{z} = h(\mathbf{t}_2^*)]$ , which matches the estimate in traditional OBS-CI.

**Implementation of LODE** LODE first estimates the conditional expectation  $\mathbb{E}[\mathbf{y} | \mathbf{t}]$ ; this can be done with model-based or nonparametric estimators. This is achieved by regressing  $\mathbf{y}$  on  $\mathbf{t}$ ,  $\hat{f} = \arg \min_{\mathbf{u} \in \mathcal{F}} \mathbb{E}_{\mathbf{y}, \mathbf{t} \sim D} (\mathbf{y} - \mathbf{u}(\mathbf{t}))^2$ , with empirical distribution  $D$ . The surrogate intervention  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  is computed using Euler integration to solve the gradient flow equation. Euler integration in this setting is equivalent to gradient descent with a fixed step size. Other, more efficient schemes like Runge–Kutta numerical integration methods [[3](#)] could also be used. The conditional effect estimate is  $\hat{f}(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)))$ . See [algorithm 1](#) for a description.

### 3.1 Estimation error of LODE in practice

To compute the surrogate intervention  $\mathbf{t}'$ , LODE uses the gradients of  $h(\cdot)$  in Euler integration. In practice, taking Euler integration steps, instead of solving the gradient flow exactly, could result in errors. Then  $\mathbf{t}'$  could lie outside the level-set of the conditional effect  $\phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) = \mathbb{E}_{\boldsymbol{\eta}}[f(\mathbf{t}^*, h(\mathbf{t}_2^*), \boldsymbol{\eta})]$ . Further, if  $h(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))) \neq h(\mathbf{t}_2^*)$ , LODE incurs error for conditioning on a value of the confounder that is different from  $h(\mathbf{t}_2^*)$ . The error due to  $\mathbf{t}'$  estimation is decoupled from the error in the estimation of  $\mathbb{E}[\mathbf{y} | \mathbf{t}]$  which adds without further amplification. We formalize this error:

**Theorem 2.** Consider the conditional effect  $\phi(\mathbf{t}^*, h(\mathbf{t}_2^*))$ . Let  $\hat{\mathbf{t}}(\mathbf{t}^*, h(\mathbf{t}_2^*))$  be the estimate of the surrogate intervention computed by LODE, computed via Euler integration of the gradient flow  $\frac{d\tilde{\mathbf{t}}(s)}{ds} = -\nabla_{\tilde{\mathbf{t}}} (h(\tilde{\mathbf{t}}(s)) - h(\mathbf{t}_2^*))^2$ , initialized at  $\tilde{\mathbf{t}}(0) = \mathbf{t}^*$ . Assume the true surrogate  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  exists and is the limiting solution to the gradient flow equation.

1. Let the finite sample estimator of  $\mathbb{E}[\mathbf{y} | \mathbf{t} = \tilde{\mathbf{t}}]$  be  $\hat{f}(\tilde{\mathbf{t}})$ . Let the error for all  $\tilde{\mathbf{t}}$  be bounded,  $|\hat{f}(\tilde{\mathbf{t}}) - \mathbb{E}[\mathbf{y} | \mathbf{t} = \tilde{\mathbf{t}}]| \leq c(N)$ , where  $N$  is the sample size and  $\lim_{N \rightarrow \infty} c(N) = 0$ .
2. Assume  $K$  Euler integrator steps were taken to find the surrogate estimate  $\hat{\mathbf{t}}(\mathbf{t}^*, h(\mathbf{t}_2^*))$ , each of size  $\ell$ . Let the maximum confounder mismatch be  $\max_{i \leq K} (h(\tilde{\mathbf{t}}_i) - h(\mathbf{t}_2^*))^2 = M$ .

<sup>5</sup>We ignore noise in the outcome for ease of exposition.



3. Let  $L_{z,\tilde{t}}$  be the Lipschitz-constant of  $\phi(\tilde{t}, h(\tilde{t}_2))$  as a function of  $h(\tilde{t}_2)$ , for fixed  $\tilde{t}$ .  
 Let  $L_e$  be the Lipschitz-constant of  $\mathbb{E}[\mathbf{y} | \mathbf{t} = \tilde{t}] = \phi(\tilde{t}, h(\tilde{t}))$  as a function of  $\tilde{t}$ .  
 Assume  $h$  has a gradient with bounded norm,  $\|\nabla h(\tilde{t})\|_2 \leq L_h$ .  
 Assume  $f$ 's Hessian has bounded eigenvalues:  $\forall \tilde{t}, \tilde{t}_2, \|\nabla_{\tilde{t}}^2 \phi(\tilde{t}, h(\tilde{t}_2))\|_2 \leq \sigma_{H\phi}$ .

The conditional effect estimate error,  $\xi(\mathbf{t}^*, h(\mathbf{t}_2^*)) = |\hat{f}(\hat{\mathbf{t}}) - \phi(\mathbf{t}^*, h(\mathbf{t}_2^*))|$ , is upper bounded by:

$$c(N) + \min (L_e \|\mathbf{t}' - \hat{\mathbf{t}}\|_2, 2K\ell^2 (\mathcal{O}(\ell) + M\sigma_{H\phi}L_h^2) + L_{z,\hat{\mathbf{t}}} \|h(\hat{\mathbf{t}}) - h(\mathbf{t}_2^*)\|_2) \quad (8)$$

See [appendix A.3](#) for the proof. [Theorem 2](#) captures the trade-off between biases due to conditioning on the wrong confounder value and due to the accumulated error in solving the gradient flow equation. This accumulated error analysis may be loose in settings where the sum of many gradient steps lead to  $\hat{\mathbf{t}} \approx \mathbf{t}'$ , even if each step individually induces large error. In such settings, the term that depends on  $\|\hat{\mathbf{t}} - \mathbf{t}'\|_2$  is a better measure of error. The maximum-mismatch  $M$  appears because Euler integrator takes steps that depend on the magnitude of the gradient which depends on the mismatch value  $(h(\tilde{t}_i) - h(\mathbf{t}_2^*))$ . If mismatch is large for some  $i$ , the Euler step could lead to a large error for a fixed step size  $\ell$ . We discuss the assumptions in [theorems 1 and 2](#) in [appendix A.1](#)

### 3.2 Effect Connectivity and the Existence of $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$

The key element in [Theorem 1](#) is the surrogate intervention  $\mathbf{t}'$  such that its conditional effect given  $h(\mathbf{t}')$ , equals that of  $\mathbf{t}^*$  and  $h(\mathbf{t}_2^*)$ . The orthogonality  $\nabla_{\tilde{t}} f^\top \nabla_{\tilde{t}} h = 0$ , is a functional condition that does not guarantee  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  exists in  $\text{supp}(\mathbf{t})$ ; a necessity to compute  $\mathbb{E}[\mathbf{y} | \mathbf{t} = \mathbf{t}']$  without additional parametric assumptions. We give a general condition called *Effect Connectivity* that guarantees the surrogate intervention exists. With conditional effect  $\phi(\mathbf{t}^*, h(\mathbf{t}_2^*))$ , for any  $\mathbf{t}^*$

$$p(h(\mathbf{t}) = h(\mathbf{t}_2^*)) > 0 \implies p(\phi(\mathbf{t}, h(\mathbf{t})) = \phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) | h(\mathbf{t}) = h(\mathbf{t}_2^*)) > 0. \quad (9)$$

In words,  $\mathbf{t}$  has a chance of setting the conditional effect to any possible value  $\text{supp}(\phi(\mathbf{t}, h(\mathbf{t}_2^*)))$  given any confounder value  $h(\mathbf{t}_2^*) \in \text{supp}(h(\mathbf{t}))$ . An equivalent statement is that every level set of the conditional effect  $\phi(\mathbf{t}^*, h(\mathbf{t}_2^*))$ , with  $h(\mathbf{t}_2^*)$  fixed, contains an intervention for each confounder value. That is, for some  $h(\mathbf{t}_2^*)$  define the level set  $A_c = \{\mathbf{t}^*; f(\mathbf{t}^*, h(\mathbf{t}_2^*)) = c\}$ , then  $\forall h(\mathbf{t}_2^*) \in \text{supp}(h(\mathbf{t}))$ ,  $p(\mathbf{t} \in A_c | h(\mathbf{t}) = h(\mathbf{t}_2^*)) > 0$ .

**Theorem 3.** Under Effect Connectivity, [eq. \(9\)](#), any surrogate intervention  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)) \in \text{supp}(\mathbf{t})$ .

We give the proof in [appendix A.4](#). Whether the intervention  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  can be found via tractable search is problem-specific. If the surrogate  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  exists  $\forall \mathbf{t}^*, h(\mathbf{t}_2^*)$ , then [eq. \(9\)](#) holds by definition of the surrogate. Effect Connectivity allows us to reason about values of  $f$  anywhere in  $\text{supp}(\mathbf{t}) \times \text{supp}(h(\mathbf{t}))$  using only samples from  $p(\mathbf{y}, \mathbf{t})$ . Further, it is necessary in EFC:

**Theorem 4.** Effect Connectivity is necessary for nonparametric effect estimation in EFC.

We prove this in [appendix A.5](#). Effect Connectivity ensures that causal models with different causal effects have different observational distributions. Then, parametric assumptions on the causal model are not necessary to estimate effects.

## 4 Experiments

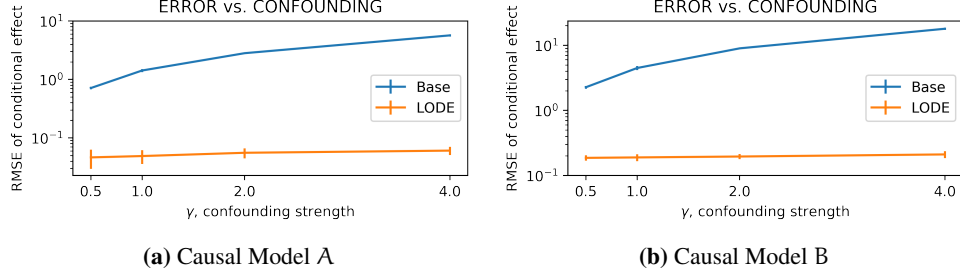
We evaluate LODE on simulated data first and show that LODE can correct for confounding. We also investigate the error induced by imperfect estimation of the surrogate intervention in LODE. Further, we run LODE on a GWAS dataset [\[6\]](#) and demonstrate that LODE is able to correct for confounding and recovers genetic variations that have been reported relevant to Celiac disease [\[8, 25, 14, 1\]](#).

### 4.1 Simulated experiments

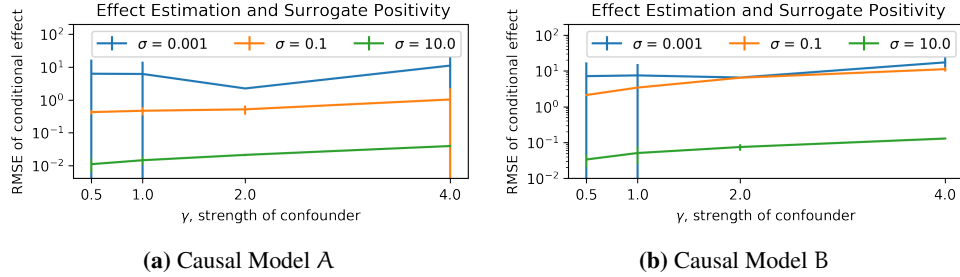
We investigate different properties of LODE on simulated data where ground truth is available. Let the dimension of  $\mathbf{t}$  (pre-outcome variables) be  $T = 20$  and outcome noise be  $\eta \sim \mathcal{N}(0, 0.1)$ . We consider two EFC causal models, denoted by A and B with different  $h(\mathbf{t})$  and  $f(\mathbf{t}, h(\mathbf{t}), \eta)$ :

$$(A) \quad h(\mathbf{t}) = \gamma \frac{\sum_i \mathbf{t}_i}{\sqrt{T}}, \quad \mathbf{t} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}^{T \times T}), \quad \mathbf{y} = \frac{\sum_i (-1)^i \mathbf{t}_i}{\sqrt{T}} + \alpha h(\mathbf{t})^2 + (1 + \alpha) h(\mathbf{t}) + \eta$$

$$(B) \quad h(\mathbf{t}) = \sum_{i: i \in 2\mathbb{Z}} \gamma \mathbf{t}_i \mathbf{t}_{i+1}, \quad \mathbf{t} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}^{T \times T}), \quad \mathbf{y} = \frac{\sum_i (-1)^i \mathbf{t}_i^2}{\sqrt{T}} + \alpha h(\mathbf{t}) + \eta$$



**Figure 3:** RMSE of estimated conditional effect vs. strength of confounding  $\gamma$ . LODE corrects for confounding and produces good effect estimates across different values of  $\gamma$ .

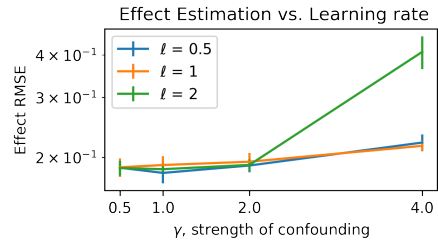


**Figure 4:** RMSE of estimated conditional effect estimate vs. the strength of confounding  $\gamma$ , for different levels of variance of  $t$ ,  $\sigma^2$ . Small  $\sigma$  leads to large conditional estimation error.

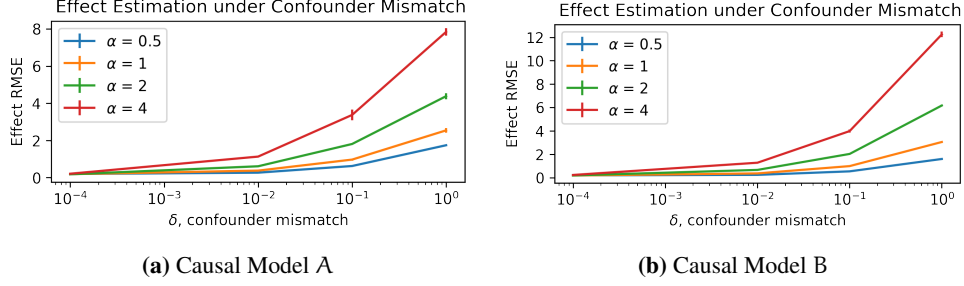
In both causal models, C-REDUNDANCY is satisfied. The constant  $\gamma$  controls the strength of the confounder and the constant  $\alpha$  controls the Lipschitz constant of the outcome as a function of the confounder. We let the variance  $\sigma^2 = 1$ , unless specified otherwise. In the following, we train on 1000 samples and report conditional effect root-mean-squared error (RMSE), computed with another 1000 samples. We used a degree-2 kernel ridge regression to fit the outcome model as a function of  $t$ . This model is correctly specified, and so the conditional  $\mathbb{E}[y | t = \tilde{t}]$  can be estimated well. We compare against a baseline estimate of conditional effect that is the same outcome model's estimate of  $\mathbb{E}[y | t = t^*]$ . This baseline fails to account for confounding and produces a biased estimate of the conditional effect of  $\text{do}(t = t^*)$ , conditional on any  $h(t_2^*) \neq h(t^*)$ .

First, we investigate how well LODE can correct for confounding for both causal models. We let  $\alpha = 1$  and obtain surrogate estimates by Euler integrating until the quantity  $\mathbb{E}_{t^*, h(t_2^*)} (h(\tilde{t}(s)) - h(t_2^*))^2$  is smaller than  $10^{-4}$  times value at initialization, where  $\mathbb{E}_{t^*, h(t_2^*)}$  is expectation over the evaluation set. In fig. 3, we plot the mean and standard deviation of conditional effect RMSE averaged over 10 seeds, for different strengths of confounding. We see that LODE is able to estimate effects well across multiple strengths of confounding while the baseline suffers.

Second, we investigate LODE's estimation when surrogate positivity holds but the probability  $p(t \approx t'(t^*, h(t_2^*)))$  is very small. This results in estimation error due to poor fitting of the outcome model in low density regions of  $\text{supp}(t)$ . We run LODE on simulated data where  $t$  is generated with different variances ( $\sigma^2$ ). For small  $\sigma$ , the outcome model error is large when using surrogate interventions  $t'(t^*, h(t_2^*))$ , where either  $h(t_2^*)$  or  $t^*$  is large. This leads to high variance effect estimation as we show in fig. 4 for both causal models. For various variances of  $t$ ,  $\sigma^2$ , we plot the mean and standard deviation of RMSE of estimated conditional effect over 10 seeds, against different  $\gamma$ .



**Figure 5:** RMSE of estimated conditional effect vs. step size in Euler Integrator in causal model B. Accumulating error due to large step size in Euler integrator increases with strength of confounding.



**Figure 6:** RMSE of estimated conditional effect vs. degree of confounder mismatch  $\delta$ . Error due to conditioning on a mismatched value of the confounder increases with strength of confounding but is mitigated by smoothness of the outcome function.

Third, we investigate the bias induced due to imperfect estimation of the surrogate intervention in LODE for both causal models. We construct surrogate interventions  $\tau'(\tau^*, h(\tau_2^*))$  by ensuring there is confounder-value mismatch  $h(\tilde{\tau}) \neq h(\tau_2^*)$ . We do this by interrupting Euler integration when the objective  $\mathbb{E}_{\tau^*, h(\tau_2^*)} (h(\tau'(\tau^*, h(\tau_2^*))) - h(\tau_2^*))^2 = \delta^2 > 0$ , where the  $\mathbb{E}_{\tau^*, h(\tau_2^*)}$  is over our evaluation set upon which we estimate conditional effects. For different  $\alpha$ , we plot in [fig. 6](#) the mean and standard deviation of RMSE of estimated conditional effect over 10 seeds, against different degrees of confounder mismatch,  $\delta$ . The error due to confounder mismatch is mitigated by small  $\alpha$ , the Lipschitz-constant of the outcome as a function of  $h(\mathbf{t})$ . Finally, we consider how step size in Euler integration affects the quality of estimated effects. Large step sizes may result in biased surrogate estimates; this bias is captured in the accumulation error in [section 3.1](#). We focus on the non-linear case in causal model B where gradient errors can accumulate (see [appendix A.3.1](#)). We demonstrate this error in [fig. 5](#) where we plot mean and standard deviation of conditional effect RMSE against the strength of confounding, for different step sizes  $\ell$ . We do not report results for larger step sizes ( $\ell > 2$ ) because Euler integration diverged for many surrogate estimates.

#### 4.2 Effects in Genetics (GWAS)

In this experiment, we explore the associations of genetic factors and Celiac disease. We utilize data from the Wellcome Trust Celiac disease GWAS dataset [8, 6] consisting of individuals with celiac disease, called cases ( $n = 3796$ ), and controls ( $n = 8154$ ). We construct our dataset by filtering from the  $\sim 550,000$  SNPs. The only preprocessing in our experiments is linkage disequilibrium pruning of adjacent SNPs (at 0.5  $R^2$ ) and PLINK [5] quality control. After this, 337,642 SNPs remain for 11,950 people. We imputed missing SNPs for each person by sampling from the marginal distribution of that SNP. No further SNP or person was dropped due to missingness. The objective of this experiment is to show that LODE corrects for confounding and recovers SNPs reported in the literature [8, 25, 14, 1]. To this end, after preprocessing, we included in our data 50 SNPs reported in [8, 25, 14, 1] and 1000 randomly sampled from the rest.

We use outcome models and functional confounders  $h(\cdot)$  traditionally employed in the GWAS literature. We choose a linear  $h(\tilde{\tau}) = A^\top \tilde{\tau}$ , where  $A$  is a matrix of the right singular vectors of a normalized Genotype matrix, that correspond to the top 10 singular values [19]. The outcome model is selected from logistic Lasso linear models with various regularization strengths, via cross validation within the training data (60% of the dataset). We defer details about the experimental setup to [appendix B](#).

We then use this outcome model in LODE to compute causal effects on the whole filtered dataset. The effects are computed one SNP at a time. First, for each person  $\tilde{\tau}$ , create  $\tilde{\tau}_i^1, \tilde{\tau}_i^0$  which correspond to the  $i$ th SNP set to 1 and 0 respectively, with all other SNPs same as  $\tilde{\tau}$ . Randomly sample a  $h(\tau_2^*)$  from the marginal  $p(h(\mathbf{t}))$  and, using the outcome model  $P_\theta$ , compute  $\phi(\tilde{\tau}, i) = \log P_\theta(y=1 | \tau'(\tilde{\tau}_i^1, h(\tau_2^*))) / P_\theta(y=1 | \tau'(\tilde{\tau}_i^0, h(\tau_2^*)))$ . The average effect of SNP  $i$  is obtained by averaging across all persons:  $\sum_{\tilde{\tau}} \phi(\tilde{\tau}, i) / N$ . Any SNP that beats a specified threshold of effect is deemed relevant to Celiac disease by LODE. We use a 60 – 40% train-test split, and outcome model selection is done via cross-validation within the training set. We did 5-fold cross-validation using just the training set. We use Scikit-learn [18] to fit the outcome models and for cross-validation.

**Results** The best outcome model was a Lasso model, trained with regularization constant 10. We select relevant SNPs by thresholding estimated effects at a magnitude  $> 0.1$ . From 1050 SNPs (1000



not reported before) LODE returned 31 SNPs, out of which 13 were previously reported as being associated with Celiac disease [8, 25, 14, 1]. In [appendix B.2](#) we plot the true positive and false negative rates of identifying previously reported SNPs, as a function of the effect threshold.

In [table 1](#), we list a few SNPs that were both deemed relevant by LODE and were reported in existing literature [8, 25, 14, 1], their effects, and their Lasso coefficients. The full list is in [table 2](#) in [appendix B](#). If LODE cannot adjust for confounding, the Lasso coefficients would dictate the effects; 0 coefficient means 0 effect. However, the two pairs of SNPs in [table 1](#) show that the effects estimated by LODE do not rely solely on the Lasso coefficients. For the first pair (rs13151961, rs2237236), the effect is the same but the coefficient of one is 0, while the other is positive. We note that rs2237236 was found to be associated with ulcerative colitis [12, 2], which is an inflammatory bowel disease that has been reported to share some common genetic basis with celiac disease [16]. For the second pair, (rs1738074, rs11221332), the magnitude of the effect is smaller for the former, but the coefficient is larger. Thus, LODE adjusts for confounding factors that the outcome model ignored.

SNP	EFFECT.	COEF.
rs13151961	0.17	0.32
rs2237236	0.17	0.00
rs1738074	−0.16	−0.23
rs11221332	−0.15	−0.24

**Table 1:** A few SNPs previously reported as relevant and recovered by LODE, with estimated effects and Lasso coefficients. LODE produces effect estimates that do not rely purely on the coefficients.

## 5 Discussion

When positivity is violated in traditional OBS-CI, not all effects are estimable without further assumptions. In such cases, practitioners have to turn to parametric models to estimate causal effects. However, parametric models can be misspecified when used without underlying causal mechanistic knowledge. We develop a new general setting of observational causal effect estimation called estimation with functional confounders (EFC) where the confounder can be expressed as a function of the data, meaning positivity is violated. Even when positivity is violated, the effects of many functional interventions are estimable. We develop a sufficient condition called functional positivity (F-POSITIVITY) to estimate effects of functional interventions. Such effects could be of independent interest; like the effect of cumulative dosage of a drug instead of joint effects of multiple dosages at different times.

Second, we prove a necessary condition for nonparametric estimation of effects of the full intervention. We propose the C-REDUNDANCY condition, under which, the effect of the full intervention on  $\mathbf{t}$  is estimable without parametric restrictions. We develop Level-set Orthogonal Descent Estimation (LODE) that computes surrogate interventions whose effects are estimable and match a conditional effect of interest. Further, we give bounds on errors ([theorem 2](#)) induced due to imperfect estimation of the surrogate intervention. Finally, we empirically demonstrate LODE’s ability to correct for confounding in both simulated and real data.

**Future.** A few directions of improvement remain which we elaborate next. First, F-POSITIVITY may not hold for all functions  $g(\mathbf{t})$  that we want to intervene on. Instead, one could compute a “projection”  $g_{\Pi}$  to the space of functions that satisfy F-POSITIVITY and inspect the effects defined by  $g_{\Pi}$  instead. A second direction of interest is to let  $h(\mathbf{t})$  only account for a part of the confounding, meaning ignorability is violated. This bias could be mitigated under smoothness conditions of the outcome function and its interaction with the degree of violation of ignorability.

Finally, LODE’s search strategy is Euler integration, which is equivalent to gradient descent with a fixed step size. Optimization techniques like momentum, rescaling the gradient using an adaptive matrix, and using second order hessian information, speed up gradient descent. However, if there are many local or global minima for  $(h(\tilde{\mathbf{t}}) - h(\mathbf{t}_2^*))^2$ , such techniques will result in a different solution than Euler integration, which could mean that effect estimates are biased. One extension of LODE would allow for search strategies that use such techniques.

## Broader Impact

Our work mainly applies to causal inference where confounders are specified as functions of observed data, such as in problems in genetics and healthcare. We choose to assess the impact of our work through its applications in these fields. A positive impact of the work is that better estimates of causal effects helps guide treatment for people and aid in understanding biological pathways of diseases. However, in healthcare, data collected in hospitals has biases. If, for instance, a certain demographic of people have more complete data collected about them, then this demographic would have better quality effect estimates, potentially meaning that they receive better treatment. This problem could be characterized by evaluating the positivity of treatment and completeness of confounders in electronic health record data split by demographics.

## Acknowledgements

The authors were partly supported by NIH/NHLBI Award R01HL148248, and by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science. The authors would like to thank Xintian Han, Raghav Singhal, Victor Veitch, Fredrik D. Johansson and the reviewers for thoughtful feedback. The authors would also like to thank Mukund Sudarshan and Prof. Sriram Sankararaman for help with running the GWAS experiments.

## References

- [1] Svetlana Adamovic, SS Amundsen, BA Lie, AH Gudjonsdottir, H Ascher, J Ek, DA Van Heel, S Nilsson, LM Sollid, and Å Torinsson Naluai. Association study of *il2/il21* and *fcgr2a*: significant association with the *il2/il21* region in scandinavian coeliac disease families. *Genes and immunity*, 9(4):364, 2008.
- [2] Carl A Anderson, Gabrielle Boucher, Charlie W Lees, Andre Franke, Mauro D’Amato, Kent D Taylor, James C Lee, Philippe Goyette, Marcin Imielinski, Anna Latiano, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics*, 43(3):246, 2011.
- [3] Uri M Ascher and Linda R Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*, volume 61. Siam, 1998.
- [4] William Astle, David J Balding, et al. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.
- [5] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.
- [6] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.
- [7] J. Correa and E. Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.
- [8] Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295, 2010.
- [9] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.
- [10] Miguel A Hernán and James M Robins. Causal inference: what if. *Boca Raton: Chapman & Hill/CRC*, 2020, 2020.
- [11] Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL <https://doi.org/10.1198/jcgs.2010.08162>.
- [12] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

- [13] Morris W Hirsch, Robert L Devaney, and Stephen Smale. *Differential equations, dynamical systems, and linear algebra*, volume 60. Academic press, 1974.
- [14] Karen A Hunt, Alexandra Zhernakova, Graham Turner, Graham AR Heap, Lude Franke, Marcel Bruinenberg, Jihane Romanos, Lotte C Dinesen, Anthony W Ryan, Davinder Panesar, et al. Novel celiac disease genetic determinants related to the immune response. *Nature genetics*, 40(4):395, 2008.
- [15] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833, 2011.
- [16] Virginia Pascual, Romina Dieli-Crimi, Natalia López-Palacios, Andrés Bodas, Luz María Medrano, and Concepción Núñez. Inflammatory bowel disease and celiac disease: overlaps and differences. *World journal of gastroenterology: WJG*, 20(17):4846, 2014.
- [17] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.
- [18] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [19] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.
- [20] Rajesh Ranganath and Adler Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.
- [21] Marc Ratkovic. Balancing within the margin: Causal effect estimation with support vector machines. *Department of Politics, Princeton University, Princeton, NJ*, 2014.
- [22] James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.
- [23] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [24] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.
- [25] Ludvig M Sollid. Coeliac disease: dissecting a complex inflammatory disorder. *Nature Reviews Immunology*, 2(9):647, 2002.
- [26] Michael Spivak. *Calculus on manifolds: a modern approach to classical theorems of advanced calculus*. CRC press, 2018.
- [27] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.
- [28] Timothy Thornton and Michael Wu. Summer institute in statistical genetics 2015.
- [29] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.
- [30] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, (just-accepted):1–71, 2019.
- [31] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203, 2006.

## A Theoretical details

### A.1 A note about the assumptions

**Note about the assumptions** In [theorem 1](#), assumption 1 consists of three parts that can all be validated on observed data: 1) that the gradient flow converges, 2) that the confounder value of the surrogate matches the confounder value whose effect is of interest, and 3) that the surrogate intervention lies in the support of the pre-outcome variables. Assumption 2 is required for expectations and their gradients to exist and be finite. In [theorem 2](#), assumption 1 requires a consistent estimator of  $\mathbb{E}[\mathbf{y} | \mathbf{t}]$ , which can be provided with regression. Assumption 3 lists regularity conditions which help control how the surrogate estimation error propagates to the effect error.

### A.2 Proof of [Theorem 1](#)

We restate the theorem for completeness:

**Theorem 1.** Assume C-REDUNDANCY holds. Assuming the following:

1. Let  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  be the limiting solution to the gradient flow equation  $\frac{d\tilde{\mathbf{t}}(s)}{ds} = -\nabla_{\tilde{\mathbf{t}}} (h(\tilde{\mathbf{t}}(s)) - h(\mathbf{t}_2^*))^2$ , initialized at  $\tilde{\mathbf{t}}(0) = \mathbf{t}^*$ ; i.e.  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)) = \lim_{s \rightarrow \infty} \tilde{\mathbf{t}}(s)$ . Further, let  $h(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))) = h(\mathbf{t}_2^*)$  and  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)) \in \text{supp}(\mathbf{t})$ .
2.  $f(\tilde{\mathbf{t}}, h(\tilde{\mathbf{t}}), \boldsymbol{\eta})$  and  $h(\tilde{\mathbf{t}})$  as functions of  $\tilde{\mathbf{t}}, h(\tilde{\mathbf{t}})$  are continuous and differentiable and the derivatives exist for all  $\tilde{\mathbf{t}}, \boldsymbol{\eta}$ . Let  $\nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, h(\tilde{\mathbf{t}}), \boldsymbol{\eta})$  exist and be bounded and integrable w.r.t. the probability measure corresponding to  $\mathbf{p}(\boldsymbol{\eta})$ , for all values of  $\tilde{\mathbf{t}}$  and  $h(\tilde{\mathbf{t}})$ .

Then the conditional effect (and therefore the average effect) is identified:

$$\phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) = \phi(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)), h(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)))) = \mathbb{E}[\mathbf{y} | \mathbf{t} = \mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))] \quad (10)$$

*Proof.* Recall definition of conditional effect  $\phi(\tilde{\mathbf{t}}, h(\tilde{\mathbf{t}})) = \mathbb{E}_{\boldsymbol{\eta}} f(\tilde{\mathbf{t}}, h(\tilde{\mathbf{t}}), \boldsymbol{\eta})$ . Recall  $\nabla_{\tilde{\mathbf{t}}}$  is the gradient with respect to the first argument of  $f$ , that is  $\tilde{\mathbf{t}}$ . First, by assumption 2,  $\mathbb{E}$  and  $\nabla$  commute, under the dominated convergence theorem. Then, by C-REDUNDANCY

$$\nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*))^\top \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}) = \nabla_{\tilde{\mathbf{t}}} \mathbb{E}_{\boldsymbol{\eta}} f(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*), \boldsymbol{\eta})^\top \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}) = \mathbb{E}_{\boldsymbol{\eta}} [\nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*), \boldsymbol{\eta})^\top \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}})] = 0.$$

Now consider the gradient flow equation  $\frac{d\tilde{\mathbf{t}}(s)}{ds} = -\nabla_{\tilde{\mathbf{t}}} (h(\tilde{\mathbf{t}}) - h(\mathbf{t}_2^*))^2$ . We refer to the gradient evaluated at  $\tilde{\mathbf{t}}$  as  $\Delta\tilde{\mathbf{t}} = -\nabla_{\tilde{\mathbf{t}}} (h(\tilde{\mathbf{t}}) - h(\mathbf{t}_2^*))^2 = -2(h(\tilde{\mathbf{t}}) - h(\mathbf{t}_2^*)) \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}})$ . We will express  $\phi(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)), h(\mathbf{t}_2^*))$  as defined by the starting point  $\phi(\mathbf{t}^*, h(\mathbf{t}_2^*))$  and the gradient flow equation.

Let the solution path to the gradient flow equation be  $C$  with  $\mathbf{t}^*, \mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  being the starting and ending points respectively. By the Gradient Theorem [\[26\]](#), we have that  $\phi(\mathbf{t}^*, h(\mathbf{t}_2^*))$  and  $\phi(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)), h(\mathbf{t}_2^*))$  are related via the line integral over  $C$ :

$$\int_C \nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*)) \cdot d\tilde{\mathbf{t}} = \phi(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)), h(\mathbf{t}_2^*)) - \phi(\mathbf{t}^*, h(\mathbf{t}_2^*))$$

Let  $\tilde{\mathbf{t}}(s)$  be a parametrization of solution path  $C$  by the scalar time  $s \in [0, \infty)$ . Now, to obtain the value of  $\phi(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*))$ , we will compute the line integral over the vector field defined by  $\nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*))$ , which exists by assumption 2 in [theorem 1](#), evaluated along the path  $C$  defined by  $\Delta\tilde{\mathbf{t}}(s)$ :

$$\begin{aligned} \phi(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)), h(\mathbf{t}_2^*)) &= \phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) + \int_C \nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}, h(\mathbf{t}_2^*)) \cdot d\tilde{\mathbf{t}} \\ &= \phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) + \int_0^\infty \nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}(s), h(\mathbf{t}_2^*))^\top \frac{d\tilde{\mathbf{t}}(s)}{ds} ds \\ &= \phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) + \int_0^\infty \nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}(s), h(\mathbf{t}_2^*))^\top \Delta\tilde{\mathbf{t}}(s) ds \\ &= \phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) \\ &\quad + \int_0^\infty -2((h(\tilde{\mathbf{t}}(s)) - h(\mathbf{t}_2^*))) \nabla_{\tilde{\mathbf{t}}} \phi(\tilde{\mathbf{t}}(s), h(\mathbf{t}_2^*))^\top \nabla_{\tilde{\mathbf{t}}} h(\tilde{\mathbf{t}}(s)) ds \\ &= \phi(\mathbf{t}^*, h(\mathbf{t}_2^*)) + 0 \quad \{\text{by C-REDUNDANCY}\} \end{aligned} \quad (11)$$

Finally, by assumption 1 in [theorem 1](#),  $h(t'(t^*, h(t_2^*))) = h(t_2^*)$ , and so

$$\phi(t^*, h(t_2^*)) = \phi(t'(t^*, h(t_2^*)), h(t_2^*)) = \phi(t'(t^*, h(t_2^*)), h(t'(t^*, h(t_2^*)))) \quad (12)$$

For clarity, the same equation, but using  $t'$  and suppressing dependence on  $t^*, h(t_2^*)$ :

$$\phi(t^*, h(t_2^*)) = \phi(t', h(t_2^*)) = \phi(t', h(t')) \quad (13)$$

Under the causal model for EFC, the outcome  $y = f(t, h(t), \eta)$ . Then,  $\forall \tilde{t} \in \text{supp}(p(t))$ ,

$$\mathbb{E}[y | t = \tilde{t}] = \mathbb{E}_\eta[f(\tilde{t}, h(\tilde{t}), \eta)] = \phi(\tilde{t}, h(\tilde{t})). \quad (14)$$

Using that  $t'(t^*, t_2^*) \in \text{supp}(p(t))$  and [eqs. \(13\) and \(14\)](#), the conditional effect is identified

$$\begin{aligned} \phi(t^*, h(t_2^*)) &= \phi(t'(t^*, h(t_2^*)), h(t'(t^*, h(t_2^*)))) \\ &= \mathbb{E}[y | t = t'(t^*, h(t_2^*))] \end{aligned} \quad (15)$$

Thus, the conditional effect, and consequently the average effect, are identified as  $\mathbb{E}[y | t'(t^*, h(t_2^*))]$  and  $\tau(t^*) = \mathbb{E}_{h(t)} \mathbb{E}[y | t'(t^*, h(t))]$  respectively.  $\square$

**Note about convergence of gradient flow** Any ODE's solution, if it exists and converges, converges to an  $\omega$ -limit set [\[27\]](#). An  $\omega$ -limit set is nonempty when the solution path lies entirely in a closed and bounded set and can consist of limit cycles, equilibrium points, or neither [\[13, 27\]](#). A gradient flow equation  $d\tilde{t}(s)/ds = -\nabla h(\tilde{t})$  (also called a gradient system) has the special property that its  $\omega$ -limit set only consists of critical points of  $h(\tilde{t})$ ; critical points of  $h(\tilde{t})$  are also equilibrium points of the gradient flow equation [\[13\]](#). Further, if  $\nabla h(\tilde{t})$  exists and is bounded and  $h(\tilde{t})$  has bounded sublevel sets ( $\{\tilde{t} : h(\tilde{t}) \leq c\}$ ), then the solution to the gradient flow equation will entirely lie within a bounded set. This is because along the solution path,  $h(\tilde{t}(s))$  always decreases meaning that the solution will remain in any sublevel set it started in. Thus, if  $h(\tilde{t})$  has bounded sublevel sets, the solution of the gradient flow equation will converge only to critical points of  $h(\tilde{t})$ .

### A.3 Estimation error in LODE

**Theorem 2.** Consider the conditional effect  $\phi(t^*, h(t_2^*))$ . Let  $\hat{t}(t^*, h(t_2^*))$  be the estimate of the surrogate intervention computed by LODE, computed via Euler integration of the gradient flow  $\frac{d\tilde{t}(s)}{ds} = -\nabla_{\tilde{t}}(h(\tilde{t}(s)) - h(t_2^*))^2$ , initialized at  $\tilde{t}(0) = t^*$ . Assume the true surrogate  $t'(t^*, h(t_2^*))$  exists and is the limiting solution to the gradient flow equation.

1. Let the finite sample estimator of  $\mathbb{E}[y | t = \tilde{t}]$  be  $\hat{f}(\tilde{t})$ . Let the error for all  $\tilde{t}$  be bounded,  $|\hat{f}(\tilde{t}) - \mathbb{E}[y | t = \tilde{t}]| \leq c(N)$ , where  $N$  is the sample size and  $\lim_{N \rightarrow \infty} c(N) = 0$ .
2. Assume  $K$  Euler integrator steps were taken to find the surrogate estimate  $\hat{t}(t^*, h(t_2^*))$ , each of size  $\ell$ . Let the maximum confounder mismatch be  $\max_{i \leq K} (h(\tilde{t}_i) - h(t_2^*))^2 = M$ .
3. Let  $L_{z, \tilde{t}}$  be the Lipschitz-constant of  $\phi(\tilde{t}, h(\tilde{t}_2))$  as a function of  $h(\tilde{t}_2)$ , for fixed  $\tilde{t}$ . Let  $L_e$  be the Lipschitz-constant of  $\mathbb{E}[y | t = \tilde{t}] = \phi(\tilde{t}, h(\tilde{t}))$  as a function of  $\tilde{t}$ . Assume  $h$  has a gradient with bounded norm,  $\|\nabla h(\tilde{t})\|_2 \leq L_h$ . Assume  $f$ 's Hessian has bounded eigenvalues:  $\forall \tilde{t}, \tilde{t}_2, \|\nabla_{\tilde{t}}^2 \phi(\tilde{t}, h(\tilde{t}_2))\|_2 \leq \sigma_{H\phi}$ .

The conditional effect estimate error,  $\xi(t^*, h(t_2^*)) = |\hat{f}(\hat{t}) - \phi(t^*, h(t_2^*))|$ , is upper bounded by:

$$c(N) + \min(L_e \|t' - \hat{t}\|_2, 2K\ell^2 (\mathcal{O}(\ell) + M\sigma_{H\phi}L_h^2) + L_{z, \hat{t}} \|h(\hat{t}) - h(t_2^*)\|_2) \quad (16)$$

*Proof.* (of [Theorem 2](#)) Recall the definition of conditional effect:  $\phi(\tilde{t}, h(\tilde{t}_2)) = \mathbb{E}_\eta f(\tilde{t}, h(\tilde{t}_2), \eta)$ .

LODE's estimate of the conditional effect is  $\hat{f}(\hat{t}(t^*, h(t_2^*)))$ . We will suppress notation for dependence on  $t^*, h(t_2^*)$ , and use  $t'$  and  $\hat{t}$  to refer to the true surrogate intervention and the estimated surrogate interventions respectively. Note  $\hat{f}$  is the estimate of the conditional expectation  $\mathbb{E}[y | t = \tilde{t}]$ , learned from  $N$  samples. We first bound the error by splitting into two parts and bounding each separately:

$$\begin{aligned} |\xi(t^*, h(t_2^*))| &= |\hat{f}(\hat{t}) - \phi(t^*, h(t_2^*))| \\ &\leq |\hat{f}(\hat{t}) - \phi(\hat{t}, h(\hat{t}))| + |\phi(\hat{t}, h(\hat{t})) - \phi(t^*, h(t_2^*))| \\ &\leq c(N) + |\phi(\hat{t}, h(\hat{t})) - \phi(t^*, h(t_2^*))| \\ &\leq |\phi(\hat{t}, h(\hat{t})) - \phi(\hat{t}, h(t_2^*))| + |\phi(\hat{t}, h(t_2^*)) - \phi(t^*, h(t_2^*))| + c(N) \end{aligned}$$



The first term is bounded via the Lipschitz-ness of  $\phi$  as a function of  $h(\tilde{\tau})$  with fixed first argument  $\tilde{\tau} = \hat{\tau}$ .

$$|\phi(\hat{\tau}, h(\hat{\tau})) - \phi(\hat{\tau}, h(\tau_2^*))| \leq L_{z,\hat{\tau}} \|h(\hat{\tau}) - h(\tau_2^*)\|$$

We now bound the remaining term. Recall that LODÉ's computation of the surrogate intervention involved  $K$  gradient steps, each of size  $\ell$ . We work with a constant step-size but the analysis can be generalized to a non-uniform step size. Indexing steps with  $i$ , let  $d_i = h(\tilde{\tau}_i) - h(\tau_2^*)$  be the confounder mismatch error at the  $i$ th iterate. Then note that  $\hat{\tau} = \tau^* - \ell \sum_{i=0}^{K-1} 2d_i \nabla_{\tilde{\tau}} h(\tilde{\tau}_i)$ . We can use this to bound the error  $\phi(\hat{\tau}, h(\tau_2^*)) - \phi(\tau^*, h(\tau_2^*))$ . With  $\tilde{\tau}_K = \hat{\tau}$  and  $\tilde{\tau}_0 = \tau^*$ , we proceed by expressing the error as a telescoping sum and using the Taylor expansion for  $\phi(\tilde{\tau}, h(\tau_2^*))$  in terms of the the first argument  $\tilde{\tau}$ .

$$\phi(\hat{\tau}, h(\tau_2^*)) - \phi(\tau^*, h(\tau_2^*)) = \sum_{i=0}^{K-1} \phi(\tilde{\tau}_{i+1}, h(\tau_2^*)) - \phi(\tilde{\tau}_i, h(\tau_2^*)) \quad (17)$$

$$= \sum_{i=0}^{K-1} \nabla_{\tilde{\tau}} \phi(\tilde{\tau}_i, h(\tau_2^*))^\top (\tilde{\tau}_{i+1} - \tilde{\tau}_i) \quad (18)$$

$$+ \frac{1}{2} (\tilde{\tau}_{i+1} - \tilde{\tau}_i)^\top \nabla_{\tilde{\tau}}^2 \phi(\tilde{\tau}_i, h(\tau_2^*)) (\tilde{\tau}_{i+1} - \tilde{\tau}_i) + \mathcal{O}(\|\tilde{\tau}_{i+1} - \tilde{\tau}_i\|_2^3) \quad (19)$$

$$= \sum_{i=0}^{K-1} 2\ell d_i \nabla_{\tilde{\tau}} \phi(\tilde{\tau}_i, h(\tau_2^*))^\top \nabla_{\tilde{\tau}} h(\tilde{\tau}_i) + 2(\ell d_i)^2 \nabla_{\tilde{\tau}} h(\tilde{\tau}_i)^\top \nabla_{\tilde{\tau}}^2 \phi(\tilde{\tau}_i, h(\tau_2^*)) \nabla_{\tilde{\tau}} h(\tilde{\tau}_i) + \mathcal{O}(\ell^3) \quad (20)$$

$$= \sum_{i=0}^{K-1} 0 + 2(\ell d_i)^2 \nabla_{\tilde{\tau}} h(\tilde{\tau}_i)^\top \nabla_{\tilde{\tau}}^2 \phi(\tilde{\tau}_i, h(\tau_2^*)) \nabla_{\tilde{\tau}} h(\tilde{\tau}_i) + \mathcal{O}(\ell^3) \quad (21)$$

$$= \mathcal{O}(K\ell^3) + \sum_{i=0}^{K-1} 2(\ell d_i)^2 \nabla_{\tilde{\tau}} h(\tilde{\tau}_i)^\top \nabla_{\tilde{\tau}}^2 \phi(\tilde{\tau}_i, h(\tau_2^*)) \nabla_{\tilde{\tau}} h(\tilde{\tau}_i) \quad (22)$$

$$\leq \mathcal{O}(K\ell^3) + \sum_{i=0}^{K-1} 2(\ell(h(\tilde{\tau}_i) - h(\tau_2^*)))^2 |\nabla_{\tilde{\tau}} h(\tilde{\tau}_i)^\top \nabla_{\tilde{\tau}}^2 \phi(\tilde{\tau}_i, h(\tau_2^*)) \nabla_{\tilde{\tau}} h(\tilde{\tau}_i)| \quad (23)$$

$$\leq \mathcal{O}(K\ell^3) + \sum_{i=0}^{K-1} 2\ell^2 M |\nabla_{\tilde{\tau}} h(\tilde{\tau}_i)^\top \nabla_{\tilde{\tau}}^2 \phi(\tilde{\tau}_i, h(\tau_2^*)) \nabla_{\tilde{\tau}} h(\tilde{\tau}_i)| \quad (24)$$

$$\leq \mathcal{O}(K\ell^3) + \sum_{i=0}^{K-1} 2\ell^2 M \sigma_{\text{H}\phi} \|\nabla_{\tilde{\tau}} h(\tilde{\tau}_i)\|_2^2 \quad (25)$$

$$\leq \mathcal{O}(K\ell^3) + \sum_{i=0}^{K-1} 2\ell^2 M \sigma_{\text{H}\phi} L_h^2 \quad (26)$$

$$= 2K\ell^2 (\mathcal{O}(\ell) + M\sigma_{\text{H}\phi} L_h^2), \quad (27)$$

where the inequalities follow by the maximum value of  $(h(\tilde{\tau}_i) - h(\tau_2^*))^2$ , bounded eigenvalues of the Hessian of  $\phi$  and the Lipschitz-ness of  $h(\tilde{\tau})$ .

Another way we bound the error is via the Lipschitz constant of the conditional expectation as a function of  $\tilde{\tau}$ . Recall this is  $L_e$ . An alternate bound on the error is as follows:

$$|\phi(\hat{\tau}, h(\hat{\tau})) - \phi(\tau^*, h(\tau_2^*))| = |\phi(\hat{\tau}, h(\hat{\tau})) - \phi(\tau', h(\tau'))| \leq L_e \|\tau' - \hat{\tau}\|_2$$

The bound follows:

$$|\xi(\tilde{\tau}, h(\tau_2^*))| \leq c(N) + \min(L_e \|\tau' - \hat{\tau}\|_2, \quad 2K\ell^2 (\mathcal{O}(\ell) + M\sigma_{\text{H}\phi} L_h^2) + L_{z,\hat{\tau}} \|h(\hat{\tau}) - h(\tau_2^*)\|_2)$$

□

### A.3.1 A note on linear confounder functions and LODE

In the proof above, the error in Euler integration accumulates due to terms like this one:  $\nabla_{\tilde{\mathbf{t}}}^\top \mathbf{h}(\tilde{\mathbf{t}}) \nabla_{\tilde{\mathbf{t}}}^2 f(\tilde{\mathbf{t}}, \mathbf{h}(\mathbf{t}^*), \boldsymbol{\eta}) \nabla_{\tilde{\mathbf{t}}} \mathbf{h}(\tilde{\mathbf{t}})$ . For a linear confounder function that satisfies  $\nabla_{\tilde{\mathbf{t}}} \mathbf{h}(\tilde{\mathbf{t}}) = \beta$ , such terms can be expressed as  $\beta^\top \nabla_{\tilde{\mathbf{t}}} (\nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, \mathbf{h}(\mathbf{t}^*), \boldsymbol{\eta})^\top \beta) = \beta^\top \nabla_{\tilde{\mathbf{t}}} (0) = 0$  under C-REDUNDANCY. Thus, such error does not accumulate even with large step sizes.

Further, note that the gradient flow equation in LODE for the causal model  $A$  in [section 4](#) is a linear ODE whose solution has a closed form expression and one can estimate the surrogate without numerical integration [\[27\]](#).

### A.4 Proof of sufficiency of Effect Connectivity

**Theorem 3.** *Under Effect Connectivity, [eq. \(9\)](#), any surrogate intervention  $\mathbf{t}'(\mathbf{t}^*, \mathbf{h}(\mathbf{t}_2^*)) \in \text{supp}(\mathbf{t})$ .*

*Proof.* Recall  $\phi(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}})) = \mathbb{E}_{\boldsymbol{\eta}} f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}), \boldsymbol{\eta})$ . We have  $\forall \mathbf{t}^* \in \text{supp}(\mathbf{p}(\mathbf{t}))$ :

$$p(\mathbf{h}(\mathbf{t}) = \mathbf{h}(\mathbf{t}_2^*)) > 0 \implies p(\phi(\mathbf{t}, \mathbf{h}(\mathbf{t})) = \phi(\mathbf{t}^*, \mathbf{h}(\mathbf{t}_2^*)) \mid \mathbf{h}(\mathbf{t}) = \mathbf{h}(\mathbf{t}_2^*)) > 0.$$

This implies  $\exists \mathbf{t}' \in \text{supp}(\mathbf{t}), \phi(\mathbf{t}', \mathbf{h}(\mathbf{t}_2^*)) = \phi(\mathbf{t}^*, \mathbf{h}(\mathbf{t}_2^*))$ , s.t.  $\mathbf{h}(\mathbf{t}') = \mathbf{h}(\mathbf{t}_2^*)$ .

Then,  $\phi(\mathbf{t}^*, \mathbf{h}(\mathbf{t}_2^*)) = \phi(\mathbf{t}', \mathbf{h}(\mathbf{t}_2^*)) = \phi(\mathbf{t}', \mathbf{h}(\mathbf{t}')) = \mathbb{E}[\mathbf{y} \mid \mathbf{t} = \mathbf{t}']$ . □

### A.5 Necessity of Effect Connectivity for Nonparametric effect estimation in EFC

**Theorem 4.** *Effect Connectivity is necessary for nonparametric effect estimation in EFC.*

*Proof.* (Proof of [Theorem 4](#)) Let the outcome be  $\mathbf{y} = f(\mathbf{t}, \mathbf{h}(\mathbf{t}))$ . Recall the joint distribution  $p(\mathbf{t}, \mathbf{y})$  and let  $\mathbf{h}(\mathbf{t})$  be the confounder. Let Effect Connectivity be violated, i.e. there exists a non-measure-zero subset  $B \in \text{supp}(\mathbf{t}) \times \text{supp}(\mathbf{h}(\mathbf{t}))$  such that <sup>6</sup>:

$$\forall \tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2) \in B, \quad p(f(\mathbf{t}, \mathbf{h}(\mathbf{t})) = f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) \mid \mathbf{h}(\mathbf{t}) = \mathbf{h}(\tilde{\mathbf{t}}_2)) = 0.$$

Now, we construct a new outcome  $\mathbf{y}_2 = f_2(\mathbf{t}, \mathbf{h}(\mathbf{t}))$  and show the conditional effects for this new outcome are different from the one defined by  $f$  on  $\forall(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) \in B$ . Let

$$f_2(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) = f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) + 10 * 1((\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) \in B).$$

We have  $f_2(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}})) = f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}})) \forall \tilde{\mathbf{t}} \in \text{supp}(\mathbf{t})$ , as the additional term in  $f_2$  is only present for  $(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) \in B$ ; this follows from the fact that  $\forall \tilde{\mathbf{t}} \in \text{supp}(\mathbf{t}), (\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}})) \notin B$  as

$$p[f(\mathbf{t}, \mathbf{h}(\mathbf{t})) = f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}})) \mid \mathbf{h}(\mathbf{t}) = \mathbf{h}(\tilde{\mathbf{t}})] = p[f(\mathbf{t}, \mathbf{h}(\mathbf{t})) = f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}))] > 0.$$

Thus,  $p(\mathbf{y}, \mathbf{t}) \stackrel{d}{=} p(\mathbf{y}_2, \mathbf{t})$  are equal in distribution since  $B \cap \text{supp}(\mathbf{t}, \mathbf{h}(\mathbf{t})) = \emptyset$ . This means that the conditional effects are different for the outcomes  $\mathbf{y}, \mathbf{y}_2$  for all  $(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2)) \in B$ :

$$\mathbb{E}[\mathbf{y} \mid \text{do}(\mathbf{t} = \tilde{\mathbf{t}}), \mathbf{h}(\mathbf{t}) = \mathbf{h}(\tilde{\mathbf{t}}_2)] \neq \mathbb{E}[\mathbf{y}_2 \mid \text{do}(\mathbf{t} = \tilde{\mathbf{t}}), \mathbf{h}(\mathbf{t}) = \mathbf{h}(\tilde{\mathbf{t}}_2)]$$

Therefore, for causal models that violates Effect Connectivity, there exist observationally equivalent causal models with different causal effects. Thus, nonparametric effect estimation is impossible. Thus, Effect Connectivity is required for EFC. □

### A.6 Algorithmic details

We give in [algorithm 1](#) pseudocode for LODE.

**Extensions of LODE** Consider that we have access to  $m(\mathbf{h}(\mathbf{t}))$  for some bijective differentiable function  $m(\cdot)$ , instead of  $\mathbf{h}(\mathbf{t})$ . The orthogonality in C-REDUNDANCY holds  $\nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2), \boldsymbol{\eta})^\top \nabla_{\tilde{\mathbf{t}}} m(\mathbf{h}(\tilde{\mathbf{t}})) = m'(\mathbf{h}(\tilde{\mathbf{t}})) \nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2), \boldsymbol{\eta})^\top \nabla_{\tilde{\mathbf{t}}} \mathbf{h}(\tilde{\mathbf{t}}) = 0$ . Then, using  $m(\mathbf{h}(\tilde{\mathbf{t}}))$  to compute the surrogate  $\mathbf{t}'(\mathbf{t}^*, \mathbf{h}(\mathbf{t}_2^*))$ , LODE would estimate valid effects. Similarly, LODE can estimate the effect on any differentiable transformation of the outcome  $m(\mathbf{y})$ , because  $\nabla_{\tilde{\mathbf{t}}} m(\mathbf{y}_{\tilde{\mathbf{t}}})^\top \nabla_{\tilde{\mathbf{t}}} \mathbf{h}(\tilde{\mathbf{t}}) = m'(\mathbf{y}_{\tilde{\mathbf{t}}}) \nabla_{\tilde{\mathbf{t}}} f(\tilde{\mathbf{t}}, \mathbf{h}(\tilde{\mathbf{t}}_2), \boldsymbol{\eta})^\top \nabla_{\tilde{\mathbf{t}}} \mathbf{h}(\tilde{\mathbf{t}}) = 0$  holds.

<sup>6</sup>Non-zero w.r.t. the product measure over  $\text{supp}(\mathbf{t}) \times \text{supp}(\mathbf{h}(\mathbf{t}))$  due to  $p$ .

---

**Algorithm 1:** LOD for  $\text{do}(\mathbf{t} = \mathbf{t}^*)$ 

---

**Input:** Functional confounder  $h(\mathbf{t})$ ; tolerance  $\epsilon$ **Output:** Conditional effects of  $\mathbf{t}^*$ ,  $h(\mathbf{t}_2^*)$ 

- 1 Regress  $\mathbf{y}$  on  $\mathbf{t}$  and compute  $\hat{f}() := \arg \min_{\mathbf{u} \in \mathcal{F}} \mathbb{E}_{\mathbf{y}, \mathbf{t}} (\mathbf{y} - \mathbf{u}(\mathbf{t}))^2$ .
- 2 To estimate effects of  $\mathbf{t}^*$ ,  $h(\mathbf{t}_2^*)$ , compute the surrogate intervention  $\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*))$  by Euler integrating the gradient flow equation, initialized at  $\tilde{\mathbf{t}} = \mathbf{t}^*$ , until  $(h(\tilde{\mathbf{t}}_s) - h(\mathbf{t}_2^*))^2 < \epsilon$ .

$$\frac{d\tilde{\mathbf{t}}(s)}{ds} = \nabla_{\tilde{\mathbf{t}}} (h(\tilde{\mathbf{t}}_s) - h(\mathbf{t}_2^*))^2,$$

- 3 Return  $\hat{f}(\mathbf{t}'(\mathbf{t}^*, h(\mathbf{t}_2^*)))$ ;
- 

## B Experimental Details

### B.1 Functional confounders in GWAS

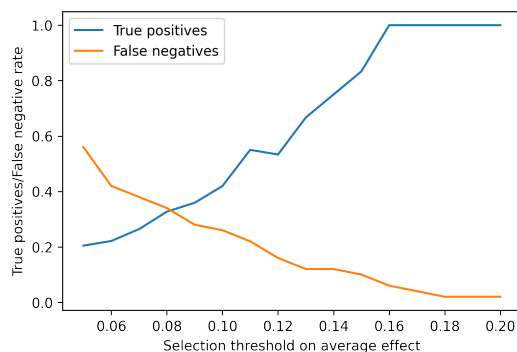
Here, we show how  $h(\mathbf{t}) = \mathbf{A}\mathbf{t}$  and  $\mathbf{A}$  reflect the traditional PCA based adjustment in GWAS. Recall population structure acts as a confounder in GWAS. Price et al. [19] demonstrated that using the principal components of the normalized genetic relationships matrix adjusts for confounding due to population structure in GWAS. Let the genotype matrix be  $\mathbf{G}$  with people as rows and SNPs as columns, such that each element is one of 0, 1/2, 1, where 1/2 and 1 refer to one and two copies of the allele respectively at the position of the SNP. With  $p_s$  as the allele frequency at SNP  $s$  [28],  $\Phi$  is the genetic relationship matrix whose elements are defined as  $\Phi_{i,j} = \frac{1}{S} \sum_{s=1}^S (G_{i,s} - p_s)(G_{j,s} - p_s) / (p_s(1 - p_s))$ . Then, Price et al. [19] compute the top  $K$  (10 suggested) principal components of  $\Phi$  to use as the axes of variation due to the population structure. The eigenvectors of  $\Phi$  are the left eigenvectors of  $\hat{\mathbf{G}}$  such that  $\Phi = \hat{\mathbf{G}}\hat{\mathbf{G}}^T$  which capture independent axes of variation of individuals.

Price et al. [19] exploit the idea that if a SNP aligns with some of the axes of variation, this is due to the population structure. These axes of variation are the top  $K$  eigenvectors  $\mathbf{U}$  of  $\Phi = \hat{\mathbf{G}}\hat{\mathbf{G}}^T \approx \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ , where  $\mathbf{U} \in \mathbb{R}^{N \times K}$ ,  $\Phi \in \mathbb{R}^{N \times N}$  and  $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$ . Here,  $\mathbf{U}$  are also the left singular vectors of  $\hat{\mathbf{G}} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$  where  $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$  is diagonal, and  $\mathbf{V} \in \mathbb{R}^{S \times K}$ . We use  $\approx$  to denote that the chosen  $K$  eigenvectors explain the variation due to population structure; what remains are random mutations.

Let the  $s$ th SNP be  $\hat{\mathbf{G}}_{:,s} \in \mathbb{R}^N$ , which is a column in  $\hat{\mathbf{G}}$ . In Price et al. [19], population structure in the  $s$ th SNP is captured in  $\hat{\mathbf{G}}_{:,s}^T \mathbf{U}$ . In words, projecting the SNP  $\hat{\mathbf{G}}_{:,s}$  onto the axes of variation in individuals gives the population structure between  $s$ th SNP and the outcome. This projection  $\hat{\mathbf{G}}_{:,s}^T \mathbf{U}$  is a row of  $\hat{\mathbf{G}}^T \mathbf{U} \in \mathbb{R}^{S \times K}$ . In turn,  $\hat{\mathbf{G}}^T \mathbf{U} \in \mathbb{R}^{S \times K}$  is the population structure in all SNPs. Projecting this population structure onto the genotype of an individual gives the confounding due to population structure amongst the SNPs present in the genotype. With  $\mathbf{G}_{j,\cdot} \in \{0, 1/2, 1\}^S$  as the genotype for an individual  $j$ , this projection is  $((\hat{\mathbf{G}}^T \mathbf{U})^T \mathbf{G}_{j,\cdot})$ . However,  $\hat{\mathbf{G}} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$  implies that  $\hat{\mathbf{G}}^T \mathbf{U} \approx \mathbf{V}\mathbf{\Sigma}$ . Reflecting this,  $h(\mathbf{t}) = \mathbf{\Sigma}\mathbf{V}^T \mathbf{t}$  is the functional confounder for an individual  $\mathbf{t}$ .

## B.2 Expanded results

In [table 2](#), we list the 13 SNPs recovered by LODE, that have been previously reported as relevant to Celiac disease. In [fig. 7](#), we plot the true positive and false negative rate amongst SNPs deemed relevant by LODE. The ground truth here are the SNPs reported associated with celiac disease in prior literature.



**Figure 7:** True positive vs. False negative rate as we vary the threshold on average effects, that determines which SNPs LODE deems relevant to the outcome.

SNP	EFFECT	LASSO COEF.
rs3748816	0.12	0.20
rs10903122	0.10	0.17
rs2816316	0.11	0.20
rs13151961	0.17	0.32
rs2237236	0.17	0.00
rs12928822	0.14	0.29
rs2187668	−0.70	−2.37
rs2327832	−0.12	−0.20
rs1738074	−0.16	−0.23
rs11221332	−0.15	−0.24
rs653178	−0.13	−0.21
rs4899260	−0.12	−0.19
rs17810546	−0.12	−0.20

**Table 2:** Full list of SNPs previously reported as relevant that were recovered by LODE, and their estimated effects and Lasso coefficients for SNPs. The effect threshold here is 0.1.