PROBABILISTIC COUNTERFACTUAL INFERENCE FOR DISCRETE OUTCOMES IN GAUSSIAN-PROCESS CAUSAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual inference in Gaussian-process structural causal models (GP-SCMs) has been developed primarily for continuous endogenous variables, limiting applicability to causal graphs that contain discrete child nodes with continuous parents. We introduce a unified probabilistic framework for counterfactual inference with heterogeneous variable types by pairing GP predictors with explicit exogenous noise mechanisms. For discrete outcomes, we derive exact conditional noise-abduction procedures using a uniform threshold for binary variables, a Gumbel-max race for nominal categories, and a latent Gaussian cut-point model for ordinal ones. In each case, we propagate abducted noise through interventions while accounting for posterior uncertainty in the GP latent functions, and prove that the resulting mechanisms reproduce the fitted model's observational and interventional distributions. On synthetic SCMs with known ground-truth counterfactuals, we evaluate estimation accuracy, consistency, and robustness to coupling misspecification. A key finding is that applying a categorical coupling to ordinal data inflates counterfactual error roughly threefold even when observational fit remains comparable, and that this error does not diminish with more data. As the training set grows, the fitted structural equation converges to the truth while the counterfactual error flattens onto a floor. In the reverse direction, forcing a false order onto nominal data instead degrades the fitted equation itself. The choice of coupling must therefore be justified on structural grounds rather than read off the fit.

1 INTRODUCTION

Incorporating causality into Machine Learning can be a crucial step towards developing realistic and explainable models. Pearl et al. (2016) has shown that structural causal models (SCMs) can be used to determine the effects of interventions and estimate counterfactual queries. Interventions compute the downstream effects of changing one or more variables in a system. Counterfactuals extend upon this by using knowledge of a factual instance to compute how the variables would have changed had an intervention been performed.

To incorporate this causal reasoning into a real-world system, researchers must develop an SCM comprised of a causal graph and structural equations. The graph structure describes which features are causal parents of others, which can be determined using domain knowledge and independence tests using observational data (Spirtes et al., 2000; Zhang et al., 2011; Hoyer et al., 2009). For this paper, we assume this step has been completed using one of many graph discovery techniques. From here, structural equations must be fitted for each parent-child relationship. Recent work has suggested that Gaussian process (GP) regression can approximate the variety of relationships that occur (linear, non-linear, additive) and provide uncertainty quantification over the resulting counterfactual (Karimi et al., 2020). This method has presently only been developed for relationships between real-valued features, and extending it to work for non-continuous child variables requires several adjustments. We explain how the probabilistic counterfactual estimation changes for binary, categorical and ordinal features. Our method is proven theoretically and validated experimentally using synthetic SCMs for ground-truth.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 COUNTERFACTUALS

We define counterfactuals and their evaluation as in Pearl et al. (2016):

Counterfactual Queries

Given observation $\mathbf{x}^F$, what would have happened if $\mathbf{X}_\mathcal{I}$ had instead taken the value $\boldsymbol{\theta}$? The counterfactual variable is written $\mathbf{x}^{\mathrm{CF}} = \mathbf{x}(\mathrm{do}(\mathbf{X}_\mathcal{I} = \boldsymbol{\theta})) \mid \mathbf{x}^F$.

Given a fully specified SCM M with d nodes and equations $\mathbf{F} = \{\mathbf{X}_r := f_r(\mathbf{X}_{\mathrm{pa}(r)}, U_r)\}_{r=1}^d$, and an observed instance $\mathbf{x}^F$, counterfactuals asking what would have happened had we set $\mathbf{X}_I = \theta$ are evaluated using the following three steps:

Three Steps for Computing Counterfactuals

1. *Abduction*: Use evidence $\mathbf{x}^F$ and the equations to solve backwards for $\{U_r\}_{r=1}^d$.
2. *Action*: Modify the model M, by performing the intervention $\mathrm{do}(\mathbf{X}_I = \theta)$, to get new structural equations $\mathbf{F}^{\mathrm{do}(\mathbf{X}_I=\theta)}$.
3. *Prediction*: Use the modified structural equations and the values of $\{U_r\}_{r=1}^d$ to compute the value of $\mathbf{x}^{\mathrm{CF}}$.

Typically, the most challenging step in this process is *Abduction*, because the method to isolate the factual noise depends on the form of the structural equations. The previous approach uses additive Gaussian noise, which allows for easy noise isolation Karimi et al. (2020), however this approach will not work for discrete features.

Consider a bank deciding whether to extend a loan. A continuous feature such as debt-to-income ratio determines an ordinal risk tier (subprime through super-prime, from thresholding a composite score), which in turn determines a nominal decision such as which loan product is offered. An applicant who is denied may reasonably ask what would have happened had their financial position been different. Answering this counterfactual query means carrying Abduction, Action, and Prediction through a causal chain whose structural equations mix continuous, ordinal, and nominal variables. Existing approaches are limited to scenarios with only real-valued downstream variables, leaving systems like the one described here out of their scope Karimi et al. (2020).

1.2 EXISTING APPROACHES

Prior work falls into three families, by which pipeline stage they address: *discovering* the graph, *fitting* equations on a given graph, or *coupling* a fixed interventional law into a counterfactual mechanism. We address the middle stage: fitting Gaussian Process equations paired with an *exact* noise-abduction identity. **Graph discovery** has begun exploiting ordinal structure to orient edges (Ni & Mallick, 2022), and Yao et al. (2025) use a categorical mechanism close to ours for graph identifiability in mixed additive noise models, without addressing counterfactual queries; we instead take the graph as given. For **fitting discrete or ordinal equations**, ordinal variables are classically modelled by discretising a latent score at fixed cutpoints (McKelvey & Zavoina, 1975; Boes, 2013), with Chu & Ghahramani (2005) giving the Gaussian Process likelihood our construction builds on; deep generative alternatives dequantise discrete variables into normalising-flow SCMs (Pawlowski et al., 2020), or relax categorical latents via a temperature-annealed softmax (Jang et al., 2017; Maddison et al., 2017) for gradient-based training rather than exact abduction, which we use instead. A separate line instead **couples a fixed interventional law**: Oberst & Sontag (2019) introduce the Gumbel-max coupling, singled out axiomatically by *counterfactual stability*, while Lorberbom et al. (2021) *learn* a coupling minimising counterfactual treatment-effect variance over a query distribution; our interventional law is itself a fitted, uncertain Gaussian Process, and we choose the coupling to match the endogenous variable's *type*, not to optimise a downstream objective.

2 PRELIMINARIES

2.1 INTRODUCING GUMBEL NOISE

We introduce the Gumbel distribution because it gives an exact, explicit exogenous noise variable for categorical outcomes. The same construction is used to relax categorical sampling into a differentiable operation for gradient-based training via the Gumbel-softmax/Concrete distribution (Jang et al., 2017; Maddison et al., 2017). We instead exploit its exact (non-relaxed) form to obtain closed-form noise abduction.

Gumbel Distribution (Gumbel, 1954)

$G \sim \mathrm{Gumbel}(\mu, 1)$ if it has CDF $F(g) = \exp\big(-e^{-(g-\mu)}\big)$ on $\mathbb{R}$.

We highlight three important properties of this distribution that we will use to evaluate the counterfactuals:

- **Reparameterisation:** $V \sim \mathrm{Uniform}(0,1) \implies \mu - \log(-\log V) \sim \mathrm{Gumbel}(\mu, 1)$.
- **Truncation** (Maddison et al., 2014): Conditioned on $G \le b$,
$$\mu - \log\big(e^{-(b-\mu)} - \log V\big) \sim \mathrm{Gumbel}(\mu, 1) \text{ for } V \sim \mathrm{Uniform}(0,1)$$
has the conditioned law.
- **Gumbel-max trick** (Yellott, 1977): For scores $\boldsymbol{\ell} \in \mathbb{R}^C$ and iid $G_1, \dots, G_C \sim \mathrm{Gumbel}(0,1)$, the perturbed maximum $M := \max_c\{\ell_c + G_c\}$ satisfies
$$M \sim \mathrm{Gumbel}\Big(\log \textstyle\sum_c e^{\ell_c},\, 1\Big) \quad \text{independently of } argmax_c\{\ell_c + G_c\} \sim \mathrm{softmax}(\boldsymbol{\ell})$$

2.2 COUPLING AND ABDUCTION

When the structural equations are approximated by fitted GPs, the uncertainty about f_r means that the exogenous noise driving X_r is never observed directly. Instead, its posterior is updated (abducted) from the factual observation, and this fixed abducted noise is carried into the counterfactual world while only the parents change. For continuous nodes as in Karimi et al. (2020), one can use additive Gaussian noise that can be uniquely solved for given a factual instance.

For a discrete outcome, this uniqueness breaks. The factual and counterfactual class distributions, $\mathbf{p}^F$ and $\mathbf{p}^{\mathrm{CF}}$, are each pinned down by the fitted model, but they alone do not determine a joint law over the pair (X_r, X_r^{CF}): infinitely many joint distributions share the same two marginals. Call any such choice a *coupling*. Abduction is how a coupling gets realised in practice: we equip X_r with an explicit exogenous noise variable, infer its posterior from the factual observation x_r^F, and reuse that same noise, unchanged, to generate X_r^{CF} under the counterfactual parents. Because the noise realisation is shared across both worlds, the resulting joint law (the coupling) is determined entirely by how the exogenous noise is defined and abducted, which is exactly the modelling choice this paper studies for each variable type.

- **Binary.** A binary outcome needs only a single scalar noise $U_r \sim \mathrm{Uniform}(0,1)$, thresholded against the class probability. This is the unique coupling consistent with two Bernoulli marginals under the natural monotonicity requirement that raising the class probability cannot lower $P(X_r = 1)$.
- **Categorical.** For $C \ge 3$ unordered classes, one natural noise is a vector of C independent Gumbel perturbations, one per class, abducted jointly from which class won the factual "race" and reused for a second race against the counterfactual class scores. This is the *Gumbel-max coupling*: it treats every pair of classes symmetrically, and is the coupling characterised by counterfactual stability (Oberst & Sontag, 2019).
- **Ordinal.** When the classes carry an order, a different noise construction is more natural: a single shared Gaussian N_r, pushed through the factual and counterfactual cutpoints of the same monotone threshold function. This is the *comonotone* coupling: since the same monotone map is applied to the same noise draw in both worlds, it respects the class ordering in a way the Gumbel-max coupling does not.

The categorical and ordinal constructions therefore abduct genuinely different exogenous noise objects (a vector of per-class Gumbels versus a single Gaussian), and so induce different couplings, even when fit to variables with the same number of classes. The one case where this distinction disappears is $C = 2$: with only two classes, both constructions collapse to thresholding a single uniform, so the Gumbel-max and comonotone couplings coincide exactly, and the choice between "categorical" and "ordinal" machinery becomes immaterial.

3 RESULTS

Both propositions below instantiate the Compatibility/Abduction/Action-and-Prediction schema of §2.2; what differs is the noise object each abducts and how it is reused.

3.1 CATEGORICAL

Proposition 1 (Categorical Probabilistic Counterfactuals with Gaussian Processes)**.** *Let $\mathbf{f}(\mathbf{x})$ denote a realisation of the fitted GP latent at $\mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}$, and let $\mathbf{p}(\mathbf{x}; \mathbf{f}) := \big(p_1(\mathbf{x}; \mathbf{f}), \dots, p_C(\mathbf{x}; \mathbf{f})\big)$ be the normalised class-probability vector it induces. We fit C independent binary Laplace GP classifiers, one per class, and renormalise: $p_c := \sigma(f_c)/\sum_{c'} \sigma(f_{c'})$. The joint multiclass softmax fit, under which $\mathbf{p} = \mathrm{softmax}(\mathbf{f})$ instead, is developed in §A.6.1, and Remark 2 records what differs between the two; nothing in the construction below depends on the choice, because the coupling is stated on $\log \mathbf{p}$ rather than on $\mathbf{f}$. Write $\boldsymbol{\ell}(\mathbf{x}) := \log \mathbf{p}(\mathbf{x}; \mathbf{f})$ for the corresponding* log-probability scores*, suppressing $\mathbf{f}$ where no ambiguity arises. Model the structural equations as:*

$$X_r := \underset{c \in \{1,\dots,C\}}{\arg\max} \big\{ \ell_c(\mathbf{X}_{\mathrm{pa}(r)}) + G_{r,c} \big\}, \qquad G_{r,c} \overset{iid}{\sim} \mathrm{Gumbel}(0,1), \tag{1}$$

with $\mathbf{G}_r := (G_{r,1}, \dots, G_{r,C})$.

(Compatibility.) *For every $\mathbf{x}$ and every realisation of the latent,*

$$P\big(X_r = c \mid \mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x},\, \mathbf{f}\big) = \mathrm{softmax}\big(\log \mathbf{p}(\mathbf{x}; \mathbf{f})\big)_c = p_c(\mathbf{x}; \mathbf{f}), \tag{2}$$

the softmax identity required by the schema above.

(Abduction.) *Write $\boldsymbol{\ell}^F := \boldsymbol{\ell}(\mathbf{x}^F_{\mathrm{pa}(r)}; \mathbf{f}^F)$ and let $c^* := x_r^F$ be the observed class, and $M := \max_c\{\ell_c^F + G_{r,c}\}$ is the* factual winning score*. Conditional on $\boldsymbol{\ell}^F$ and on $X_r = c^*$, the posterior of $\mathbf{G}_r$ is sampled exactly by*

$$M \sim \mathrm{Gumbel}\Big(\log \textstyle\sum_c e^{\ell_c^F},\, 1\Big), \quad G_{r,c^*} := M - \ell^F_{c^*}, \quad G_{r,c} := -\log\big(e^{-(M-\ell_c^F)} - \log V_c\big),\ c \neq c^*, \tag{3}$$

with $V_c \overset{iid}{\sim} \mathrm{Uniform}(0,1)$.

(Action and Prediction.) *The intervention replaces $\mathbf{x}^F_{\mathrm{pa}(r)}$ by $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ leaving the latent process and $\mathbf{G}_r$ fixed, so with $\tilde{\boldsymbol{\ell}} := \boldsymbol{\ell}(\tilde{\mathbf{x}}_{\mathrm{pa}(r)})$,*

$$X_r^{\mathrm{CF}} = \underset{c}{\arg\max} \big\{ \tilde{\ell}_c + G_{r,c} \big\}. \tag{4}$$

As the schema requires, $(\mathbf{f}^F, \tilde{\mathbf{f}})$, from which $\boldsymbol{\ell}^F$ and $\tilde{\boldsymbol{\ell}}$ are formed, are drawn from their joint *posterior, which for each class c is the bivariate Gaussian with means $\mathbf{k}_F^\top(\mathbf{e}_{r,c} - \hat{\boldsymbol{\pi}}_{r,c})$ and $\tilde{\mathbf{k}}^\top(\mathbf{e}_{r,c} - \hat{\boldsymbol{\pi}}_{r,c})$ and covariance*

$$\mathrm{Cov}\big(f_c^F, \tilde{f}_c\big) = k_r\big(\mathbf{x}^F_{\mathrm{pa}(r)}, \tilde{\mathbf{x}}_{\mathrm{pa}(r)}\big) - \mathbf{k}_F^\top\big(\mathbf{K} + \hat{W}_c^{-1}\big)^{-1}\tilde{\mathbf{k}}, \qquad \mathbf{k}_F := \big(k_r(\mathbf{x}^F_{\mathrm{pa}(r)}, \mathbf{x}^i_{\mathrm{pa}(r)})\big)_{i=1}^n. \tag{5}$$

Under the one-versus-rest fit the C classifiers are independent, so the pairs $(f_c^F, \tilde{f}_c)$ are drawn independently across classes. Within a class, however, the pair must be a single *draw from equation 5: $f_{r,c}$ is one function evaluated at two inputs, and drawing the two marginally would discard the covariance and with it Corollary 1. The counterfactual class probabilities are*

$$P\big(X_r^{\mathrm{CF}} = c \mid \mathbf{x}^F\big) = \mathbb{E}_{(\mathbf{f}^F, \tilde{\mathbf{f}})}\, \mathbb{E}_{\mathbf{G}_r \mid x_r^F, \mathbf{f}^F}\Big[\mathbb{I}\big\{\underset{c'}{\arg\max}(\tilde{\ell}_{c'} + G_{r,c'}) = c\big\}\Big], \tag{6}$$

approximated by Monte Carlo over draws of $(\mathbf{f}^F, \tilde{\mathbf{f}})$ and, for each, a draw of $\mathbf{G}_r$ from equation 3.

Corollary 1 (Consistency under the trivial intervention)**.** *In the setting of Proposition 1, intervening at the factual parent value itself,* $\mathrm{do}\big(\mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)}\big)$*, recovers the observed factual outcome with certainty:*

$$P\big(X_r^{\mathrm{CF}} = x_r^F \mid \mathbf{x}^F,\ \mathrm{do}(\mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)})\big) = 1. \tag{7}$$

Corollary 2 (Binary case: closed form)**.** *Let* $C = 2$ *and put* $p(\mathbf{x}) := p_1(\mathbf{x}) = \sigma(f_r(\mathbf{x}))$ *with* $\sigma(t) = (1 + e^{-t})^{-1}$*. Then equation 1 is equivalent to*

$$X_r := \mathbb{I}\big[\, p(\mathbf{X}_{\mathrm{pa}(r)}) > U_r \,\big], \qquad U_r \sim \mathrm{Uniform}(0, 1), \tag{8}$$

the exogenous noise being $U_r = \sigma(G_{r,2} - G_{r,1})$*.*

Abduction is exact: writing $p^F := p(\mathbf{x}^F_{\mathrm{pa}(r)})$*,*

$$U_r \mid \mathbf{x}^F, f_r \ \sim\ \begin{cases} \mathrm{Uniform}(0,\, p^F), & x_r^F = 1, \\ \mathrm{Uniform}(p^F,\, 1), & x_r^F = 0, \end{cases} \tag{9}$$

and with $p^{\mathrm{CF}} := p(\tilde{\mathbf{x}}_{\mathrm{pa}(r)})$*,*

$$P\big(X_r^{\mathrm{CF}} = 1 \mid \mathbf{x}^F, f_r\big) = \begin{cases} \min(p^{\mathrm{CF}},\, p^F) \,/\, p^F, & x_r^F = 1, \\ \max(0,\ p^{\mathrm{CF}} - p^F) \,/\, (1 - p^F), & x_r^F = 0. \end{cases} \tag{10}$$

Marginalising over the joint Laplace posterior of $(f^F, \tilde{f}) := \big(f_r(\mathbf{x}^F_{\mathrm{pa}(r)}), f_r(\tilde{\mathbf{x}}_{\mathrm{pa}(r)})\big)$ *gives*

$$P\big(X_r^{\mathrm{CF}} = 1 \mid \mathbf{x}^F\big) = \mathbb{E}_{(f^F, \tilde{f})}\Big[\, \textit{equation } 10 \textit{ evaluated at } p^F = \sigma(f^F),\ p^{\mathrm{CF}} = \sigma(\tilde{f}) \Big]. \tag{11}$$

3.2 Ordinal

The categorical approach assumes an *unordered* (nominal) set of classes, thus the Gumbel-max coupling of Proposition 1 treats every pair of classes symmetrically. We give the ordinal analogue here, built on the classical Gaussian-process ordinal-regression likelihood (Chu & Ghahramani, 2005). This is exactly the continuous ANM-GP approach of Karimi et al. (2020), observed through a coarsened, binned channel rather than directly.

Cutpoints as a learned mechanism parameter. If the cutpoints of a variable are unknown, class boundaries can be inferred from data on the same footing as the latent function f_r itself. Write $\theta = (\theta_1, \ldots, \theta_{C-1}) \in \mathbb{R}^{C-1}$ for an unconstrained parameterisation of the cutpoints,

$$b_1(\theta) := \theta_1, \qquad b_c(\theta) := b_{c-1}(\theta) + e^{\theta_c} \quad (c = 2, \ldots, C-1), \tag{12}$$

under which every $\theta \in \mathbb{R}^{C-1}$ maps to a strictly increasing tuple of cutpoints and every strictly increasing tuple arises this way, and place an independent Gaussian prior $\theta \sim p(\theta) := \prod_c \mathcal{N}(\theta_c; 0, \tau_c^2)$, a priori independent of f_r and of N_r.

Proposition 2 (Ordinal Probabilistic Counterfactuals with Gaussian Processes and Learned Cutpoints)**.** *Model the structural equation as*

$$\begin{aligned} Y_r &:= f_r(\mathbf{X}_{\mathrm{pa}(r)}) + N_r, \quad f_r \sim \mathcal{GP}(0, k_r), \quad N_r \sim \mathcal{N}(0,1), \quad \theta \sim p(\theta), \\ X_r &:= \sum_{c=1}^{C} c \cdot \mathbb{I}\big\{ b_{c-1}(\theta) \le Y_r < b_c(\theta) \big\}, \end{aligned} \tag{13}$$

with $b_0(\theta) := -\infty$*,* $b_C(\theta) := \infty$*, and* (f_r, N_r, θ) *mutually independent a priori.*

(Compatibility.) *Recall that for* $N \sim \mathcal{N}(0,1)$ *we have the CDF:* $\Phi(z) = P(N \le z)$*. For every* $\mathbf{x}$ *and every realisation of* (f_r, θ)*,*

$$P\big(X_r = c \mid \mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}, f_r, \theta\big) = \Phi\big(b_c(\theta) - f_r(\mathbf{x})\big) - \Phi\big(b_{c-1}(\theta) - f_r(\mathbf{x})\big), \tag{14}$$

the identity required by the schema above; nothing in it uses θ*'s non-randomness.*

(Abduction.) *Let* $c^* := x_r^F$ *and, for any realisation of* θ*, write* $(a^F(\theta), b^F(\theta)) := (b_{c^*-1}(\theta), b_{c^*}(\theta))$ *for the interval implied by the observed class, and* $f_r^F := f_r(\mathbf{x}^F_{\mathrm{pa}(r)})$*. Conditional on* θ*, on* f_r^F*, and on* $X_r = c^*$*, the posterior of* N_r *is exact:*

$$N_r \mid \mathbf{x}^F, f_r^F, \theta, X_r = c^* \ \sim\ \mathcal{N}(0,1) \textit{ truncated to } \big(a^F(\theta) - f_r^F,\ b^F(\theta) - f_r^F\big), \tag{15}$$

by exactly the fixed-cutpoint argument, since θ enters only through which interval is being conditioned on. Marginalising over the joint posterior of (f_r^F, θ), the noise's posterior is the continuous mixture of truncated normals obtained by averaging equation 15 over θ.

(Action and Prediction.) *The intervention replaces $\mathbf{x}^F_{\mathrm{pa}(r)}$ by $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$, leaving the latent process, θ, and N_r fixed, so with $\tilde{f}_r := f_r(\tilde{\mathbf{x}}_{\mathrm{pa}(r)})$,*

$$P\big(X_r^{\mathrm{CF}} = k \mid \mathbf{x}^F, f_r^F, \tilde{f}_r, \theta\big) = \frac{\big[\Phi\big(\min(\beta, \gamma_k)\big) - \Phi\big(\max(\alpha, \gamma_{k-1})\big)\big]_+}{\Phi(\beta) - \Phi(\alpha)}, \tag{16}$$

where $\alpha := a^F(\theta) - f_r^F$, $\beta := b^F(\theta) - f_r^F$, $\gamma_k := b_k(\theta) - \tilde{f}_r$ and $(\cdot)_+ := \max(0, \cdot)$. Since f_r is unknown, $(f_r^F, \tilde{f}_r)$ must be drawn from their joint *posterior given θ, approximated as Gaussian by the Laplace approximation to the ordinal likelihood equation 14 (Chu & Ghahramani, 2005), the single-latent analogue of the joint Laplace posterior in eq. equation 5, and θ is itself unknown, drawn from its own Laplace-Gaussian posterior (Lemma 6). The full estimator therefore marginalises over both:*

$$P\big(X_r^{\mathrm{CF}} = k \mid \mathbf{x}^F\big) = \mathbb{E}_\theta\Big[\mathbb{E}_{(f_r^F, \tilde{f}_r)\mid\theta}\big[\,equation\ 16\,\big]\Big], \tag{17}$$

approximated by nested Monte Carlo: draw $\theta^{(s)}$ from its Laplace-Gaussian posterior, refit the single-latent joint posterior of $(f_r^F, \tilde{f}_r)$ at the resulting cutpoints $\mathbf{b}(\theta^{(s)})$, draw from it, and average eq. equation 16 over both layers.

Corollary 3 (Fixed cutpoints)**.** *When the cutpoints $\mathbf{b} = \mathbf{b}(\theta_0)$ are known rather than inferred, this means $p(\theta) = \delta_{\theta_0}$ is a point mass and the outer expectation in equation 17 collapses to a single term at $\theta = \theta_0$. Then Proposition 2 reduces exactly to fitting a single-latent GP-ordinal model at the given, fixed $\mathbf{b}$: this is the case used throughout Chu and Ghahramani's original construction (Chu & Ghahramani, 2005) and, prior to this section, throughout this paper.*

Remark 1 (Which coupling each proposition selects). §2.2 sketches which coupling each proposition selects; here we make Proposition 2's precise. Fix a realisation of θ and write $H(y) := \sum_c c \cdot \mathbb{I}\{b_{c-1}(\theta) \le y < b_c(\theta)\}$, non-decreasing. Eq. equation 13 gives $X_r = H(f_r^F + N_r)$ and $X_r^{\mathrm{CF}} = H(\tilde{f}_r + N_r)$, both non-decreasing functions of the *same* scalar; with $U := \Phi(N_r)$ this realises (X_r, X_r^{CF}) as $\big(F_p^{-1}(U), F_q^{-1}(U)\big)$ for the factual and counterfactual class distributions p, q — the comonotone coupling (matching the monotonicity assumption used for identification in the binary (Pearl, 1999) and categorical (Lu et al., 2020) settings).

At $C = 2$ this coincides with the Gumbel-max coupling of Proposition 1: for Bernoulli marginals the comonotone coupling is the unique one, which is why Corollary 2 admits a closed form where the general case does not — but only as couplings. The link functions still differ (probit here, logit in Corollary 2), so the two propositions agree at $C = 2$ as couplings, not as fitted models, and they diverge for $C \ge 3$ (the Experiments section measures the cost). The identification above also holds only at a fixed θ: after marginalising over the joint posterior of $(f_r^F, \tilde{f}_r, \theta)$ in equation 17, the estimator is a mixture of comonotone couplings over varying marginals, which need not itself be comonotone.

4 Experiments

4.1 Setup

Two fully worked examples in Appendix A.2 (Examples 1 and 2, one by hand, one against a fitted GP) walk through Abduction, Action and Prediction on a single SCM whose *true* mechanism happens to be exactly the Gumbel-max form of eq. equation 1. To conduct a more thorough survey of our method, we use categorical and ordinal approaches (Algorithm 1 and Algorithm 2) to compute counterfactuals on varying causal graphs depicted in Figure 1. The true structural equations of the graphs are known to us, but for our experiments we fit a `sklearn` Gaussian Process classifier to synthetic data generated from the true SCMs. The SCMs vary in both structural complexity and in the datatype of the endogenous nodes. We deliberately test whether each method's accuracy is confined to the case it was derived for. Using the Binary SCM, we explore the claim, from Remark 1, that both the categorical and ordinal approaches are equivalent.

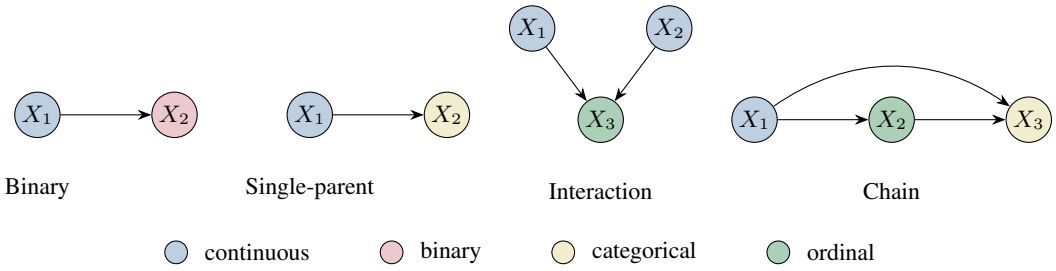


Figure 1: The four synthetic SCM structures used in Section 4 (Binary, Single-parent, Interaction and Chain), increasing in graph complexity and node type. Node fill color indicates each variable's data type.

These four structures can naturally occur in many systems, for example in the loan scenario (§1). The Binary SCM could be a bank's approve/deny decision from debt-to-income ratio; the Single-parent SCM, an applicant's routing to one of several loan products via an argmax over per-product suitability scores; the Interaction SCM, a risk tier from thresholding debt-to-income and loan-to-value non-additively; and the Chain SCM, the full pipeline, where debt-to-income sets the tier and both together determine which product is offered.

For each SCM we draw $n = 500$ training samples, fit the classifier(s) with no knowledge of the true structural equations, and evaluate on 8 factual/intervention scenarios.

During the GP fitting, we bound each RBF length scale to $[0.1\hat{\sigma}_d,\ 10\hat{\sigma}_d]$ rather than `sklearn`'s default $[10^{-5}, 10^{5}]$, whose ten-decade range causes most optimiser restarts to start far outside the data support, where the kernel is flat and fits get stuck. Under the default, the Single-parent SCM's baseline-class latent came out exactly constant ($P \approx 0.20$ vs. a true 0.05) and stayed there even with five times more restarts. The bounded fit attains a *higher* marginal likelihood, so this is a correction, not a modelling choice.

An ordinal cut-point mechanism. Our Interaction and Chain SCMs use the ordinal mechanism of Proposition 2 (eq. equation 13), with true cutpoints $b_1 = -0.5$, $b_2 = 0.5$. Ground truth is closed form: since the true score g and (b_1, b_2) are both known, eq. equation 16 is evaluated directly at $f_r^F = g(\mathbf{x}^F_{\mathrm{pa}(r)})$, $\tilde{f}_r = g(\tilde{\mathbf{x}}_{\mathrm{pa}(r)})$, with no Monte Carlo needed. Both fitted estimators below instead learn g from data via a GP classifier, as throughout, and differ only in whether they are also given the true cutpoints or must infer them (discussed further below).

Binary

$$X_1 := U_1,\ U_1 \sim \mathcal{N}(0,1),$$
$$X_2 := \mathbb{I}\big[2X_1{+}U_2{>}0\big],$$
$$U_2 \sim \mathcal{N}(0,1).$$

Single-parent

$$X_1 := U_1,\ U_1 \sim \mathcal{N}(0,1),$$
$$X_2 := \arg\max_{c\in\{1,2,3\}}\{\eta_c(X_1){+}G_c\},$$
$$\eta(x_1){=}(0, 1.2x_1, -x_1{+}0.4).$$

Interaction

$$X_1, X_2 \sim \mathcal{N}(0,1) \text{ iid},$$
$$Y_3 := X_1X_2{+}N_3,$$
$$N_3 \sim \mathcal{N}(0,1),$$
$$X_3 := 1{+}\mathbb{I}\{Y_3{>}b_1\}$$
$$+\mathbb{I}\{Y_3{>}b_2\}.$$

Chain (mixed mechanisms).

$$\begin{aligned} X_1 &:= U_1, \\ X_2 &:= 1{+}\mathbb{I}\{X_1{+}N_2{>}b_1\}{+}\mathbb{I}\{X_1{+}N_2{>}b_2\}, \\ &\quad N_2 \sim \mathcal{N}(0,1), \\ X_3 &:= \arg\max_c\{\tilde{\eta}_c(X_1,X_2) + G_{3,c}\}, \\ &\quad \tilde{\eta}(x_1,x_2) = (0,\ x_1{+}(x_2{-}2),\ -(x_1{+}(x_2{-}2))) \end{aligned}$$

Computing the ground-truth counterfactual for X_3 under $do(X_1 = \tilde{x}_1)$ chains two different per-node abductions through the mediator X_2: a truncated-normal posterior for X_2's noise (eq. equation 15) and a truncated-Gumbel posterior for X_3's (eq. equation 3).

For each experiment, we report two total-variation distances representing how well the *fitted structural equation* matches the true one ($\mathrm{TV}_{\mathrm{int}}$), and the counterfactual accuracy ($\mathrm{TV}_{\mathrm{cf}}$). $\mathrm{TV}_{\mathrm{int}}$ represents the distance between $\hat{P}(X_r \mid \mathbf{X}_{\mathrm{pa}(r)})$ and the true mechanism's own class probabilities,

Table 1: Counterfactual error $\mathrm{TV}_{\mathrm{cf}}$ and interventional error $\mathrm{TV}_{\mathrm{int}}$ across all four synthetic SCMs, under a categorical fit and under an ordinal fit with learned cutpoints; mean $\pm$ sd over 10 seeds at $n = 500$. Mediator applies only to the Chain SCM's X_3: whether the upstream mediator X_2 was fit as a categorical or ordinal node, isolating propagated error from X_3's own.

			Categorical fit		Ordinal fit	
SCM	Node type	Mediator	$\mathrm{TV}_{\mathrm{int}}$	$\mathrm{TV}_{\mathrm{cf}}$	$\mathrm{TV}_{\mathrm{int}}$	$\mathrm{TV}_{\mathrm{cf}}$
Binary	binary		0.021 ± 0.010	0.021 ± 0.014	$\mathbf{0.018 \pm 0.009}$	$\mathbf{0.015 \pm 0.007}$
Single-parent	categorical		$\mathbf{0.030 \pm 0.010}$	$\mathbf{0.031 \pm 0.015}$	0.184 ± 0.031	0.173 ± 0.033
Interaction	ordinal		0.078 ± 0.016	0.231 ± 0.100	$\mathbf{0.045 \pm 0.011}$	$\mathbf{0.068 \pm 0.040}$
learned cutpoints	ordinal		—	—	0.049 ± 0.010	0.070 ± 0.041
Chain, X_3	categorical	categorical	0.070 ± 0.012	0.091 ± 0.017	0.223 ± 0.021	0.221 ± 0.062
Chain, X_3	categorical	ordinal	$\mathbf{0.070 \pm 0.012}$	$\mathbf{0.044 \pm 0.020}$	0.223 ± 0.021	0.213 ± 0.057
Chain, X_2	ordinal		0.037 ± 0.010	0.274 ± 0.074	$\mathbf{0.027 \pm 0.007}$	$\mathbf{0.045 \pm 0.016}$
learned cutpoints	ordinal		—	—	0.029 ± 0.010	0.041 ± 0.024

averaged over the factual and the counterfactual parent configuration. This is the rung-two quantity that the Compatibility identities equation 2 and equation 14 guarantee any fitted model reproduces. Each entry is a mean $\pm$ standard deviation over 10 seeds at $n = 500$, so the spread reflects seed-to-seed variability across the whole pipeline (data, kernel fit and Monte Carlo draws together), not Monte Carlo noise alone. Appendix A.3 gives the full protocol and a Monte Carlo convergence check. Consistency error is exactly zero in every row, as Corollary 1 requires, whether or not the estimator's assumed mechanism matches the truth.

4.2 Findings

Misspecification is invisible to the fit — but only in one direction, and only partly correctable. Fitting the categorical estimator to data whose true mechanism is the ordinal cut-point model (the Interaction SCM, the Chain SCM's mediator) leaves $\mathrm{TV}_{\mathrm{int}}$ at the same order as the correctly specified rows, while $\mathrm{TV}_{\mathrm{cf}}$ is still off by roughly 0.25 (Table 1). Refitting these nodes with Proposition 2's ordinal estimator recovers most of that gap, by factors of 3.4 and 6.1, while $\mathrm{TV}_{\mathrm{int}}$ moves by far less. The Chain SCM's *terminal* node makes this sharpest: its own mechanism was never misspecified and its fitted equation is unchanged between the two arms, yet its counterfactual accuracy improves because it is fed a more accurate mediator counterfactual. What residual error remains once the coupling is right is ordinary GP estimation difficulty. The Interaction SCM's higher error against the Single-parent SCM's reflects the interaction term $x_1 x_2$ being a harder two-dimensional surface to learn from 500 samples than the Single-parent SCM's additive score. In the reverse direction, fitting the ordinal estimator to genuinely nominal Gumbel-max data (the Single-parent SCM, the Chain SCM's terminal node) degrades $\mathrm{TV}_{\mathrm{int}}$ itself, by a factor of six on the Single-parent SCM ($0.030 \rightarrow 0.184$), because forcing a monotone latent score onto classes with no true order is a harder regression problem outright, not merely the wrong label on a well-fit one. Imposing order costs the fit; imposing nominality does not.

The binary control. As §2.2 and Remark 1 predict for the case where the two couplings coincide, refitting the Binary SCM with the matching probit link barely moves $\mathrm{TV}_{\mathrm{cf}}$, well inside one standard deviation across seeds, in sharp contrast to the large drop that correcting a genuine mismatch produces at the Interaction SCM (Table 1).

Learned cutpoints cost almost nothing. The indented "learned cutpoints" rows infer $\mathbf{b}$ jointly with g (Proposition 2's general construction) instead of being given the true $b_1 = -0.5$, $b_2 = 0.5$ (Corollary 3's fixed-cutpoint case), yet the two are indistinguishable within seed noise. This is valuable for cases when a variable is known to be ordinal, but the cutpoints are not defined, such as the risk-tiers in the loan example.

More data does not fix the wrong coupling. Repeating the sweep at four training-set sizes, $n =$ 100 to 1000 (Table 2, Appendix A.4), shows $\mathrm{TV}_{\mathrm{int}}$ and $\mathrm{TV}_{\mathrm{cf}}$ falling together where the coupling is correct, as ordinary GP asymptotics predict, but diverging where it is wrong: $\mathrm{TV}_{\mathrm{int}}$ keeps improving while $\mathrm{TV}_{\mathrm{cf}}$ flattens onto a floor near 0.21–0.24, so their ratio grows with n under misspecification instead of staying near 1. The error is not a small-sample artefact more data would remove: the fit would pass any diagnostic applied to it, yet the counterfactual remains wrong by a fixed amount. The coupling family is a modelling decision that must be made on structural grounds, not one the observational fit can be trusted to reveal at any sample size.

Ablating the estimator confirms the coupling, not the GP, is responsible. Table 3 (Appendix A.4) separates abduction, coupling and posterior averaging on the two single-node SCMs: abduction is what does the work (skipping it costs an order of magnitude in accuracy on the Single-parent SCM), posterior averaging contributes almost nothing to the point estimate, and an oracle given the *true* class probabilities still recovers only about a tenth of the misspecification error at the Interaction SCM — roughly nine tenths of it is the coupling, not the GP.

5 Discussion

Propositions 1 and 2 agree on every observational and interventional quantity and disagree only about the counterfactual, by construction. This is the content of the two Compatibility identities, and Remark 1 identifies the two couplings responsible. §4 puts numbers on this gap and its direction-dependent visibility to the fit. Either way, the mechanism is a modelling choice made above the fitted model, not a parameter the fit can be relied on to recover.

This raises the question of how the choice should be made. Lorberbom et al. (2021) answer it by optimising a downstream objective, since no experiment can adjudicate the choice directly. Our results argue an objective alone is not enough: §4 shows a good observational and interventional fit does not imply a correct counterfactual, since mismatched couplings can agree on both while still diverging sharply downstream. We instead let the variable's type, known before any objective is specified, constrain the admissible couplings, and find this constraint the more consequential of the two: the gap between type-appropriate couplings dwarfs the gains reported from tuning within a single family.

Two limitations follow. First, our validation is against synthetic SCMs with known mechanisms; on real data the exact mechanism is not observable, though the gap is narrower than it sounds: choosing the coupling only requires knowing the variable's *type* — whether its categories carry an order — not its structural equation, and type is ordinary domain knowledge even when the mechanism is not. Second, the constraint is structural: Lorberbom et al. (2021) show no mechanism that abducts noise once and reuses it across interventions can be optimal simultaneously for every pair of interventional distributions, so neither proposition is uniformly best, each being the natural choice for its variable type. Proposition 2 already relaxes one candidate restriction by inferring cutpoints jointly with the latent function, with Corollary 3 recovering the fixed-cutpoint case exactly and costing almost nothing in practice (§4). Extending the construction to latent functions shared across correlated ordinal nodes remains a natural next step.

6 Conclusion

We introduced a unified probabilistic framework for counterfactual inference in GP-SCMs with heterogeneous variable types, deriving exact noise-abduction procedures for binary, categorical, and ordinal outcomes and proving that each reproduces its fitted model's observational and interventional distributions exactly. Across a suite of synthetic SCMs with known ground truth, we showed that the choice of coupling, not the quality of the underlying GP fit, is what determines counterfactual accuracy: two estimators that agree on every observational and interventional quantity can disagree by a total variation distance of roughly 0.25 on the counterfactual, an error that does not shrink with more data. This gap is invisible to goodness-of-fit when ordinal data is mistaken for categorical, but not in reverse: forcing a false order onto categorical data degrades the fit itself, so the two directions of misspecification are detected differently but corrected the same way. Because a variable's type constrains the admissible couplings before any data is seen, that structural knowledge, not a downstream fitting criterion, is what should guide the choice.

AI use statement

In this work, we used generative AI tools to help develop the theoretical framework and propositions, formulate and refine mathematical claims, provide critical ingredients for their proofs and assist in writing them, propose and refine hypotheses, design and give feedback on the experimental methodology, implement the counterfactual estimators and experiments, generate the synthetic SCMs used for evaluation, and interpret the resulting numbers. Assisting with translation, cleaning or reformatting a dataset, and supporting qualitative or thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools to draft and edit sections of the paper for readability and structure, create and edit the accompanying code, create and modify the paper's figures, search for and summarise related literature, and format references. We have reviewed all AI-assisted work: code was read and tested by the authors for correctness, and all AI-suggested claims, proofs, and related work were checked by the authors against the literature and re-derived or verified independently before inclusion. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

Our theoretical claims — the two Compatibility identities, the Consistency corollary, and the binary and ordinal special cases — are proved in full in Appendix A.6, with assumptions stated alongside each result and the Gaussian process background they build on recalled in Appendix A.1. All experiments use synthetic SCMs whose generating equations are given explicitly in §4 and Appendix A.2, so exact ground-truth counterfactuals are available in closed form rather than estimated; the full training and evaluation protocol (sample sizes, seeds, and Monte Carlo budgets) is specified in Appendix A.3, and a Monte Carlo convergence check is included there. Anonymized source code implementing both estimators, the full experiment suite, and the scripts that regenerate every table and figure in this paper is provided in the supplementary material.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Stefan Boes. Nonparametric analysis of treatment effects in ordered response models. *Empirical Economics*, 44(1):81–109, 2013.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

Emil Julius Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*, volume 33 of *National Bureau of Standards Applied Mathematics Series*. U.S. Government Printing Office, Washington, D.C., 1954.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NeurIPS)*, pp. 689–696, 2009.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 265–277, 2020.

Guy Lorberbom, Daniel D. Johnson, Chris J. Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized Gumbel-max causal mechanisms. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*, 2020.

David J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

Chris J. Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*, pp. 3086–3094, 2014.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*, 2017.

Richard D. McKelvey and William Zavoina. A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, 4(1):103–120, 1975.

Yang Ni and Bani Mallick. Ordinal causal discovery. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 4881–4890, 2019.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149, 1999.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

Ruicong Yao, Tim Verdonck, and Jakob Raymaekers. Causal discovery in mixed additive noise models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *PMLR*, 2025.

John I. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2): 109–144, 1977.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 804–813, 2011.

A Appendix

A.1 Definitions

Definition 1 (Gaussian Process Prior (Rasmussen & Williams, 2006))**.** A *Gaussian process* (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. A GP is completely specified by its *mean function* $m(\mathbf{x})$ and *covariance function* $k(\mathbf{x}, \mathbf{x}')$, defined as:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{18}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\big[(f(\mathbf{x}) - m(\mathbf{x}))\,(f(\mathbf{x}') - m(\mathbf{x}'))\big]. \tag{19}$$

We denote the GP prior over functions as

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}),\ k(\mathbf{x}, \mathbf{x}')\big). \tag{20}$$

For notational convenience the mean function is often taken to be zero.

Definition 2 (Gaussian Process Posterior (Rasmussen & Williams, 2006))**.** Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be training observations with $y_i = f(\mathbf{x}_i) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$, and let $f \sim \mathcal{GP}(0,\, k(\mathbf{x}, \mathbf{x}'))$. The joint prior over training targets $\mathbf{y}$ and test function values $\mathbf{f}_*$ at inputs X_* is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0},\ \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right). \tag{21}$$

Conditioning on the observations yields the GP posterior predictive distribution:

$$\mathbf{f}_* \mid X, \mathbf{y}, X_* \;\sim\; \mathcal{N}\big(\bar{\mathbf{f}}_*,\; \mathrm{Cov}(\mathbf{f}_*)\big), \tag{22}$$

$$\bar{\mathbf{f}}_* \;\triangleq\; K(X_*, X)\big[K(X, X) + \sigma_n^2 I\big]^{-1}\mathbf{y}, \tag{23}$$

$$\mathrm{Cov}(\mathbf{f}_*) \;=\; K(X_*, X_*) - K(X_*, X)\big[K(X, X) + \sigma_n^2 I\big]^{-1}K(X, X_*). \tag{24}$$

For a single test point $\mathbf{x}_*$ with $\mathbf{k}_* \triangleq K(X, \mathbf{x}_*)$, these reduce to:

$$\bar{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1}\mathbf{y}, \qquad \mathrm{Var}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1}\mathbf{k}_*. \tag{25}$$

A.2 Worked Numerical Example

Example 1 (Worked numerical example of the three-step procedure). We walk through Abduction, Action and Prediction for a single categorical node X_r with $C = 3$ classes $\{1, 2, 3\}$, treating the fitted GP latent as known exactly at the factual and counterfactual parent values (the *Compatibility* case of Proposition 1), so that $\boldsymbol{\ell}^F$ and $\tilde{\boldsymbol{\ell}}$ below are fixed score vectors rather than posterior draws.

Setup. Let X_r have a single scalar parent, so that $\mathbf{x}_{\mathrm{pa}(r)}$ is one number, and fix the intervention up front: the factual parent value is $\mathbf{x}^F_{\mathrm{pa}(r)} = 0.5$ and we intervene with $\mathrm{do}(\mathbf{X}_{\mathrm{pa}(r)} = \tilde{\mathbf{x}}_{\mathrm{pa}(r)})$ at $\tilde{\mathbf{x}}_{\mathrm{pa}(r)} = 2.5$. The intervention is on the *parent*, not on X_r itself; the class taken by X_r is what the three steps below compute. Both score vectors are *outputs* of the already-fitted classifier, read off at these two inputs through $\boldsymbol{\ell}(\mathbf{x}) = \log \mathbf{p}(\mathbf{x}; \mathbf{f})$, and neither is chosen at counterfactual time.

At the factual input the classifier gives $\mathbf{p}^F := \mathbf{p}(\mathbf{x}^F_{\mathrm{pa}(r)}; \mathbf{f}) = (0.7,\, 0.2,\, 0.1)$, so $\boldsymbol{\ell}^F = \log \mathbf{p}^F = (-0.3567,\, -1.6094,\, -2.3026)$, and the observed factual outcome is $x_r^F = c^* = 1$. At the intervened input it gives $\tilde{\mathbf{p}} := \mathbf{p}(\tilde{\mathbf{x}}_{\mathrm{pa}(r)}; \mathbf{f}) = (0.2,\, 0.3,\, 0.5)$, i.e. $\tilde{\boldsymbol{\ell}} = (-1.6094,\, -1.2040,\, -0.6931)$ — moving the parent from 0.5 to 2.5 makes class 3, not class 1, the most probable outcome.

1. Abduction (eq. equation 3). Since $\mathbf{p}^F$ is normalised, $\mu := \log \sum_c e^{\ell_c^F} = 0$. Suppose the sampler draws $V_M = 0.6$ for the factual winning score M:

$$M = \mu - \log(-\log V_M) = -\log(-\log 0.6) = 0.6717.$$

This fixes $G_{r,c^*} = M - \ell^F_{c^*} = 0.6717 - (-0.3567) = 1.0284$. For the two losing classes, drawing $V_2 = 0.5$ and $V_3 = 0.9$ and writing $b_c := M - \ell_c^F$,

$$b_2 = 2.2812, \quad G_{r,2} = -\log\big(e^{-b_2} - \log V_2\big) = -\log(0.1022 + 0.6931) = 0.2290,$$
$$b_3 = 2.9743, \quad G_{r,3} = -\log\big(e^{-b_3} - \log V_3\big) = -\log(0.0511 + 0.1054) = 1.8551.$$

The abducted noise vector is $\mathbf{G}_r = (1.0284,\, 0.2290,\, 1.8551)$, held fixed for the remaining steps.

2. Action. The intervention replaces $\boldsymbol{\ell}(\mathbf{X}_{\mathrm{pa}(r)})$ by $\tilde{\boldsymbol{\ell}}$ in the structural equation equation 1, giving the modified equation $X_r^{\mathrm{CF}} := \arg\max_c\{\tilde{\ell}_c + G_{r,c}\}$ with $\mathbf{G}_r$ unchanged.

3. Prediction (eq. equation 4). Adding the fixed noise to the post-intervention scores,

$$\tilde{\ell}_c + G_{r,c} = (-0.5810,\; -0.9750,\; \mathbf{1.1619}), \qquad c = 1, 2, 3,$$

so $X_r^{\mathrm{CF}} = 3$: although class 1 was the observed factual outcome and class 3 was the *least* likely class factually, the noise realisation abducted from the factual observation, carried through to the new scores under the intervention, makes class 3 the counterfactual outcome. As a check, intervening at the factual value itself, $\mathrm{do}(\mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)})$, replaces $\tilde{\boldsymbol{\ell}}$ by $\boldsymbol{\ell}^F$ and recovers $X_r^{\mathrm{CF}} = c^* = 1$ for every draw, as required by consistency (Corollary 1, eq. equation 2).

Example 2 (Empirical validation against a known ground truth). Example 1 fixes the scores $\boldsymbol{\ell}^F, \tilde{\boldsymbol{\ell}}$ by hand. Here we instead generate data from a known SCM, fit a genuine Gaussian Process classifier to it, run the full estimator of Proposition 1 (including the joint-latent Monte Carlo average of eq. equation 6, not the point-estimate simplification of Example 1), and compare the result against the exact counterfactual computed from the known structural equations.

The SCM. One continuous parent and one 3-way categorical child, the simplest instance of Proposition 1 that is not binary:

$$X_1 := U_1, \quad U_1 \sim \mathcal{N}(0,1),$$
$$X_2 := \underset{c\in\{1,2,3\}}{\arg\max}\{\eta_c(X_1) + G_c\}, \quad G_c \overset{iid}{\sim} \text{Gumbel}(0,1),$$

with $\boldsymbol{\eta}(x_1) = (0,\, x_1,\, -x_1)$: positive x_1 favours class 2, negative x_1 favours class 3, and class 1 is a constant baseline.

Pipeline. We draw $n = 400$ samples (x_1^i, x_2^i) from this SCM and fit an `sklearn GaussianProcessClassifier` with an RBF kernel to $X_1 \to X_2$. For the factual instance $x_1^F = 1.7641$, $x_2^F = 2$ and the intervention $do(X_1 = -1.7641)$, the fitted model's counterfactual class distribution (Monte Carlo over $4{,}000$ draws of the joint latent and the abducted noise) is compared against the exact ground-truth distribution obtained by abducting $\mathbf{G}$ from the *known* $\boldsymbol{\eta}$ (the same identity as eq. equation 3, applied to $\boldsymbol{\eta}$ instead of a fitted $\boldsymbol{\ell}$):

The GP estimate evaluates the counterfactual as:

$$\hat{P}(X_2^{\text{CF}} = c|do(X_1 = -1.7641)) = (0.0635, 0.0515, \mathbf{0.8850})$$

While the ground truth is:

$$P(X_2^{\text{CF}} = c|do(X_1 = -1.7641)) = (0.0774, 0.0293, \mathbf{0.8933})$$

Both agree that the counterfactual class is 3 and the total variation distance between the fitted estimate and the ground truth is only 0.022. As a consistency check, re-running the same fitted estimator with the trivial intervention $do(X_1 = x_1^F)$ places probability 1.0 on class 2, confirming eq. equation 2 holds empirically for the fitted model as well as for the true one.

A.3 Experimental Protocol

Each entry in Table 1 is a mean $\pm$ standard deviation over 10 random seeds at $n = 500$: each seed redraws the training data, refits every Gaussian Process, and re-runs all 8 factual/intervention scenarios, so the reported spread reflects sampling variability across the whole pipeline (data, kernel fit and Monte Carlo draws together). For example, the mismatched arm of the Interaction SCM ranges from 0.11 to 0.41 across the ten seeds.

Every counterfactual probability is itself a Monte Carlo average over $4{,}000$ draws of the joint latent and the abducted noise. This budget is comfortably converged: re-running the Interaction SCM at n_{mc} from 250 to $16{,}000$ moves TV_{cf} by only $+0.012$ between the first two budgets and by at most 0.0015 beyond $2{,}000$, two orders of magnitude below the seed-to-seed standard deviation of 0.100. Monte Carlo error is therefore not what the reported spread is measuring.

A.4 Additional Ablations

More data does not fix the wrong coupling. Table 2 repeats the sweep at four training-set sizes, and separates the two kinds of error cleanly. Where the coupling is correct, both quantities fall with n as ordinary GP asymptotics predict: the Single-parent SCM goes from 0.061 to 0.021 between $n = 100$ and $n = 1000$, the Chain SCM's refitted mediator from 0.055 to 0.022. Where the coupling is wrong, the fit keeps converging while the counterfactual does not. The Interaction SCM's TV_{int} falls by 61% over that range ($0.144 \to 0.056$) and the Chain SCM's mismatched mediator by 74% ($0.091 \to 0.024$), but their TV_{cf} barely moves, $0.254 \to 0.207$ and $0.279 \to 0.243$, flattening onto a floor near 0.21–0.24.

The ratio $\text{TV}_{\text{cf}}/\text{TV}_{\text{int}}$ therefore *grows* with sample size when the coupling is misspecified (from 1.8 to 3.7 on the Interaction SCM, and from 3.1 to 10.2 on the Chain SCM's mediator), while staying near 1 in every correctly specified row. This is the sharpest form of the point. The error is not a small-sample artefact that more data would remove: the structural equation converges to the truth, the goodness of fit improves monotonically and would reassure any diagnostic applied to it, and the counterfactual remains wrong by a fixed amount. The coupling family is a modelling decision that must be made on structural grounds; it is not one the observational fit can be trusted to reveal, at any sample size.

Table 2: Counterfactual error $\mathrm{TV}_{\mathrm{cf}}$ against training-set size, mean over 10 seeds. Rows marked † use an estimator whose coupling does not match the true mechanism; their error flattens rather than falling with n.

arm	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Binary (probit)	0.040	0.022	0.021	0.015
Single-parent (Gumbel-max)	0.061	0.045	0.031	0.021
Interaction (ordinal) †	0.253	0.226	0.231	0.207
Chain, X_3 (Gumbel-max)	0.119	0.084	0.091	0.071
mediator X_2 (ordinal) †	0.279	0.249	0.274	0.242
Binary, refit	0.039	0.021	0.014	0.013
Interaction, refit	0.083	0.094	0.068	0.055
Chain, refit, X_3	0.081	0.056	0.044	0.024
mediator X_2	0.055	0.036	0.045	0.022

What each ingredient contributes. The table above varies the SCM and holds the estimator fixed; Table 3 does the reverse, on the two single-node SCMs, Single-parent and Interaction, and separates the three things the estimator does: abduct an exogenous noise, couple it through the intervention, and average over the GP posterior.

Abduction is what does the work. Evaluating the fitted classifier at the intervened parents and ignoring the factual outcome entirely (the "resample independently" baseline) gives $\mathrm{TV}_{\mathrm{cf}} = 0.295$ on the Single-parent SCM, an order of magnitude worse than the 0.031 of the full estimator. Conditioning the latent on the factual outcome without abducting an exogenous noise scarcely helps (0.300): it lands on top of the interventional baseline, which is the behaviour one expects, since a latent shared with the factual world constrains the *probabilities* but leaves the realised outcome free to be redrawn.

Averaging over the GP posterior, by contrast, contributes almost nothing to the point estimate: replacing the draws by the posterior *mean* latent leaves $\mathrm{TV}_{\mathrm{cf}}$ unchanged to three decimals on both SCMs (0.031 against 0.031, 0.232 against 0.231). We report this as a negative result about our own construction. It does not license dropping the joint draw: the pair $(f_c^F, \tilde{f}_c)$ must still come from the joint posterior, since it is the shared latent that makes Corollary 1 hold exactly, and a plug-in mean discards the posterior spread that any credible interval on the counterfactual would need. What the comparison shows is that marginalising buys calibration, not accuracy, at these sample sizes.

The oracle isolates the coupling. The last row replaces the fitted class probabilities with the true ones, leaving everything else alone, so any error it still makes is attributable to the coupling and to nothing else. On the Single-parent SCM, whose true mechanism is Gumbel-max, the oracle error falls to 0.005 ± 0.001, the Monte Carlo floor at $4{,}000$ draws, which is the check that the oracle is doing what it claims. On the Interaction SCM it falls only from 0.231 to 0.205. Exact probabilities recover about a tenth of the misspecification error and leave the rest untouched: roughly nine tenths of it is the coupling, not the Gaussian Process. Together with Table 2 this closes the argument from both directions: neither more data nor perfect probabilities repairs the wrong coupling.

Table 3: Counterfactual error $\mathrm{TV}_{\mathrm{cf}}$ for five estimators on the same data, mean $\pm$ sd over 10 seeds at $n = 500$. The Single-parent SCM's true mechanism is Gumbel-max, matching the coupling; the Interaction SCM's is the ordinal cut-point model, so the coupling is misspecified.

estimator	Single-parent	Interaction
interventional (no abduction)	0.295 ± 0.079	0.434 ± 0.051
latent-only abduction	0.300 ± 0.080	0.430 ± 0.051
plug-in posterior mean	0.031 ± 0.014	0.232 ± 0.102
full estimator (ours)	0.031 ± 0.015	0.231 ± 0.100
oracle: true $\mathbf{p}$, same coupling	0.005 ± 0.001	0.205 ± 0.092

A.5 Algorithms

Algorithms 1 and 2 give the two estimators in the form they are implemented. Both follow the same skeleton — fit, compute the joint posterior moments once, then repeat abduction, action and prediction inside a Monte Carlo loop that marginalises over the uncertainty in the latent — and differ only in the exogenous noise and how it is abducted.

Algorithm 1 Counterfactual for a categorical node (Proposition 1)

Require: sample $\{(\mathbf{x}^i_{\mathrm{pa}(r)}, x^i_r)\}_{i=1}^n$; factual $(\mathbf{x}^F_{\mathrm{pa}(r)}, x^F_r = c^*)$; intervention $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$; draws S
Ensure: $\hat{P}(X^{\mathrm{CF}}_r = c \mid \mathbf{x}^F)$ for $c = 1, \dots, C$

Training
for $c = 1, \dots, C$ **do**
  fit a binary Laplace GP classifier for class c against the rest ▷ one-versus-rest; Remark 2
end for
Posterior moments — independent of the draw, so computed once
for $c = 1, \dots, C$ **do**
  $(\mu^F_c, s^{F2}_c, \tilde{\mu}_c, \tilde{s}^2_c, \rho_c) \leftarrow$ joint two-point Laplace moments at $(\mathbf{x}^F_{\mathrm{pa}(r)}, \tilde{\mathbf{x}}_{\mathrm{pa}(r)})$, eq. equation 5
end for
$\mathbf{n} \leftarrow \mathbf{0} \in \mathbb{R}^C$
for $s = 1, \dots, S$ **do** ▷ marginalise over the GP posterior
  for $c = 1, \dots, C$ **do**
    draw $(f^F_c, \tilde{f}_c)$ jointly from the bivariate Gaussian above ▷ one draw, both inputs; classes independent
  end for
  $\boldsymbol{\ell}^F \leftarrow \log \mathrm{norm}\big(\sigma(\mathbf{f}^F)\big), \quad \tilde{\boldsymbol{\ell}} \leftarrow \log \mathrm{norm}\big(\sigma(\tilde{\mathbf{f}})\big)$ ▷ coupling lives on $\log \mathbf{p}$
  Abduction — eq. equation 3
  $V \sim \mathrm{Uniform}(0, 1); \quad M \leftarrow \log \sum_c e^{\ell^F_c} - \log(-\log V)$ ▷ the factual winning score
  $b_c \leftarrow M - \ell^F_c; \quad E_c \overset{iid}{\sim} \mathrm{Exp}(1)$
  $G_{c^*} \leftarrow b_{c^*}; \quad G_c \leftarrow -\,\mathrm{logaddexp}(-b_c, \log E_c)$ for $c \neq c^*$
  Action and prediction
  $k \leftarrow \arg\max_c\{\tilde{\ell}_c + G_c\}; \quad n_k \leftarrow n_k + 1$
end for
return $\mathbf{n}/S$

Algorithm 2 Counterfactual for an ordinal node with learned cutpoints (Proposition 2)

Require: sample; cutpoint prior $p(\theta)$ (eq. equation 12); factual $(\mathbf{x}^F_{\mathrm{pa}(r)}, x^F_r = c^*)$; intervention $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$; outer draws S_θ; inner draws S_f
Ensure: $\hat{P}(X^{\mathrm{CF}}_r = k \mid \mathbf{x}^F)$ for $k = 1, \dots, C$

Training
fit $\hat{\theta}, \Sigma_\theta$ by Laplace-approximating the profiled cutpoint posterior ▷ Lemma 6
$\mathbf{q} \leftarrow \mathbf{0} \in \mathbb{R}^C$
for $t = 1, \dots, S_\theta$ **do** ▷ marginalise over the cutpoint posterior
  draw $\theta^{(t)} \sim \mathcal{N}(\hat{\theta}, \Sigma_\theta)$; set $b_c \leftarrow b_c(\theta^{(t)})$ via eq. equation 12
  fit one latent GP under the ordinal-probit likelihood equation 14 at these cutpoints, by Laplace approximation ▷ Lemma 5
  $(\mu^F, s^{F2}, \tilde{\mu}, \tilde{s}^2, \rho) \leftarrow$ joint two-point moments at $(\mathbf{x}^F_{\mathrm{pa}(r)}, \tilde{\mathbf{x}}_{\mathrm{pa}(r)})$ under this fit
  for $s = 1, \dots, S_f$ **do** ▷ marginalise over the GP posterior
    draw $(f^F, \tilde{f})$ jointly from the bivariate Gaussian above
    Abduction — the factual class confines N_r to one interval, eq. equation 15
    $\alpha \leftarrow b_{c^*-1} - f^F; \quad \beta \leftarrow b_{c^*} - f^F$
    Action and prediction — eq. equation 16; N_r is integrated out in closed form rather than sampled
    for $k = 1, \dots, C$ **do**
      $q_k \leftarrow q_k + \big[\Phi(\min(\beta,\, b_k - \tilde{f})) - \Phi(\max(\alpha,\, b_{k-1} - \tilde{f}))\big]_+ \big/ \big(\Phi(\beta) - \Phi(\alpha)\big) \big/ S_f$
    end for
  end for
end for
return $\mathbf{q}/S_\theta$

Fixing $p(\theta)$ to a point mass at some θ_0 (so $S_\theta = 1$ and $\theta^{(1)} = \theta_0$ deterministically) recovers the fixed-cutpoint algorithm of Corollary 3 exactly, the Training step reducing to a single Laplace fit at the given $\mathbf{b}(\theta_0)$.

Two implementation details are worth stating, because both are easy to get wrong and neither is visible in the equations. First, the pair $(f_c^F, \tilde{f}_c)$ on line 9 of Algorithm 1 must be a *single* draw from the joint posterior; drawing the two marginally discards the covariance equation 5 and destroys the exact consistency of Corollary 1. When a whole response curve is traced over a grid of interventions, the draws should additionally share one random-number stream across grid points, for the reason given in Remark 3. Second, the abduction on line 15 is written with $\mathrm{logaddexp}$ rather than as $-\log(e^{-b_c} - \log V_c)$: the two are identical in exact arithmetic, but the first is a sum inside the logarithm and so neither cancels nor overflows, whereas the naive form loses precision exactly for classes that were probable factually yet lost the $\arg\max$ — the individuals whose counterfactuals are most interesting.

For a multi-node SCM the two algorithms are applied per node in topological order, each node abducting its own exogenous noise from its own factual parents and outcome, and the sampled counterfactual of a parent feeding the next node's intervened input.

A.6 Proofs

Throughout, $\{\mathbf{x}^i\}_{i=1}^n$ denotes the observational sample, $\mathbf{x}^F$ the factual individual, and where $\mathbf{x}^F \in \{\mathbf{x}^i\}_{i=1}^n$ we write $F \in [n]$ for its index. We write $\mathbf{e}_F \in \mathbb{R}^n$ for the F-th standard basis vector, $\mathbf{1}_n$ for the all-ones vector, and reserve $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ for the counterfactual parent configuration. All statements are made for a fixed $r \in [d]$ with non-empty parent set $|\mathrm{pa}(r)| > 0$; the subscript r is suppressed wherever no ambiguity arises.

For the discrete results of §A.6.2, $\boldsymbol{\ell} = (\ell_1, \dots, \ell_C)$ denotes a vector of *log-probability scores* entering a $\arg\max$ over classes, and M denotes the value of that maximum. We deliberately avoid a and M for these objects: in the recourse literature (Karimi et al., 2020) $a = \mathrm{do}(\mathbf{X}_\mathcal{I} = \boldsymbol{\theta}) \in \mathbb{A}^F$ is an *action* and $\mathcal{M}$ is the SCM itself, and neither is what is meant here. The Gumbel vector $\mathbf{G}_r$ plays the role of the exogenous U_r of that paper's equation (1); we retain G because U is used in Corollary 2 for the induced scalar uniform noise of the binary mechanism.

A.6.1 Laplace machinery for the multiclass GPC

Notation and ordering conventions. For $C > 2$ the latent object is the nC-vector collecting C latent values per training point, and two orderings are in circulation. Write

$$\mathbf{f} := \left(\mathbf{f}_{r,1}^\top, \dots, \mathbf{f}_{r,C}^\top\right)^\top \in \mathbb{R}^{nC} \quad \text{(class-ordered)}, \qquad \mathbf{f}^\pi := P\mathbf{f} \quad \text{(point-ordered)}, \tag{26}$$

where $\mathbf{f}_{r,c} := \left(f_{r,c}(\mathbf{x}^1_{\mathrm{pa}(r)}), \dots, f_{r,c}(\mathbf{x}^n_{\mathrm{pa}(r)})\right)^\top$ and P is the permutation sending index $(c-1)n+i$ to $(i-1)C+c$. Since $f_{r,1}, \dots, f_{r,C}$ are independent GPs sharing the kernel k_r, the prior covariance of $\mathbf{f}$ is

$$\boldsymbol{\mathcal{K}}_r := \mathbf{I}_C \otimes \mathbf{K}, \qquad \mathbf{K} := \left(k_r(\mathbf{x}^i_{\mathrm{pa}(r)}, \mathbf{x}^j_{\mathrm{pa}(r)})\right)_{ij}, \tag{27}$$

Let $\mathbf{e}_{r,c} \in \{0,1\}^n$ have i-th entry $\mathbb{I}\{x_r^i = c\}$, let $\mathbf{e} \in \mathbb{R}^{nC}$ be the class-ordered stack of the $\mathbf{e}_{r,c}$, and let $\boldsymbol{\pi} \in \mathbb{R}^{nC}$ be the corresponding stack of $\boldsymbol{\pi}_{r,c}$, where $\boldsymbol{\pi}_i(\mathbf{f}) = \mathrm{softmax}(\mathbf{f}_r(\mathbf{x}^i_{\mathrm{pa}(r)})) \in \mathbb{R}^C$, the softmax reduction noted in Proposition 1's preamble below.

The following lemma is the machinery behind the multiclass Laplace posterior used throughout Proposition 1 below, whose joint covariance is displayed there as eq. equation 5; this general machinery is not itself exposed as a separate main-text equation, so it is derived here in full, and used directly in the proof of Proposition 1 in §A.6.2.

Lemma 1 (Strict concavity, the MAP equation, and the Laplace predictive)**.** *Let $\Psi(\mathbf{f}) := \log p(\mathbf{x}_r \mid \mathbf{f}) + \log p(\mathbf{f} \mid \mathbf{X}_{\mathrm{pa}(r)})$ be the unnormalised log posterior of the multiclass GPC model, i.e. of the bare softmax likelihood $P(X_r = c \mid \mathbf{f}) = \mathrm{softmax}(\mathbf{f}_r(\mathbf{x}))_c$* without *the Gumbel-max coupling of Proposition 1's eq. equation 1, with $\mathbf{K} \succ 0$. Then:*

(i) $\nabla\Psi(\mathbf{f}) = (\mathbf{e} - \boldsymbol{\pi}) - \boldsymbol{\mathcal{K}}_r^{-1}\mathbf{f}$ and $\nabla^2\Psi(\mathbf{f}) = -W(\mathbf{f}) - \boldsymbol{\mathcal{K}}_r^{-1}$, where $W(\mathbf{f})$ is the $nC \times nC$ matrix with $P\,W(\mathbf{f})\,P^\top = \mathrm{blockdiag}\big(\Lambda_1(\mathbf{f}), \dots, \Lambda_n(\mathbf{f})\big)$ and $\Lambda_i(\mathbf{f}) := \mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^\top$ (derived below);

(ii) $W(\mathbf{f}) \succeq 0$*, with null space of dimension* n*;*

(iii) Ψ *is strictly concave, hence has a unique maximiser* $\hat{\mathbf{f}}_r$*, characterised for each* $c \in \{1, \dots, C\}$ *by*

$$\hat{\mathbf{f}}_{r,c} = \mathbf{K}\big(\mathbf{e}_{r,c} - \hat{\boldsymbol{\pi}}_{r,c}\big), \qquad \hat{\boldsymbol{\pi}}_{r,c} := \boldsymbol{\pi}_{r,c}(\hat{\mathbf{f}}_r); \tag{28}$$

(iv) the Laplace approximation to the latent posterior is $q(\mathbf{f}_r \mid \cdot) = \mathcal{N}\big(\hat{\mathbf{f}}_r, (\boldsymbol{\mathcal{K}}_r^{-1} + \hat{W})^{-1}\big)$ *(the general multiclass Laplace posterior, of which the joint covariance needed in Proposition 1's eq. equation 5 is a two-point specialisation), and the induced predictive law at any input* $\mathbf{x}^\star$ *has, per class,*

$$\mathbb{E}_q\big[f_{r,c}(\mathbf{x}^\star)\big] = \mathbf{k}_\star^\top\big(\mathbf{e}_{r,c} - \hat{\boldsymbol{\pi}}_{r,c}\big), \qquad \mathrm{Var}_q\big[f_{r,c}(\mathbf{x}^\star)\big] = k_\star - \mathbf{k}_\star^\top\big(\mathbf{K} + \hat{W}_c^{-1}\big)^{-1}\mathbf{k}_\star, \tag{29}$$

with $\mathbf{k}_\star := (k_r(\mathbf{x}^\star, \mathbf{x}^i_{\mathrm{pa}(r)}))_{i=1}^n$*,* $k_\star := k_r(\mathbf{x}^\star, \mathbf{x}^\star)$ *and* $\hat{W}_c := \mathrm{diag}(\hat{\pi}_{i,c}(1 - \hat{\pi}_{i,c}))_{i=1}^n$*. More generally, for two inputs* $\mathbf{x}^\star, \mathbf{x}^\dagger$*,*

$$\mathrm{Cov}_q\big[f_{r,c}(\mathbf{x}^\star), f_{r,c}(\mathbf{x}^\dagger)\big] = k_r(\mathbf{x}^\star, \mathbf{x}^\dagger) - \mathbf{k}_\star^\top\big(\mathbf{K} + \hat{W}_c^{-1}\big)^{-1}\mathbf{k}_\dagger. \tag{30}$$

Proof. (i) The bare softmax likelihood factorises over training points, so

$$\Psi(\mathbf{f}) = \sum_{i=1}^n \Big[f_{r,x_r^i}(\mathbf{x}^i_{\mathrm{pa}(r)}) - \log \textstyle\sum_{c=1}^C \exp f_{r,c}(\mathbf{x}^i_{\mathrm{pa}(r)})\Big] - \tfrac{1}{2}\mathbf{f}^\top \boldsymbol{\mathcal{K}}_r^{-1}\mathbf{f} + \text{const}, \tag{31}$$

the constant absorbing $-\frac{1}{2}\log|\boldsymbol{\mathcal{K}}_r| - \frac{nC}{2}\log 2\pi$. Differentiating the i-th bracket in $f_{r,c}(\mathbf{x}^i_{\mathrm{pa}(r)})$ gives $\mathbb{I}\{x_r^i = c\} - \pi_{i,c}$, which stacks to $\mathbf{e} - \boldsymbol{\pi}$; differentiating $-\frac{1}{2}\mathbf{f}^\top\boldsymbol{\mathcal{K}}_r^{-1}\mathbf{f}$ gives $-\boldsymbol{\mathcal{K}}_r^{-1}\mathbf{f}$. Differentiating once more, the second derivative of the i-th bracket with respect to $\big(f_{r,c}(\mathbf{x}^i_{\mathrm{pa}(r)})\big)_{c=1}^C$ is $-\big(\mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^\top\big) = -\Lambda_i(\mathbf{f})$, and mixed derivatives across distinct training points vanish because the likelihood factorises.

(ii) For $\mathbf{v} \in \mathbb{R}^{nC}$ with point-ordered blocks $\mathbf{v}_i \in \mathbb{R}^C$,

$$\mathbf{v}^\top W\mathbf{v} = \sum_{i=1}^n \mathbf{v}_i^\top\Lambda_i\mathbf{v}_i = \sum_{i=1}^n \Big[\sum_c \pi_{i,c}v_{i,c}^2 - \Big(\sum_c \pi_{i,c}v_{i,c}\Big)^2\Big] = \sum_{i=1}^n \mathrm{Var}_{c\sim\boldsymbol{\pi}_i}\big[v_{i,c}\big] \;\geq\; 0. \tag{32}$$

The variance in equation 32 vanishes iff $v_{i,c}$ is constant in c for each i (as $\pi_{i,c} > 0$ for all c), i.e. iff $\mathbf{v}_i \in \mathrm{span}\{\mathbf{1}_C\}$; the null space of W therefore has dimension exactly n. This degeneracy is the shift-invariance of the softmax, $\mathrm{softmax}(\mathbf{f}_i + a\mathbf{1}_C) = \mathrm{softmax}(\mathbf{f}_i)$.

(iii) By (i) and (ii), $\nabla^2\Psi = -W - \boldsymbol{\mathcal{K}}_r^{-1} \prec 0$ everywhere, since $W \succeq 0$ and $\boldsymbol{\mathcal{K}}_r^{-1} = \mathbf{I}_C \otimes \mathbf{K}^{-1} \succ 0$. Hence Ψ is strictly concave and, being coercive, attains a unique global maximum $\hat{\mathbf{f}}_r$. Note that it is the GP prior, not the likelihood, that removes the softmax degeneracy and makes the MAP well defined. Setting $\nabla\Psi = 0$ gives $\boldsymbol{\mathcal{K}}_r^{-1}\hat{\mathbf{f}}_r = \mathbf{e} - \hat{\boldsymbol{\pi}}$, i.e. $\hat{\mathbf{f}}_r = \boldsymbol{\mathcal{K}}_r(\mathbf{e} - \hat{\boldsymbol{\pi}})$. Since $\boldsymbol{\mathcal{K}}_r = \mathbf{I}_C \otimes \mathbf{K}$ is block diagonal in the class-ordered layout, this decouples across c into equation 28.

(iv) Since $\nabla\Psi(\hat{\mathbf{f}}_r) = 0$, the second-order Taylor expansion of Ψ at its mode is $\Psi(\mathbf{f}) = \Psi(\hat{\mathbf{f}}_r) - \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}}_r)^\top(\boldsymbol{\mathcal{K}}_r^{-1} + \hat{W})(\mathbf{f} - \hat{\mathbf{f}}_r) + O(\|\mathbf{f} - \hat{\mathbf{f}}_r\|^3)$; exponentiating and normalising the quadratic part gives the stated Gaussian (Williams & Barber, 1998; Rasmussen & Williams, 2006), whose covariance is well defined because $\boldsymbol{\mathcal{K}}_r^{-1} + \hat{W} \succ 0$ by (ii)–(iii) notwithstanding the singularity of $\hat{W}$. For the predictive moments, $f_{r,c}(\mathbf{x}^\star) \mid \mathbf{f}_{r,c} \sim \mathcal{N}(\mathbf{k}_\star^\top\mathbf{K}^{-1}\mathbf{f}_{r,c},\, k_\star - \mathbf{k}_\star^\top\mathbf{K}^{-1}\mathbf{k}_\star)$ by the GP prior; marginalising over q with the laws of total expectation and total variance and using equation 28 gives the mean in equation 29, while the variance follows on approximating the c-th diagonal block of $(\boldsymbol{\mathcal{K}}_r^{-1} + \hat{W})^{-1}$ by $(\mathbf{K}^{-1} + \hat{W}_c)^{-1}$ and applying the Woodbury identity $\mathbf{K}^{-1} - \mathbf{K}^{-1}(\mathbf{K}^{-1} + \hat{W}_c)^{-1}\mathbf{K}^{-1} = (\mathbf{K} + \hat{W}_c^{-1})^{-1}$, legitimate because $0 < \hat{\pi}_{i,c} < 1$ makes $\hat{W}_c$ invertible. Equation equation 30 is the same computation with $\mathbf{k}_\star$ and $\mathbf{k}_\dagger$ in the two slots. □

Remark 2 (One-versus-rest predictive moments). Lemma 1 concerns the joint multiclass Laplace approximation with a softmax likelihood. The implementation evaluated in §4 instead fits C independent binary Laplace classifiers and renormalises, and equation 29–equation 30 do not apply to it verbatim: there each class has its own bivariate Gaussian posterior over $(f_c^F, \tilde{f}_c)$, obtained from

that class's own binary Laplace fit, and the classes are independent. As established in the proof of Proposition 1, this changes *how* $\mathbf{p}$ is estimated but not the counterfactual construction, because the coupling is stated on $\log \mathbf{p}$. The compatibility identity, the abduction of Lemma 3 and Corollary 1 all go through unchanged; only the accuracy of $\mathbf{p}$ is affected — which is exactly what the $\mathrm{TV}_{\mathrm{int}}$ column of §4's table measures. At $C = 2$ the two approximations coincide and Corollary 2 applies as stated.

A.6.2 Proof of Proposition 1, Corollary 1 and Corollary 2

Lemma 2 (Gumbel-max representation (Gumbel, 1954; Yellott, 1977)). *Let $\boldsymbol{\ell} = (\ell_1, \dots, \ell_C) \in \mathbb{R}^C$ be any score vector and $G_1, \dots, G_C \overset{iid}{\sim} \mathrm{Gumbel}(0, 1)$. Then*

$$P\Big(\arg\max_c\{\ell_c + G_c\} = c^*\Big) = \mathrm{softmax}(\boldsymbol{\ell})_{c^*}, \qquad \max_c\{\ell_c + G_c\} \sim \mathrm{Gumbel}\Big(\log \textstyle\sum_c e^{\ell_c},\, 1\Big), \tag{33}$$

and the $\arg\max$ *and the* $\max$ *are independent.*

Proof. $\ell_c + G_c \sim \mathrm{Gumbel}(\ell_c, 1)$ with CDF $\exp(-e^{-(t-\ell_c)})$. By independence, $P(\max_c\{\ell_c + G_c\} \le t) = \prod_c \exp(-e^{-(t-\ell_c)}) = \exp\big(-e^{-t}\sum_c e^{\ell_c}\big) = \exp\big(-e^{-(t-\log\sum_c e^{\ell_c})}\big)$, which is the second claim. For the first, condition on G_{c^*} and integrate:

$$P\big(\arg\max = c^*\big) = \int_{\mathbb{R}} e^{-(g-\ell_{c^*})} e^{-e^{-(g-\ell_{c^*})}} \prod_{c \neq c^*} \exp\big(-e^{-(g-\ell_c)}\big)\, dg. \tag{34}$$

Substituting $z := e^{-g}$, $dz = -e^{-g}dg$, the integrand becomes $e^{\ell_{c^*}} \exp\big(-z\sum_c e^{\ell_c}\big)dz$ on $z \in (0, \infty)$, whence equation 34 equals $e^{\ell_{c^*}} / \sum_c e^{\ell_c}$. Independence of the $\arg\max$ and the $\max$ is the standard consequence of the same computation, the conditional law of the maximum given $\arg\max = c^*$ being free of c^*. □

Lemma 3 (Posterior of the Gumbel vector given the realised class (Maddison et al., 2014)). *In the setting of Lemma 2, condition on $\arg\max_c\{\ell_c + G_c\} = c^*$. Let $M := \max_c\{\ell_c + G_c\}$. Then $M \sim \mathrm{Gumbel}(\log\sum_c e^{\ell_c}, 1)$, and given M the coordinates are conditionally independent with*

$$G_{c^*} = M - \ell_{c^*}, \qquad G_c \mid M \;\sim\; \mathrm{Gumbel}(0, 1) \text{ conditioned on } G_c \le M - \ell_c, \quad c \neq c^*. \tag{35}$$

Consequently equation 3 samples the posterior exactly.

Proof. The event $\{\arg\max = c^*, M \in dm\}$ is $\{\ell_{c^*} + G_{c^*} \in dm\} \cap \bigcap_{c \neq c^*}\{\ell_c + G_c \le m\}$. Given $M = m$, the coordinate c^* is determined, $G_{c^*} = m - \ell_{c^*}$, and the remaining coordinates are independent (by independence of the G_c) each conditioned on the event $\{\ell_c + G_c \le m\}$, i.e. $\{G_c \le m - \ell_c\}$. That $M \sim \mathrm{Gumbel}(\log\sum_c e^{\ell_c}, 1)$ unconditionally, and independently of the $\arg\max$, is Lemma 2; hence its conditional law given $\arg\max = c^*$ is the same. For the sampler, if $G \sim \mathrm{Gumbel}(0, 1)$ conditioned on $G \le b$ then its CDF is $\exp(-e^{-g})/\exp(-e^{-b})$ for $g \le b$; setting this equal to $V \sim \mathrm{Uniform}(0, 1)$ and solving gives $e^{-g} = e^{-b} - \log V$, i.e. $g = -\log\big(e^{-b} - \log V\big)$, which with $b = m - \ell_c$ is the expression in equation 3. Note $e^{-b} - \log V > 0$ almost surely, so the logarithm is well defined. □

Proof of Proposition 1. Compatibility. Apply Lemma 2 with $\ell_c = \log p_c(\mathbf{x})$. For any probability vector $\mathbf{p}$,

$$\mathrm{softmax}(\log \mathbf{p})_c = \frac{p_c}{\sum_{c'} p_{c'}} = p_c, \tag{36}$$

so equation 33 gives equation 2 directly, for arbitrary $\mathbf{p}$. Since $\mathbf{G}_r$ is independent of the latent and of everything upstream, and $\mathrm{do}(\mathbf{X}_{\mathrm{pa}(r)} = \tilde{\mathbf{x}}_{\mathrm{pa}(r)})$ removes only the equations for $\mathbf{X}_{\mathrm{pa}(r)}$,

$$P\big(X_r = c \mid \mathrm{do}(\mathbf{X}_{\mathrm{pa}(r)} = \tilde{\mathbf{x}}_{\mathrm{pa}(r)})\big) = p_c(\tilde{\mathbf{x}}_{\mathrm{pa}(r)}), \tag{37}$$

so equation 1 matches the fitted model on all observational and interventional distributions.

The coupling must sit on $\log \mathbf{p}$, not on $\mathbf{f}_r$. Under the joint Laplace fit, $\mathrm{softmax}(\mathbf{f}_r)_c = p_c$, so

$$\log p_c = f_{r,c} - \log\textstyle\sum_{c'} e^{f_{r,c'}} = f_{r,c} - Z(\mathbf{x}) \tag{38}$$

with $Z(\mathbf{x}) := \log \sum_{c'} e^{f_{r,c'}}$ a shift invisible to $\arg\max$, so coupling on $\mathbf{f}_r$ or on $\log \mathbf{p}$ agree. Under a one-versus-rest fit, $p_c \propto \sigma(f_{r,c})$, so $\mathrm{softmax}(\mathbf{f}_r) \neq \mathbf{p}$ in general and coupling on $\mathbf{f}_r$ would violate equation 2; only $\log \mathbf{p}$ works for both.

Abduction. Equation equation 1 is a structural equation with $\mathbf{G}_r$ exogenous, unit-specific and independent across units, so Step 1 asks for $P_{\mathbf{G}_r|\mathbf{x}^F}$. Given $\boldsymbol{\ell}^F$, Lemma 3 with $\ell_c = \ell_c^F = \log p_c(\mathbf{x}^F_{\mathrm{pa}(r)})$, $c^* = x_r^F$ gives exactly equation 3. Since $\boldsymbol{\ell}^F = \log \mathbf{p}$ is normalised, $\sum_c e^{\ell_c^F} = 1$ and $M \sim \mathrm{Gumbel}(0,1)$ here; equation 3 retains the general location because it also holds for unnormalised scores.

Action and prediction. Fixing $f_r, \mathbf{G}_r$ and substituting the argument gives equation 4. As only the Laplace posterior of f_r is known, $(\mathbf{f}^F, \tilde{\mathbf{f}})$ is drawn jointly, with moments given by Lemma 1(iv) — equation 29, equation 30 at $\mathbf{x}^\star = \mathbf{x}^F_{\mathrm{pa}(r)}$, $\mathbf{x}^\dagger = \tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ — the covariance being equation 5. Taking expectation of the indicator in equation 4 over $\mathbf{G}_r \mid x_r^F, \mathbf{f}^F$ then $(\mathbf{f}^F, \tilde{\mathbf{f}})$,

$$P(X_r^{\mathrm{CF}} = c \mid \mathbf{x}^F) = \mathbb{E}_{(\mathbf{f}^F, \tilde{\mathbf{f}})}\, \mathbb{E}_{\mathbf{G}_r | x_r^F, \mathbf{f}^F}\big[\mathbb{I}\{\arg\max_{c'}\{\tilde{\ell}_{c'} + G_{r,c'}\} = c\}\big], \tag{39}$$

gives equation 6. The integrand is a bounded indicator, so the Monte Carlo estimator is unbiased and converges with probability 1 by the strong law of large numbers; each draw is a vertex of Δ^{C-1}, so the empirical average lies in Δ^{C-1} at every finite sample size and is a valid categorical distribution.

The pair $(\mathbf{f}^F, \tilde{\mathbf{f}})$ in abduction equation 35 and prediction equation 4 must be the *same* draw: f_r is one function evaluated at two inputs. Independent draws discard the covariance equation 5, severing the shared latent needed for Corollary 1. □

Proof of Corollary 1. Set $\tilde{\mathbf{x}}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)}$, so $\mathbf{k}_\star = \mathbf{k}_\dagger$ in equation 30. Then

$$\mathrm{Var}_q[f_c^F - \tilde{f}_c] = \mathrm{Var}_q[f_c^F] + \mathrm{Var}_q[\tilde{f}_c] - 2\mathrm{Cov}_q[f_c^F, \tilde{f}_c] = 0, \qquad \text{hence} \qquad \tilde{\mathbf{f}} = \mathbf{f}^F \ \text{ almost surely.} \tag{40}$$

Since $\boldsymbol{\ell} = \log \mathbf{p}$ is a deterministic function of $\mathbf{f}$, equation 40 gives

$$\tilde{\boldsymbol{\ell}} = \boldsymbol{\ell}^F \quad \text{almost surely.} \tag{41}$$

By Lemma 3, conditioning the abduction on $x_r^F = c^*$ gives, with $M = \max_c\{\ell_c^F + G_{r,c}\}$,

$$\ell^F_{c^*} + G_{r,c^*} = M, \qquad \ell_c^F + G_{r,c} < M \quad (c \neq c^*), \tag{42}$$

the inequality strict almost surely, since the conditioned Gumbel distribution is continuous and so assigns zero probability to any exact value — an exact tie between two classes would be one such value, so ties simply do not occur. Substituting equation 41 into equation 42, the comparison is unchanged under $\tilde{\boldsymbol{\ell}}$:

$$\tilde{\ell}_{c^*} + G_{r,c^*} = \ell^F_{c^*} + G_{r,c^*} = M \ > \ \ell_c^F + G_{r,c} = \tilde{\ell}_c + G_{r,c}, \qquad c \neq c^*, \tag{43}$$

so

$$\arg\max_c \big\{\tilde{\ell}_c + G_{r,c}\big\} = c^* = x_r^F \quad \text{almost surely.} \tag{44}$$

Since this indicator equals x_r^F almost surely, averaging it in equation 6 returns the point mass at x_r^F, which is equation 7.

Remark. Two points bear on the implementation. Equation equation 40 requires $\mathbf{f}^F$ and $\tilde{\mathbf{f}}$ to be the *same* posterior draw; sampling them independently from their marginals breaks it. And equation 44 uses no property of $\mathbf{p}$ beyond being a probability vector, so an empirical failure of consistency diagnoses the coupling, not the fit. □

Proof of Corollary 2 (binary closed form). Reduction. At $C = 2$, shift-invariance of the $\arg\max$ under $\mathbf{f} \mapsto \mathbf{f} + a\mathbf{1}_2$ lets us write the event $\{X_r = 1\}$ as $\{f_{r,1} + G_{r,1} > f_{r,2} + G_{r,2}\} = \{G_{r,2} - G_{r,1} < f_r\}$ with $f_r := f_{r,1} - f_{r,2}$. The difference of two independent standard Gumbels is standard logistic: for $L := G_{r,2} - G_{r,1}$,

$$P(L \leq t) = \int_{\mathbb{R}} \exp\big(-e^{-(g+t)}\big)\, e^{-g} \exp\big(-e^{-g}\big)\, dg = \frac{1}{1 + e^{-t}} = \sigma(t), \tag{45}$$

by the substitution $z = e^{-g}$, which turns the integral into $\int_0^\infty e^{-z(1+e^{-t})}dz = (1+e^{-t})^{-1}$. Setting $U_r := \sigma(L)$, which is $\mathrm{Uniform}(0,1)$ because σ is the CDF of L, and using strict monotonicity of σ,

$$\{X_r = 1\} = \{L < f_r\} = \{\sigma(L) < \sigma(f_r)\} = \{U_r < p\}, \qquad p := \sigma(f_r), \tag{46}$$

which is equation equation 8.

Abduction. By equation 46 the observation $x_r^F = 1$ is equivalent to $U_r < p^F$, and $x_r^F = 0$ to $U_r \geq p^F$. Since U_r is uniform and independent of f_r, its posterior is the uniform law restricted to the corresponding interval, which is equation equation 9. Note that this abduction is *exact*: no approximation intervenes between the observation and the posterior of the exogenous variable.

Prediction. Conditional on f_r, and hence on p^F and p^{CF}, the counterfactual event is $\{U_r < p^{\mathrm{CF}}\}$ with U_r drawn from equation 9. If $x_r^F = 1$ then $U_r \sim \mathrm{Uniform}(0, p^F)$ and

$$P\big(U_r < p^{\mathrm{CF}}\big) = \frac{|(0, p^{\mathrm{CF}}) \cap (0, p^F)|}{p^F} = \frac{\min(p^{\mathrm{CF}}, p^F)}{p^F}, \tag{47}$$

while if $x_r^F = 0$ then $U_r \sim \mathrm{Uniform}(p^F, 1)$ and

$$P\big(U_r < p^{\mathrm{CF}}\big) = \frac{|(p^F, 1) \cap (0, p^{\mathrm{CF}})|}{1 - p^F} = \frac{\max(0,\, p^{\mathrm{CF}} - p^F)}{1 - p^F}, \tag{48}$$

which together are equation equation 10. Marginalising over the joint Laplace posterior of $(f^F, \tilde{f})$, whose moments are equation 29–equation 30 at $C = 2$, gives equation equation 11.

Consistency and monotonicity. Setting $\tilde{\mathbf{x}}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)}$ gives $p^{\mathrm{CF}} = p^F$ almost surely, so equation 47 equals 1 and equation 48 equals 0, the direct consequence of Corollary 1. Moreover both expressions are non-decreasing in p^{CF}, so the counterfactual probability is monotone in the intervention whenever $p(\cdot)$ is, and here holds exactly and pathwise. □

Remark 3 (Common random numbers in the Monte Carlo estimator). When equation 6 or equation 11 is evaluated over a grid of intervention values $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ in order to trace a counterfactual response curve, the draws of $(\mathbf{f}^F, \tilde{\mathbf{f}})$ should share a common random number stream across grid points. The per-point marginals are unaffected — the draws are merely correlated across the grid — but the resulting curve is smooth rather than jittery, which preserves the exact monotonicity established in the proof of Corollary 2 at practical sample sizes. Independent draws per grid point leave a Monte Carlo residual that can locally violate monotonicity by $O(M^{-1/2})$.

A.6.3 The probit approximation

The following is used when a closed form is wanted in place of the Monte Carlo average in equation 11, at the cost of an additional approximation. It is stated separately because, unlike Corollary 2, it is not exact.

Lemma 4 (Gaussian–probit convolution). *For $\lambda \in \mathbb{R}$, $m \in \mathbb{R}$ and $s^2 > 0$,*

$$\int_{\mathbb{R}} \Phi(\lambda f)\, \mathcal{N}(f; m, s^2)\, df = \Phi\left(\frac{\lambda m}{\sqrt{1 + \lambda^2 s^2}}\right). \tag{49}$$

Proof. Let $Z \sim \mathcal{N}(0,1)$ be independent of $F \sim \mathcal{N}(m, s^2)$. Since $\Phi(\lambda f) = P(Z \leq \lambda f)$, conditioning on F and applying the tower rule gives

$$\int_{\mathbb{R}} \Phi(\lambda f)\, \mathcal{N}(f; m, s^2)\, df = \mathbb{E}_F\big[P(Z \leq \lambda F \mid F)\big] = P\big(Z - \lambda F \leq 0\big). \tag{50}$$

By independence, $Z - \lambda F \sim \mathcal{N}(-\lambda m,\, 1 + \lambda^2 s^2)$, so the right-hand side of equation 50 equals $\Phi\big(\lambda m / \sqrt{1 + \lambda^2 s^2}\big)$. □

Remark 4 (Where the probit approximation may and may not be used). Lemma 4 with $\lambda = \sqrt{\pi/8}$ — the scale matching $\Phi(\lambda \cdot)$ to σ to first order at the origin, since $\sigma'(0) = \frac{1}{4} = \lambda\varphi(0) = \lambda/\sqrt{2\pi}$ (MacKay, 1992; Bishop, 2006) — yields the familiar closed form $\mathbb{E}[\sigma(f)] \approx \Phi\big(\lambda m / \sqrt{1 + \lambda^2 s^2}\big)$ for the *interventional* class probability itself — the quantity obtained by evaluating the fitted mechanism at $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ and averaging over the latent posterior. It does *not* give a closed form for equation 11.

The reason is that equation 47–equation 48 are non-linear in the pair (p^F, p^{CF}) through $\min$, $\max$ and the ratio, so the expectation over the joint posterior does not reduce to a one-dimensional Gaussian integral of a probit. Applying Lemma 4 separately to p^F and p^{CF} and substituting the results into equation 10 computes $g\big(\mathbb{E}[p^F], \mathbb{E}[p^{\mathrm{CF}}]\big)$ in place of $\mathbb{E}\big[g(p^F, p^{\mathrm{CF}})\big)$, which by Jensen's inequality is biased in general; in particular it destroys the exact consistency of Corollary 1, since the substitution no longer forces $p^{\mathrm{CF}} = p^F$ pathwise. We therefore retain the Monte Carlo form in equation 11. Where differentiability of the estimator is required, the reparametrised draws of $(f^F, \tilde{f})$ are themselves differentiable in $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$, so pathwise gradients are available without the probit step.

A.6.4 Proof of the ordinal-GP counterfactual proposition

Numbering. This subsection proves Proposition 2. Equation equation 12 is the cutpoint reparameterisation, equation 13 the ordinal structural equation, equation 14 the compatibility identity, equation 15 the abduction of N_r, equation 16 the prediction given the latent and θ, and equation 17 the marginalisation over the joint posterior of (f_r, θ). Lemma 6, proved below alongside Lemma 5, supplies the Laplace-Gaussian posterior over θ that this marginalisation draws from. Corollary 3 is proved at the end of this subsection.

Throughout, Φ, φ denote the standard normal CDF and density. For a fixed realisation of θ, write $b_0(\theta) = -\infty < b_1(\theta) < \cdots < b_{C-1}(\theta) < b_C(\theta) = \infty$ for the cutpoints of eq. equation 13 at that θ via equation 12; as elsewhere in this appendix, all statements are for a fixed r with the subscript suppressed.

Lemma 5 (Laplace approximation for the ordinal-probit likelihood (Chu & Ghahramani, 2005))**.** *Fix a realisation of θ and write $b_c := b_c(\theta)$. Let $\pi(f; c) := \Phi(b_c - f) - \Phi(b_{c-1} - f)$ be the likelihood of observing class c given latent value f, and let $\Psi(\mathbf{f}; \theta) := \sum_{i=1}^{n} \log \pi(f_i; x_r^i) - \frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} +$ const be the unnormalised log posterior of the single-latent GP-ordinal model equation 13 at that θ. Because $\pi(\cdot; c)$ is a function of f_i alone, $\nabla_{\mathbf{f}}\Psi(\mathbf{f}; \theta) = \boldsymbol{\ell}'(\mathbf{f}; \theta) - \mathbf{K}^{-1}\mathbf{f}$ and $\nabla^2_{\mathbf{f}}\Psi(\mathbf{f}; \theta) = -W(\mathbf{f}; \theta) - \mathbf{K}^{-1}$ with $W(\mathbf{f}; \theta)$ diagonal, $W_{ii} = -\ell_i''(f_i)$, where*

$$\ell_i'(f_i) = \frac{\varphi(z_0^i) - \varphi(z_1^i)}{\pi(f_i; x_r^i)}, \qquad \ell_i''(f_i) = \frac{z_0^i \varphi(z_0^i) - z_1^i \varphi(z_1^i)}{\pi(f_i; x_r^i)} - \big(\ell_i'(f_i)\big)^2, \tag{51}$$

$z_1^i := b_{x_r^i} - f_i$, $z_0^i := b_{x_r^i - 1} - f_i$ (with $\varphi(\pm\infty) := 0$, and $z\varphi(z) \to 0$ as $z \to \pm\infty$). Chu & Ghahramani (2005) give this construction — their Laplace approximation — together with a treatment of the resulting Newton iteration; where $W(\hat{\mathbf{f}}; \theta) \succeq 0$ at the mode $\hat{\mathbf{f}}(\theta)$ (as it is for every fit reported in the validation section of the main text, verified numerically), the mode is characterised by $\hat{\mathbf{f}}(\theta) = \mathbf{K}\boldsymbol{\ell}'(\hat{\mathbf{f}}(\theta); \theta)$, the Laplace posterior is $q(\mathbf{f} \mid \theta, \cdot) = \mathcal{N}\big(\hat{\mathbf{f}}(\theta), (\mathbf{K}^{-1} + \hat{W})^{-1}\big)$, and the induced predictive law at any two inputs $\mathbf{x}^\star, \mathbf{x}^\dagger$ is jointly Gaussian with

$$\mathbb{E}_q[f(\mathbf{x}^\star) \mid \theta] = \mathbf{k}_\star^\top \boldsymbol{\ell}'(\hat{\mathbf{f}}(\theta); \theta), \qquad \mathrm{Cov}_q\big[f(\mathbf{x}^\star), f(\mathbf{x}^\dagger) \mid \theta\big] = k(\mathbf{x}^\star, \mathbf{x}^\dagger) - \mathbf{k}_\star^\top (\mathbf{K} + \hat{W}^{-1})^{-1} \mathbf{k}_\dagger, \tag{52}$$

the single-latent specialisation of Lemma 1(iv) with the softmax likelihood's $(\mathbf{e} - \hat{\boldsymbol{\pi}})$ and per-class $\hat{W}_c$ replaced by $\boldsymbol{\ell}'(\hat{\mathbf{f}}(\theta); \theta)$ and $\hat{W}$ above — the two constructions coincide once the block-diagonal structure of the multiclass case is set aside, since both are Newton–Laplace approximations to a likelihood that factorises over training points. Nothing above uses that θ is fixed rather than random: the statement holds conditionally on any realisation of θ, cutpoints $\mathbf{b}(\theta)$ playing the same role as any other given, fixed constant vector.

We take Lemma 5 as given and do not re-derive the concavity or convergence properties of the Newton iteration; those are exactly the content of Chu & Ghahramani (2005) and are orthogonal to what follows, which is the counterfactual construction *on top of* the fitted latent – the same division of labour adopted for the one-versus-rest categorical fit.

Lemma 6 (Laplace approximation for the cutpoint posterior)**.** *Let $\hat{\mathbf{f}}(\theta)$ and $\Psi(\cdot; \theta)$ be as in Lemma 5, and let $\mathrm{LML}(\theta) := \Psi(\hat{\mathbf{f}}(\theta); \theta) - \frac{1}{2}\log\det\big(I + \hat{W}^{1/2}\mathbf{K}\hat{W}^{1/2}\big)$ denote the Laplace log marginal likelihood of Lemma 5 evaluated at θ — the quantity Lemma 5's Newton iteration already produces for any fixed θ, since profiling out $\mathbf{f}$ requires no assumption on how θ itself was chosen. Let*

$$L(\theta) := \log p(\theta) + \mathrm{LML}(\theta), \qquad \hat{\theta} := \arg\max_{\theta} L(\theta).$$

By the envelope theorem, since $\partial\Psi/\partial\mathbf{f} = 0$ at $\mathbf{f} = \hat{\mathbf{f}}(\theta)$ for every θ, the total derivative $\nabla_\theta L(\theta)$ equals the partial derivative of $\log p(\theta) + \Psi(\mathbf{f};\theta)$ in θ holding $\mathbf{f} = \hat{\mathbf{f}}(\theta)$ fixed: the gradient of the profiled objective sees θ only through the explicit appearance of $\mathbf{b}(\theta)$ in the per-point likelihoods, with no term from $d\hat{\mathbf{f}}/d\theta$. The Hessian does not enjoy this simplification: differentiating the stationarity condition $\hat{\mathbf{f}}(\theta) = \mathbf{K}\boldsymbol{\ell}'(\hat{\mathbf{f}}(\theta);\theta)$ implicitly gives

$$\frac{d\hat{\mathbf{f}}}{d\theta_c} = \left(\mathbf{K}^{-1} + \hat{W}\right)^{-1}\mathbf{K}\,\frac{\partial\boldsymbol{\ell}'}{\partial\theta_c}\Big|_{\hat{\mathbf{f}}(\theta)},$$

*the single-latent analogue of the implicit-function step used to differentiate a GP's marginal likelihood with respect to its kernel hyperparameters (**?**, §5.5.1), and $\nabla^2_\theta L(\theta)$ carries a corresponding correction term through this Jacobian. The Laplace approximation to the profiled cutpoint posterior is then*

$$\theta \mid \mathcal{D} \;\approx\; \mathcal{N}\big(\hat{\theta}, \Sigma_\theta\big), \qquad \Sigma_\theta = -\big[\nabla^2_\theta L(\theta)\big]^{-1}_{\theta=\hat{\theta}}. \tag{53}$$

Remark 5 (Numerical versus symbolic Hessian). The correction term in $\nabla^2_\theta L$ above is exact but need not be coded symbolically: since $\theta \in \mathbb{R}^{C-1}$ is low-dimensional throughout this paper's experiments ($C \le 3$), Σ_θ in equation 53 is instead obtained by numerically differentiating $L(\theta)$ (central finite differences), which is the same quantity up to $O(\epsilon^2)$ truncation error and avoids re-deriving and re-implementing $d\hat{\mathbf{f}}/d\theta$ in code.

Proof of the ordinal proposition. Compatibility. Fix any realisation of θ and write $b_c := b_c(\theta)$. Since $N_r \sim \mathcal{N}(0,1)$ is independent of (f_r, θ) by construction of equation 13, $Y_r \mid f_r(\mathbf{x}), \theta \sim \mathcal{N}(f_r(\mathbf{x}), 1)$, and $X_r = c$ iff $Y_r \in [b_{c-1}, b_c)$. Hence $P(X_r = c \mid \mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}, f_r, \theta) = \Phi(b_c - f_r(\mathbf{x})) - \Phi(b_{c-1} - f_r(\mathbf{x}))$, which is equation 14; this is the defining likelihood of Lemma 5 at that θ, so no additional argument is needed. As N_r is exogenous and independent of everything upstream, the same identity evaluated at $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ is the interventional law, so equation 13 is compatible with any fitted f_r exactly as Proposition 1 is compatible with any fitted $\mathbf{p}$: nothing about Φ being the ordinal link, as opposed to a fitted probit or any other calibrated classifier's own class probabilities, is used in this half of the argument, and nothing in it uses θ's non-randomness – averaging equation 14 over the joint posterior of (f_r, θ) still reproduces the fitted model's own law by the tower rule.

Abduction. Fix any realisation of θ and write $c^* := x_r^F$, $(a^F, b^F) := (b_{c^*-1}(\theta), b_{c^*}(\theta))$, and condition on $f_r^F := f_r(\mathbf{x}^F_{\mathrm{pa}(r)})$. The event $X_r^F = c^*$ is, given f_r^F and θ, exactly $\{N_r \in (a^F - f_r^F,\, b^F - f_r^F)\}$, since $Y_r^F = f_r^F + N_r$. Because N_r is a priori $\mathcal{N}(0,1)$ and independent of (f_r, θ), and the only information the conditioning event carries about N_r is that it lies in this interval, Bayes' rule gives the posterior of N_r as its prior restricted and renormalised to the interval, i.e. a truncated standard normal. This is equation 15, and it is exact conditional on θ: no approximation intervenes between the observation and the posterior of the exogenous variable at a given θ, exactly as in the Abduction paragraph of Corollary 2's proof, with the interval-restricted standard normal in place of the interval-restricted uniform. Since θ itself is not known, the noise's actual posterior is obtained by the tower rule, averaging equation 15 over the joint posterior of (f_r^F, θ) given the data and the factual observation – a continuous mixture of truncated normals rather than a single one, since different θ draws truncate to different intervals.

Action and prediction. Fix any realisation of θ. The intervention replaces $\mathbf{x}^F_{\mathrm{pa}(r)}$ by $\tilde{\mathbf{x}}_{\mathrm{pa}(r)}$ and leaves f_r, θ, and N_r untouched, so with $\tilde{f}_r := f_r(\tilde{\mathbf{x}}_{\mathrm{pa}(r)})$ the counterfactual latent is $\tilde{Y}_r = \tilde{f}_r + N_r$, the same N_r as abducted in equation 15. Writing $\gamma_k := b_k(\theta) - \tilde{f}_r$, $\{X_r^{\mathrm{CF}} = k\} = \{\tilde{Y}_r \in [b_{k-1}(\theta), b_k(\theta))\} = \{N_r \in (\gamma_{k-1}, \gamma_k)\}$, so with $\alpha := a^F - f_r^F$, $\beta := b^F - f_r^F$,

$$P\big(X_r^{\mathrm{CF}} = k \mid \mathbf{x}^F, f_r^F, \tilde{f}_r, \theta\big) = P\big(N_r \in (\gamma_{k-1}, \gamma_k) \mid N_r \in (\alpha, \beta)\big) = \frac{\big|(\gamma_{k-1}, \gamma_k) \cap (\alpha, \beta)\big|_{\mathcal{N}(0,1)}}{\big|(\alpha, \beta)\big|_{\mathcal{N}(0,1)}}, \tag{54}$$

which, writing the intersection's endpoints as $\max(\alpha, \gamma_{k-1})$ and $\min(\beta, \gamma_k)$ and clipping a possibly empty intersection to probability zero via $(\cdot)_+$, is exactly equation 16 — the same overlap-of-two-intervals computation as equation 47–equation 48, now for two general intervals under a standard normal rather than two half-lines under a uniform, and conditional on θ exactly the fixed-cutpoint computation. Since f_r is not known but only its Laplace posterior given θ is (Lemma 5), the pair $(f_r^F, \tilde{f}_r)$ must be drawn from that joint posterior and the indicator averaged over it. And since θ is

not known either, only its Laplace-Gaussian posterior equation 53 (Lemma 6) is, θ itself must be drawn and the whole construction – refit $(f_r^F, \tilde{f}_r)$'s joint posterior at the resulting $\mathbf{b}(\theta)$, draw from it, evaluate equation 16 – repeated per draw, giving the nested average equation 17. The nested Monte Carlo estimator is unbiased and converges almost surely by the same argument as in the proof of Proposition 1, applied twice: at each fixed $\theta^{(s)}$ the inner average over $(f_r^F, \tilde{f}_r)$ is bounded with every draw contributing a vertex of the simplex, so the inner running average is itself a bounded, unbiased estimate of $\mathbb{E}_{(f_r^F, \tilde{f}_r)|\theta^{(s)}}[equation\ 16]$ for every s; averaging these in turn over $\theta^{(s)} \sim \mathcal{N}(\hat{\theta}, \Sigma_\theta)$ is again a running average of a bounded quantity, hence unbiased and almost-surely convergent to equation 17 by the same law-of-large-numbers argument one level up.

Consistency (specialisation). Fix any realisation of θ. Setting $\tilde{\mathbf{x}}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)}$ gives $\tilde{f}_r = f_r^F$ a.s. by the same joint-posterior-on-the-diagonal argument as Corollary 1's proof, hence $\gamma_k = b_k(\theta) - f_r^F$ for every k; taking $k = c^*$ gives $(\gamma_{c^*-1}, \gamma_{c^*}) = (\alpha, \beta)$, so the interval in the numerator of equation 16 *contains* (α, β) and the ratio in equation 16 equals 1, while for $k \neq c^*$ the intervals (γ_{k-1}, γ_k) and (α, β) are disjoint (the γ's partition the real line the same way the factual class's own interval does). Hence $do(\mathbf{X}_{\mathrm{pa}(r)} = \mathbf{x}^F_{\mathrm{pa}(r)})$ returns the point mass at $c^* = x_r^F$ for every realisation of θ, and equation 17 inherits this since it is a convex combination, over both θ and the inner posterior draws, of point masses at c^*.

Corollary 3. If $p(\theta) = \delta_{\theta_0}$, then $\hat{\theta} = \theta_0$ and $\Sigma_\theta = 0$ maximise and (degenerately) characterise the posterior equation 53 exactly, so every draw $\theta^{(s)} = \theta_0$ in equation 17: the outer expectation over θ collapses to evaluation at the single point θ_0, and the nested estimator of the Action-and-prediction paragraph above reduces to a single Laplace fit of f_r at the fixed cutpoints $\mathbf{b}(\theta_0)$ followed by the ordinary Monte Carlo average over $(f_r^F, \tilde{f}_r)$ – exactly Chu and Ghahramani's construction (Chu & Ghahramani, 2005), used throughout this paper prior to this section. □

Remark 6 (Why this is not the multiclass Laplace approximation of §A.6.1, restricted to two classes). The ordinal likelihood at $C = 2$ reduces to a single cutpoint b_1 and coincides with a probit link on $Y_r - b_1$, not with the logistic (sigmoid) link of Corollary 2. The two links are close in shape but not identical (MacKay, 1992; Bishop, 2006), so the $C = 2$ ordinal-GP estimator is a distinct (probit) binary classifier from the project's default (logistic) GPC, even though both couple a single latent to a single scalar noise. Nothing in the counterfactual argument above depends on which of the two links is used; only the per-point derivatives of Lemma 5 would change.