Binary Regression Revisited: Universal Ising Model, Binary Expansion and Beyond

Yi-Chen Lee[1*] **Mingcheng Lu**[2*] **En-Jui Kuo**[3] **Han Liu**[4] **Jerry Yao-Chieh Hu**[4,5]

[1] Johns Hopkins University [2] Fudan University [3] National Yang Ming Chiao Tung University
[4] Northwestern University [5] Vanderbilt University

ylee268@jh.edu, 2860215400pp@gmail.com, kuoenjui@nycu.edu.tw
hanliu@northwestern.edu, jhu@u.northwestern.edu

Abstract

We introduce a constructive universality principle for structural interpretability in binary regression. Let $A \in \{-1,1\}^p$ be binary covariates and $B \in \{-1,1\}$ be a binary response. We represent the binary-regression target $\mathbb{E}[B|A]$ by two-body Ising Hamiltonians over observed and auxiliary binary variables. This construction encodes covariate configurations as Boolean states and maps their input-output relationships through faithful reductions from Boolean satisfiability (SAT) to ground-state-energy Ising problems. This gives an explicit graph representation of high-order binary dependence. We showcase this principle with the *Universal Ising* (UI) framework. Unlike full power-vector expansion, UI does not enumerate high-order product features. It recasts these effects through auxiliary variables and pairwise interactions on an augmented binary graph. For a fixed representation, we establish least-squares convergence to the population minimizer set. Proof-of-concept experiments corroborate our theory and assess finite-sample prediction.

1 Introduction

We introduce a constructive universality principle for structural interpretability in binary regression. Given the binary regression target $m(A) = \mathbb{E}[B|A]$ with $A \in \{-1,1\}^p$ and $B \in \{-1,1\}$, our goal is to construct an explicit pairwise Ising Hamiltonian representation of $m(A)$ on an augmented binary graph. That is, we seek a graph $G = (V, E)$ whose vertices contain the observed covariates A and auxiliary binary variables A_{aux}. Writing $X := (A, A_{\text{aux}})$, we construct an Ising Hamiltonian

$$H_G(X) = \sum_{(i,j)\in E} J_{ij} X_i X_j + \sum_{i\in V} r_i X_i \quad \text{such that} \quad m(A) = H_G(X) = H_G(A, \alpha(A)),$$

for all configurations of A. Here $\alpha : \{-1,1\}^p \to \{-1,1\}^{N_{\text{aux}}}$ is an auxiliary-state selector: it assigns to each observed configuration A the auxiliary configuration used in the representation, where N_{aux} denotes the number of auxiliary variables. We call such a construction *universal* when it represents arbitrary binary-input conditional means. Our use of universality follows the spirit of universal Ising spin models [1], but the statistical object of interest here is the conditional mean of a binary response.

The benefit of this representation is structural. An arbitrary binary conditional mean $m(A)$ may encode hard-to-interpret high-order dependence among observed covariates. Our key idea is to use Ising models to convert such dependence into an interpretable pairwise graph representation. Specifically, we encode covariate configurations by Boolean variables and construct SAT instances whose satisfying assignments preserve the desired input-output structure. Through faithful reductions from SAT to ground-state-energy Ising problems, these assignments are mapped into ground-state configurations of pairwise Ising Hamiltonians [1, 2]. Closure of the Ising Hamiltonian class then combines these gadgets to reproduce the target conditional mean. Consequently, the high-order

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

dependence among observed covariates is mediated through auxiliary states, while the Hamiltonian itself remains a graph-based pairwise model. In this sense, the interpretability is structural: the interpretable object is the pairwise factor graph, namely its vertices, edges, and couplings.

This construction yields a *universal Ising representation* for binary regression using only unary and pairwise terms on an augmented graph. At the population level, every binary conditional mean admits such a representation with values in $[-1, 1]$, without requiring a nonlinear link function. In contrast, the standard generalized linear model (GLM) connects the conditional mean to a linear predictor via a specified link function. For example, logistic regression models conditional response probability as

$$\mathbb{P}(B = 1|A) = g^{-1}\left(\gamma_0 + \gamma_1 A_1 + \gamma_2 A_2 + \ldots + \gamma_p A_p\right) \coloneqq g^{-1}\left(\gamma_0 + A^\top \gamma\right), \tag{1.1}$$

where $\gamma_0 \in \mathbb{R}$, $\gamma \in \mathbb{R}^p$ and $g(\mu) = \log(\mu/(1-\mu))$ is the logit link function [3][2]. Logistic regression is naturally interpreted on the log-odds scale, while changes on the conditional-mean scale are mediated by the non-linear inverse link. The Ising route instead represents $m(A)$ directly by a pairwise Hamiltonian on an augmented binary graph. This gives a representation bridge among GLMs [3], binary expansion [4, 5], and Ising-type models [1, 2]: it keeps the target on the conditional-mean scale while expressing dependence through interpretable pairwise graph structure.

To situate this representation goal, we connect it to two existing lines of work:

1. **Linearity via Binary Expansion.** A first route to direct linearity in binary regression is the *binary expansion* approach [4, 5]. Binary expansion maps $A \in \{-1, 1\}^p$ to the power vector $A^\otimes$ containing all interaction terms among observed covariates. Since these features form a basis for functions on $\{-1, 1\}^p$, any binary-input conditional mean can be written as
$$\mathbb{E}[B|A] = \beta^\top A^\otimes$$
for a suitable β. This representation is exact and avoids the GLM link, but it introduces explicit high-order product features up to order p. Thus, binary expansion gives the desired representational capacity through high-order observed features. Please see Section G for details.
2. **Auxiliary-Variable Pairwise Reductions.** Auxiliary-variable reductions from higher-order binary objectives to pairwise forms are classical in pseudo-Boolean optimization, MRF inference, QUBO, and Ising formulations [6–8, 2, 9, 10]. This literature shows how high-order binary structure can be represented through pairwise interactions after introducing auxiliary variables.

Our universal Ising representation unifies these views: it directly represents conditional means, replaces high-order products of observed covariates with auxiliary states, and constructs a pairwise Hamiltonian through faithful reductions. Motivated by this expressivity-structure tension, we ask:

Question 1. Can arbitrary binary conditional means be represented exactly as linear combinations of binary covariates without committing to a specific non-linear link function or to full binary expansion?

Question 2. Can such a representation be implemented by a two-body Ising Hamiltonian on an augmented binary graph that supports consistent estimation of the binary conditional mean?

Contributions. We address these questions and make four key contributions:

- **Universality through Faithful Reductions.** We address Question 1 by formalizing a universality construction for binary regression. For Hamiltonian model classes that satisfy a closure property and whose ground-state-energy problems admit faithful reductions from SAT, we construct exact representations of arbitrary binary conditional means over observed and auxiliary binary variables.
- **Two-Body Binary Linear Model with a Universal Ising Representation.** We address Question 2 through a universal Ising representation whose Hamiltonian contains only unary and pairwise terms. For any function $f : \{-1, 1\}^p \to \mathbb{R}$, we construct an augmented binary graph and a Hamiltonian H_G such that $f(A) = H_G(A, \alpha(A))$ for all A, with explicit upper bounds on the size of the graph.
- **Relation to GLMs, Binary Expansion, and Pairwise Reductions.** We position the universal Ising representation relative to GLMs, binary expansion, and classical auxiliary-variable pairwise reductions. We describe conditional-independence properties through the augmented graph features and establish least-squares convergence to the population minimizer set for a fixed representation.

[2] Equivalently, $m(A) = \mathbb{E}[B|A] = 2g^{-1}(\gamma_0 + A^\top \gamma) - 1$.

- **Extensions and Experiments.** We discuss extensions beyond binary covariates via discrete encoding and discretization. We evaluate finite-sample prediction in synthetic binary regression and examine how auxiliary variable counts affect the approximation of random Boolean functions.

Organization. Section 2 presents the construction of a universal model class by leveraging the hardness equivalence among computational problems. Section 3 presents a concrete example of universal model: the two-body, linear, universal representation for all binary-input models. We present conclusions in Section A. We defer the proof-of-concept experiments in Section H.

Notations. Let $\mathbb{N}$ denote natural numbers. We denote a statistical model by an upper-case M. We denote a family of statistical models consisting of models with identical structures and different numbers of covariates by $\mathcal{M}$. We denote the cardinality of a finite set S by $|S|$. We denote the modulo operation by mod. We denote the index set $\{1, \ldots, I\}$ by $[I]$. We denote the degree of a vertex in a graph by $\deg$. We denote the ceiling function by $\lceil \cdot \rceil$.

2 Construction of Universality

In this section, we present a high-level overview of the universality principle, and its role in binary regression. Section 2.1 formalizes the underlying computational problems required for the universality construction. Section 2.2 connects these ideas to a universality-based binary-regression framework.

2.1 Preliminaries

We revisit two computational problems: (i) the Boolean satisfiability problem (SAT) and (ii) the ground-state-energy (GSE) problem. We further formalize the faithful reduction (Definition 2.6).

Formulation of SAT. We begin with several basic definitions in Boolean algebra following [1].

Definition 2.1 (Boolean Notation)**.** Boolean variables take values 0 (false) or 1 (true). The *Boolean connectives* $\wedge$, $\vee$, and $\neg$ denote AND, OR, and NOT (negation) operation, respectively. A *literal* is a Boolean variable or its negation. For $k \in \mathbb{N}$ and $x, \ell \in \{0,1\}^k$, the *elementary basis functions* are $e_\ell(x) := \mathbb{1}\{x = \ell\}$. For $x_1, x_2 \in \{0,1\}$, the *exclusive-nor operation* is $x_1 \odot x_2 := \mathbb{1}\{x_1 = x_2\}$.

These basic elements and operations serve as the building blocks of *Boolean expressions*, which we formalize below together with the notion of satisfiability.

Definition 2.2 (Boolean Expression, Conjunction and Disjunction)**.** A Boolean expression, denoted by ϕ, is a finite combination of Boolean literals and connectives. The expression ϕ is *satisfiable* if there exists an assignment of values in $\{0,1\}$ to its variables under which ϕ evaluates to 1. Given Boolean expressions ϕ_1 and ϕ_2, their *conjunction* and *disjunction* are $\phi_1 \wedge \phi_2$ and $\phi_1 \vee \phi_2$, respectively.

Building upon these notions, we introduce *conjunctive normal form* for Boolean expressions:

Definition 2.3 (Conjunctive Normal Form)**.** A *clause* C is a disjunction of literals[a]. Then, we say that a Boolean expression ϕ is in *conjunctive normal form* (CNF) if $\phi = \bigwedge_{i=1}^{m} C_i$ for some $m \in \mathbb{N}$, where each C_i is a clause. If each clause contains exactly three literals, we call ϕ a *3CNF expression*.

[a] Note that every clause is a Boolean expression, but not every Boolean expression qualifies as a clause.

Importantly, Boolean expressions specify logical relations among binary variables, allowing us to encode constraints linking observed and auxiliary variables. To preserve satisfying assignments in the resulting Hamiltonian models, our construction relies on faithful reductions from SAT and 3SAT:

Definition 2.4 (SAT and 3SAT)**.** The Boolean satisfiability problem (SAT) asks whether a given Boolean expression is satisfiable, and the 3SAT problem further restricts the input to 3CNF.

The second computational problem we use in our construction is *ground-state-energy* (GSE) problem.

Definition 2.5 (Hamiltonian, Energy Spectrum, and GSE)**.** A *Hamiltonian* H on p binary variables is a function $H : \{-1,1\}^p \to \mathbb{R}$ that assigns an energy $H(A)$ to each configuration $A \in \{-1,1\}^p$. Its *energy spectrum* is the set $\{H(A) : A \in \{-1,1\}^p\}$, whose smallest value is the *ground-state energy*. For a family $\mathcal{M}$ of Hamiltonians, we consider the following version of the *GSE problem*: given a nonnegative Hamiltonian $H \in \mathcal{M}$, does there exist a configuration A such that $H(A) = 0$?

We remark that *a family of models* is a collection of models, each characterized by a Hamiltonian. The number of covariates and parameter values may differ across models, but all models in the collection share a common functional form. Within a fixed family, each Hamiltonian defines an instance of GSE, which, like SAT, asks whether a solution exists. At this decision level, polynomial-time reductions connect different problems by transforming input instances while preserving their YES/NO answers. NP-complete problems admit such reductions to one another and are equivalent in this sense. For our binary regression construction, however, this equivalence alone is **insufficient** because the Ising representation must also retain the logical relations encoded by individual assignments. This need for assignment information parallels the distinction between decision and function versions of a problem. For example, in the function version of SAT (FSAT), we return a satisfying assignment when one exists and NO otherwise. To retain these assignments in the GSE problem, we require a one-to-one correspondence between satisfying assignments and its solutions restricted to a designated subset of variables. This stronger requirement motivates the notion of *faithful reduction* introduced by [1].

Definition 2.6 (Faithful Reduction)**.** Consider two decision problems $\mathcal{L}_1$ and $\mathcal{L}_2$ with solution sets $\mathcal{S}_1$ and $\mathcal{S}_2$ where $|\mathcal{S}_2| \geq |\mathcal{S}_1|$. We say that a reduction from $\mathcal{L}_1$ to $\mathcal{L}_2$ is faithful if there exists a bijective mapping h such that $s_{2,i} = h(s_{1,i})$ for all $s_{1,i} \in \mathcal{S}_1$ and $s_{2,i} \in \mathcal{S}_2$ with $i = 1, \ldots, |\mathcal{S}_1|$.

Given any reduction from problem $\mathcal{L}_1$ to $\mathcal{L}_2$, the number of variables in $\mathcal{L}_2$ are larger than those in $\mathcal{L}_1$. The faithful reduction preserves the structure of solutions via the one-to-one correspondence between elements in the solution set of $\mathcal{L}_1$ and elements in the subset of the solution set of $\mathcal{L}_2$.

Lastly, we define the *closure* property of a family of models.

Definition 2.7 (Closure)**.** Let $\mathcal{M}$ denote a family of models with covariates A. Let $H_1, H_2 \in \mathcal{M}$ be the Hamiltonian acting on $A_1, A_2 \subseteq A$ respectively. Then, we say $\mathcal{M}$ is closed if for any H_1, H_2:

1. There exists a Hamiltonian $H \in \mathcal{M}$ acting on $A_3 \subseteq A$ such that $H(A_3) = H_1(A_1) + H_2(A_2)$.
2. Let $A_3^\star$ denote the configuration such that the first condition holds. There exists a subset $B_3^\star \subseteq A_3^\star$ whose elements form a one-to-one correspondence with every $a \in A_1 \cup A_2$.

The closure property ensures that Hamiltonians constructed for individual configurations can be combined into a single Hamiltonian. If $\mathcal{M}$ imposes no constraints on interaction patterns, closure follows naturally. It only becomes a non-trivial property when $\mathcal{M}$ restricts the interaction patterns to some specific structures, such as graphical models characterized by a 2D lattice or a 3D cubic.

2.2 Binary Regression via Universality

We now connect the computational ingredients above to binary regression by establishing universality.

Theorem 2.1 (Binary Regression via Universality)**.** Let $\mathcal{M}_{\text{uni}} = \{H_\theta\}_\theta$ be a family of models parametrized by θ. Suppose $\mathcal{M}_{\text{uni}}$ is closed and its GSE problem admits a polynomial-time faithful reduction from SAT. Then, for any $A = (A_1, \ldots, A_p)^\top \in \{-1, 1\}^p$ and $B \in \{-1, 1\}$, there exists a Hamiltonian $H_\theta \in \mathcal{M}_{\text{uni}}$, a $N \in \mathbb{N}_+$ and a binary vector $A_{\text{aux}} \in \{-1, 1\}^N$ such that:

$$\mathbb{E}[B|A] = H_\theta(A, A_{\text{aux}}), \tag{2.1}$$

where H_θ and A_{aux} are dependent on A, and N is determined by the structure of the faithful reduction specific to the model class $\mathcal{M}_{\text{uni}}$. Furthermore, (2.1) holds with $|H_\theta(A, A_{\text{aux}})| \leq 1$ for all A.

Theorem 2.1 establishes a general construction for applying a model class to binary regression distinct from GLMs or any binary expansion-based models. In the next section, we present a concrete example demonstrating the existence of $\mathcal{M}_{\text{uni}}$ characterized by linear H_θ with respect to all interaction terms while avoiding complex product terms. Crucially, our results on the universality extends existing framework to encompass a broader model class with well-structured ground state energy problems.

Proof Sketch. We perform faithful reductions from SAT to GSE problem of $H_\theta \in \mathcal{M}_{\text{uni}}$ three times.

Step 1. We introduce a model with a non-linear transformation as an intermediary between $\mathbb{E}[B|A]$ and $\mathcal{M}_{\text{uni}}$, ensuring that the outputs lies in $[-1, 1]$. Next. we introduce p Boolean variables to encode states of A. Last, we decompose $\mathbb{E}[B|A]$ into a sum of 2^p outputs in the Boolean elementary basis.

Step 2. We identify the solution to achieve (2.1) in the form of Boolean algebra. Next, we construct three SAT formulas that collectively encode such solution. Last, we introduce auxiliary covariates A_{aux} to perform three faithful reductions from SAT to GSE problems for models in $\mathcal{M}_{\text{uni}}$.

Step 3. We construct final Hamiltonian via GSE problems. Next, we equate the output values from our Hamiltonian with the decomposed $\mathbb{E}[B|A]$ in binary expansion form. Last, we identify configurations such that the equality holds. Please see to Section D.1 for a detailed proof. □

The next result demonstrates that Theorem 2.1 provides a direct modeling alternative to all GLMs.

Corollary 2.1.1 (Rethinking GLMs through Universality). Let A_{aux} and $\mathcal{M}_{\text{uni}} = \{H_\theta\}_\theta$ be a binary vector and a family of models in Theorem 2.1. Let $A \in \{-1, 1\}^p$ and $B \in \{-1, 1\}$. Then, for any GLMs with link function $g : (-1, 1) \to \mathbb{R}$ and parameter $\gamma \in \mathbb{R}^p$, there exist a H_θ such that:

$$\mathbb{E}[B|A] = g^{-1}(\gamma^\top A) = H_\theta(A, A_{\text{aux}}),$$

where the dimension of θ and A_{aux} are determined by the faithful reduction specific to $\mathcal{M}_{\text{uni}}$.

Further, such universality-based construction addresses the issue of perfect separation via universality.

Corollary 2.1.2 (Addressing Perfect Separation via Universality). Let A_{aux} and $\mathcal{M}_{\text{uni}} = \{H_\theta\}_\theta$ be a binary vector and a family of models defined in Theorem 2.1. Let $B \in \{-1, 1\}$ be response and $A \in \{-1, 1\}^p$ be covariates. Let $\{B_i, A_i\}_{i=1}^n$ be a set of samples. Then, there exists a H_θ such that:

$$B_i = H_\theta(A_i, A_{\text{aux}}) \quad \text{for all} \quad i \in [n],$$

where the dimension of θ and A_{aux} are determined by the faithful reduction specific to $\mathcal{M}_{\text{uni}}$.

Proof. Please see Sections D.2 and D.3 for proofs. □

3 Universal Ising Representation for Binary Regression

In this section, we showcase the universality principle given in Section 2.2 with a concrete model family. Specifically, we introduce a universal Ising representation for binary regression using unary and pairwise terms. Section 3.1 presents the relevant computational background. Next, Section 3.2 establishes the universality through an explicit Ising construction. Finally, Section 3.3 examines the conditional independence and gives least-squares statistical consistency for a fixed representation.

3.1 Related Computational Problems

A graph consists of a pair of sets $G = (V, E)$ with V representing the vertex set and E representing the edge set. Each element in E is a 2-element subset of V. First, we have:

Definition 3.1 (Vertex Cover Problem). Let $G = (V, E)$ be a graph. $\mathcal{C} \subseteq V$ is a vertex cover if for all edge set $\{u, v\} \in E$, at least one of its endpoints belongs to $\mathcal{C}$. The vertex cover problem asks: given a graph G and an integer K_1, does G has vertex cover of size K_1 or less?

Trivially, we have vertex cover $\mathcal{C}$ of size $|V|$. On top of that, the minimum vertex cover problem asks: what is the smallest possible size of a vertex cover? Next, we introduce the independent set problem.

Definition 3.2 (Independent Set Problem). Let $G = (V, E)$ be a graph. We say $I \subseteq V$ is an independent set if for any $a, b \in I$, $\{a, b\} \notin E$. The independent set problem asks: given a graph G and an integer K_2, does G has independent set of size K_2 or more?

The maximum independent set problem asks: what is the largest possible size of a independent set?

3.2 Two-Body Ising Representation of Binary Conditional Means

We construct a universal Ising representation for functions of binary covariates.

Definition 3.3 (Two-Body Binary Linear Model). The *two-body binary linear family*, which we call the *universal Ising* (UI) family $\mathcal{M}_{\text{UI}}$, is defined on some graph $G = (V, E)$, where the vertex set $V = \{A_1, \ldots, A_N\}$ represents binary covariates $A \in \{-1, 1\}^N$. The edge set E encodes the interaction patterns between distinct pairs $A_i, A_j \in V$. Explicitly, every model $H_G \in \mathcal{M}_{\text{UI}}$ follows:

$$H_G(A) = \sum_{(i,j)\in E} J_{i,j} A_i A_j + \sum_{i \in V} r_i A_i + K_0, \tag{3.1}$$

where the pairwise coupling $J_{i,j} \in \mathbb{R}$, local field $r_i \in \mathbb{R}$, and global shift $K_0 \in \mathbb{R}$ are parameters.

Universality framework holds for larger model class and achieve the reproduction for any energy spectrum by a general argument. In contrast, universal Ising framework includes a subfamily of $\mathcal{M}_{\text{uni}}$, restricting the universality-based model class to binary, linear and two-body interaction terms.

Theorem 3.1 (Universal Ising Representation). Let $A = (A_1, \ldots, A_p)^\top \in \{-1, 1\}^p$. Then, for any function $f : A \to \mathbb{R}$, there exists a $A_{\text{aux}} = (A_{p+1}, \ldots, A_{|V|})^\top \in \{-1, 1\}^{|V|-p}$ dependent on A, and Hamiltonian $H_G \in \mathcal{M}_{\text{UI}}$ fixed across all inputs A defined on a graph $G = (V, E)$ such that

$$f(A) = H_G(A, A_{\text{aux}}),$$

where $V = \{A_1, \ldots, A_p, A_{p+1}, \ldots A_{|V|}\}$. Furthermore, the cardinality of V and E are bounded by

$$|V| \leq 2^p \left(7p - 5 + 9\lceil p/2 \rceil\right), \quad |E| \leq 2^p \left(8p - 7 + 18\lceil p/2 \rceil\right).$$

Theorem 3.1 guarantees that any function over binary covariates admits a two-body linear representation (3.1) within $\mathcal{M}_{\text{UI}}$. The construction encodes higher-order effects of the observed covariates through auxiliary variables, allowing H_G to remain a linear combination of unary and pairwise graph features. This provides a alternative representation without specifying a nonlinear link function.

Proof Sketch. The construction consists of three steps:

Step 1: We introduce a flag covariate $F \in \{0, 1\}$ and Boolean variables $x = (x_1, x_2, x_3) \in \{0, 1\}^3$ to encode the states of A. Next, we define a SAT with a Boolean formula such that F signifies the states of x among all solutions of the SAT. Last, we reiterate the process for all configurations.

Step 2: We reduce each SAT to a vertex cover problem. Next, we reduce each vertex cover problem to a GSE problem. Last, we ensure the reduction from SAT to GSE problem is faithful.

Step 3: We construct Hamiltonian via GSE problems. Next, we equate the output values from our Hamiltonian with elements in $\mathcal{F}_{\text{energy}}$. Last, we identify configurations of every variables in each problem such that the equality holds. Please see Section E.1 for a detailed proof. □

We provide an example with special case $p = 3$ with $A = (A_1, A_2, A_3)^\top \in \{-1, 1\}^3$, to illustrate the idea. We denote the energy spectrum of f by $\mathcal{F}_{\text{energy}} = \{f^1, f^2, \ldots, f^n\}$ with $n \leq 2^3 = 8$.

Proof. We prove Theorem 3.1 for special case $p = 3$ with the following three steps.

Step 1: Encoding Desired Solutions in SAT. First, we handle a specific configuration; without loss of generality, we fix $A = (-1, 1, -1)^\top$. Further, we define $A_i = 2x_i - 1$ for $i \in [3]$

- **Step 1.1: Construct the Boolean Formula.** For a $x = (x_1, x_2, x_3)^\top \in \{0, 1\}^3$, we construct a corresponding Boolean formula ϕ_ℓ, where ℓ denotes the the configuration of interest. For example, for $x = (0, 1, 0)$, we construct a ϕ_{010} with two auxiliary covariates $F_{010}, u_{010} \in \{0, 1\}$:

$$\begin{aligned}
&\phi_{010}\left(x_1, x_2, x_3; F_{010}, u_{010}\right) \qquad (3.2)\\
&= \underbrace{(x_1 \vee \neg x_2 \vee u_{010})}_{\text{Clause1}} \wedge \underbrace{(\neg u_{010} \vee x_3 \vee F_{010})}_{\text{Clause2}} \wedge \underbrace{(\neg x_1 \vee \neg x_2 \vee \neg F_{010})}_{\text{Clause3}} \wedge \underbrace{(\neg x_1 \vee x_2 \vee \neg F_{010})}_{\text{Clause4}}\\
&\wedge \underbrace{(x_1 \vee x_2 \vee \neg F_{010})}_{\text{Clause5}} \wedge \underbrace{(\neg x_1 \vee \neg x_3 \vee \neg F_{010})}_{\text{Clause6}} \wedge \underbrace{(\neg x_1 \vee x_3 \vee \neg F_{010})}_{\text{Clause7}} \wedge \underbrace{(x_1 \vee \neg x_3 \vee \neg F_{010})}_{\text{Clause8}}.
\end{aligned}$$

  In particular, we refer to F_{010} as the *flag covariate*, as it signifies the state $x = (0, 1, 0)$ in the following sense. Specifically, we consider the two scenarios for the solutions to ϕ_{010}.

  Scenario 1: $F_{010} = 1$**.** Clauses 3 to 5 force $(x_1, x_2) = (0, 1)$, while Clauses 6 to 8 force $(x_1, x_3) = (0, 0)$. Together, these clauses restrict the input to $x = (0, 1, 0)$ in any satisfying assignment of (3.2). For this input, Clause 1 requires $u_{010} = 1$, which also satisfies Clause 2.
  Scenario 2: $F_{010} = 0$**.** Clauses 3 to 8 are automatically satisfied. For $x = (0, 1, 0)$, Clause 1 requires $u_{010} = 1$, whereas Clause 2 requires $u_{010} = 0$, so no satisfying assignment exists. For every other input, Clauses 1 and 2 can both be satisfied: choose $u_{010} = 0$ when $x_1 = 1$ or $x_2 = 0$, and the only remaining input is $x = (0, 1, 1)$, for which choosing $u_{010} = 1$ satisfies both clauses.

  Combining Scenario 1 and Scenario 2, every input x admits a choice of (F_{010}, u_{010}) that satisfies (3.2). Most importantly, in every satisfying assignment, $F_{010} = 1$ if and only if $x = (0, 1, 0)$.
- **Step 1.2: Extension to All Configurations.** Let $\ell \in \{x = (x_1, x_2, x_3) | x \in \{0, 1\}^3\}$. The corresponding flag variable $F_\ell \in \{0, 1\}$ and Boolean formula ϕ_ℓ for an arbitrary ℓ takes the form:

$$\begin{aligned}
&\phi_\ell\left(x_1, x_2, x_3; F_\ell, u_\ell\right) \qquad (3.3)\\
&= \underbrace{(\xi_1 \vee \eta_1 \vee u_\ell)}_{\text{Clause1}} \wedge \underbrace{(\xi_2 \vee \eta_2 \vee F_\ell)}_{\text{Clause2}} \wedge \underbrace{(\xi_3 \vee \eta_3 \vee \neg F_\ell)}_{\text{Clause3}} \wedge \underbrace{(\xi_4 \vee \eta_4 \vee \neg F_\ell)}_{\text{Clause4}}
\end{aligned}$$

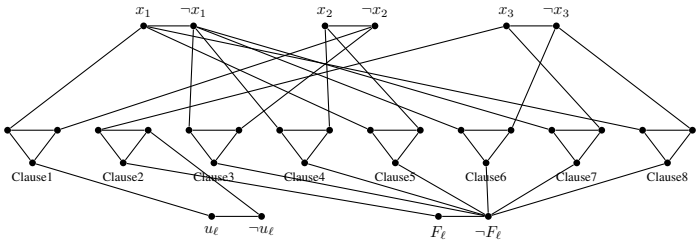


Figure 1: Vertex Cover Problem in (3.2) for $p = 3$.

$$\wedge \underbrace{(\xi_5 \vee \eta_5 \vee \neg F_\ell)}_{\text{Clause5}} \wedge \underbrace{(\xi_6 \vee \eta_6 \vee \neg F_\ell)}_{\text{Clause6}} \wedge \underbrace{(\xi_7 \vee \eta_7 \vee \neg F_\ell)}_{\text{Clause7}} \wedge \underbrace{(\xi_8 \vee \eta_8 \vee \neg F_\ell)}_{\text{Clause8}},$$

where $u_\ell \in \{0, 1\}$ is an additional auxiliary variable, and

$$(\xi_1, \eta_1) = (\neg x_1, \neg x_2); \quad (\xi_2, \eta_2) = (\neg u_\ell, \neg x_3); \quad (\xi_3, \eta_3) = (x_1, \neg x_2); \quad (\xi_4, \eta_4) = (x_1, x_2); \tag{3.4}$$

$$(\xi_5, \eta_5) = (\neg x_1, x_2); \quad (\xi_6, \eta_6) = (x_1, x_3); \quad (\xi_7, \eta_7) = (x_1, \neg x_3); \quad (\xi_8, \eta_8) = (\neg x_1, x_3). \tag{3.5}$$

Here, ξ_j and η_j are both literals for all $j \in [8]$, each consisting of a Boolean variable or its negation, chosen according to the given target configuration ℓ. For example, when $\ell = (1, 0, 0)$, we keep x_1 unchanged, swap x_2 with $\neg x_2$, and swap x_3 with $\neg x_3$ throughout the expressions given by (3.4).

Step 2: Faithful Reduction to the GSE Problem. This step builds on existing results. First, we perform a faithful reduction from the SAT to a vertex cover problem [1, 11]. Second, we perform another faithful reduction from the vertex cover problem to a GSE problem [1, 12].

- **Step 2.1: From SAT to Vertex Cover.** For any ℓ with corresponding SAT formula ϕ_ℓ in 3CNF (3.3), we faithfully reduce it to a minimum vertex cover problem on a graph $G_\ell = (V_\ell, E_\ell)$.

  Construction of Vertex Set. For Boolean variables $(x_1, x_2, x_3; F_\ell, u_\ell)$, we define a vertex set $V_B = \{v_1, v_2, v_3, \bar{v}_1, \bar{v}_2, \bar{v}_3, v_4, \bar{v}_4, v_5, \bar{v}_5\}$ by $h_1 : x_i \to v_i$; $\neg x_i \to \bar{v}_i$ for $i = [3]$, and $F_\ell \to v_4$; $\neg F_\ell \to \bar{v}_4$; $u_\ell \to v_5$; $\neg u_\ell \to \bar{v}_5$. We define vertex set $V_C = \bigcup_{j=1}^{8} \{c_j^1, c_j^2, c_j^3\}$ by $h_2 : \xi_j \to c_j^1$; $\eta_j \to c_j^2$ for $j \in [8]$. For the third literal, we map $u_\ell \to c_1^3$, $F_\ell \to c_2^3$, $\neg F_\ell \to c_j^3$ for $j = 3, \ldots, 8$. Here, the index j denotes each clause in (3.3). Altogether, we have $V_\ell = V_B \cup V_C$ with $|V| = 34$.

  Construction of Edge Set. We define E as the union of three subsets, including (i): edges between vertices within V_B, denoted by E_B (ii): edges between vertices within V_C, denoted by E_C (iii): edges between vertices in V_B and V_C, denoted by $E_{B,C}$, with the following expression:

$$E_B = \bigcup_{i=1}^{5} \Big\{\{v_i, \bar{v}_i\}\Big\}, \; E_C = \bigcup_{j=1}^{8} \Big\{\{c_j^1, c_j^2\}, \{c_j^1, c_j^3\}, \{c_j^2, c_j^3\}\Big\}, \; E_{B,C} = \bigcup_{j=1}^{8} \Big\{E_j^1, E_j^2, E_j^3\Big\}, \tag{3.6}$$

  where E_j^k connects vertex k of the j-th Clause to the vertex in V_B representing the same literal:

$$E_j^k = \{c_j^k, a^k \mid a^k \in V_B, \; h_1^{-1}(a^k) = h_2^{-1}(c_j^k)\}, \quad \text{for} \quad k = 1, 2, 3$$

  Then, we have a graph $G_\ell = (V_\ell, E_\ell)$ constructed from the Boolean formula ϕ_ℓ, where $V_\ell = V_B \cup V_C$ and $E_\ell = E_B \cup E_C \cup E_{B,C}$, as illustrated in Figure 1. This construction faithfully reduces SAT for ϕ_ℓ to the minimum vertex cover problem on G_ℓ. For the correspondence between satisfying assignments and minimum vertex covers, please see Section C.1 for a detailed argument.

- **Step 2.2: From Vertex Cover to GSE.** Given a graph $G_\ell = (V_\ell, E_\ell)$, we faithfully reduce it to a GSE problem of a two-body binary model. Specifically, for vertices in V_B, we define $h_3 : v_i \to s_{2i-1}$ and $\bar{v}_i \to s_{2i}$ for $i = 1, \ldots, 5$. For vertices in V_C, we define $h_4 : c_j^1 \to s_{3j+8}$; $c_j^2 \to s_{3j+9}$; $c_j^3 \to s_{3j+10}$ for $j = 1, \ldots, 8$. Then, we assign s_i to 1 if its pre-image is in the independent set, and to 0 otherwise. For the edge set, we include a two-body term $s_i s_j$ whenever the corresponding vertices form an edge. By the previous steps (3.6), there are $5 + 24 + 24 = 53$ such two-body terms. Combining these terms, we have the following two-body Ising Hamiltonian:

$$H_\ell(s) = -\sum_{i\in V_\ell} s_i + 2\sum_{(i,j)\in E_\ell} s_i s_j + K_0. \tag{3.7}$$

where $K_0 = 13$ shifts the ground-state energy to zero, since the maximum independent set has size $5 + 8 = 13$. Then, we define $A_i := 1 - 2s_i$ for all $i \in [34]$. This gives:

$$\begin{aligned} H_\ell(A) &= \underbrace{-\frac{1}{2}\sum_{i\in V_\ell}(1 - A_i)}_{(\mathrm{T}_1)} + \underbrace{\frac{1}{2}\sum_{(i,j)\in E_\ell}(1 - A_i)(1 - A_j)}_{(\mathrm{T}_2)} + K_0 \\ &= \underbrace{-\frac{|V_\ell|}{2} + \frac{1}{2}\sum_{i\in V_\ell} A_i}_{(\mathrm{T}_1)} + \underbrace{\frac{|E_\ell|}{2} + \frac{1}{2}\sum_{(i,j)\in E_\ell} A_i A_j}_{(\mathrm{T}_{2.1})} + K_0 - \underbrace{\frac{1}{2}\sum_{(i,j)\in E_\ell}(A_i + A_j)}_{(\mathrm{T}_{2.2})}. \end{aligned} \tag{3.8}$$

Here, (T_1) and $(\mathrm{T}_{2.1})$ match the form of (3.1). For $(\mathrm{T}_{2.2})$, each A_i appears once for every incident edge, giving $-0.5\sum_{i\in V_\ell} \mathsf{deg}(A_i)A_i$, where $\mathsf{deg}(\cdot)$ denotes the degree of the vertex. Since each degree depends only on the graph, this term can be absorbed into the one-body coefficients in (T_1). Therefore, (3.8) is a two-body binary linear model. For the one-to-one correspondence between minimum vertex covers and ground states, please refer to Section C.2 for a detailed argument.

To this end, every input x admits an extension to a spin configuration with $H_\ell = 0$ by Step 1 and the faithful reductions throughout. In every such extension, $A_7 = 2F_\ell - 1$, so $A_7 = 1$ if and only if $x = \ell$. Moreover, since H_ℓ takes nonnegative integer values, every nonzero energy is at least 1.

Step 3: Universality Construction. For any $f^\ell \in \mathcal{F}_{\text{energy}}$, we define an corresponding Hamiltonian:

$$\mathcal{H}_\ell := \Big(\Delta + \sum_{\ell'} |f^{\ell'}|\Big) \cdot H_\ell + f^\ell \cdot \Big(\frac{A_7 + 1}{2}\Big), \tag{3.9}$$

where $\Delta > \max_{\ell'}\{f^{\ell'} | f^{\ell'} \in \mathcal{F}_{\text{energy}}\}$ and the sum ranges over all eight target configurations. When the eight terms $\mathcal{H}_\ell$ are summed, the flag terms contribute at least $-\sum_{\ell'} |f^{\ell'}|$. Since each H_ℓ is a nonnegative integer, any configuration with at least one $H_\ell > 0$ therefore has total energy at least Δ.

Since all reductions are faithful, every input admits an extension to the remaining spins with $H_\ell = 0$. In every such extension, $A_7 = 1$ if and only if the input matches ℓ, under the identification $0 \leftrightarrow -1$ and $1 \leftrightarrow 1$. Thus, whenever $H_\ell = 0$, we have $\mathcal{H}_\ell = f^\ell$ for the matching input and $\mathcal{H}_\ell = 0$ for every other input. Repeating this construction for all eight configurations ℓ, we define the Hamiltonian:

$$H_G(A, A_{\text{aux}}) = \sum_\ell \mathcal{H}_\ell$$

where H_ℓ is defined on G_ℓ by (3.8) and $\mathcal{H}_\ell$ by (3.9). Since every input A admits an auxiliary assignment with all $H_\ell = 0$, this gives $H_G(A, A_{\text{aux}}) = f(A)$, while any configuration with some $H_\ell > 0$ has total energy at least Δ. Therefore, H_G reproduces the values of any $f : \{-1, 1\}^3 \to \mathbb{R}$ below Δ. Each $\mathcal{H}_\ell$ contains only constant, one-body, and two-body terms, so their sum retains the form of (3.1) after combining these terms. Hence, H_G on $G = \bigcup_\ell G_\ell$ still satsifies Definition 3.3.

Further, recall Step 2. For any configuration ℓ, the corresponding graph $G_\ell = (V_\ell, E_\ell)$ has size (i) $|V_\ell| = |V_B| + |V_C| = 10 + 24 = 34$. (ii) $|E_\ell| = |E_B| + |E_C| + |E_{B,C}| = 5 + 8 \cdot 3 + 8 \cdot 3 = 53$.

Since the argument is identical for all ℓ, the size of the graph $G = (V, E) := \bigcup_\ell G_\ell$ is upper-bounded by $|V| \le 8 \cdot |V_\ell| = 272$ and $|E| \le 8 \cdot |E_\ell| = 424$. This completes the construction for $p = 3$. □

Full binary expansion and the UI construction both represent arbitrary functions of binary covariates. Their distinction lies in the feature structure: binary expansion uses products of observed covariates of all orders, while UI encodes higher-order effects through auxiliary variables and retains only unary and pairwise graph terms. Applying Theorem 3.1 to $f(A) = \mathbb{E}[B \mid A]$ gives the following corollary.

Corollary 3.1.1 (Universal Ising for Binary Regression)**.** There exists a Hamiltonian $H_G \in \mathcal{M}_{\text{UI}}$ defined on a graph $G = (V, E)$ specified in Theorem 3.1 such that

$$\mathbb{E}[B|A] = H_G(A, A_{\text{aux}})$$

where the vertex set, the edge set and A_{aux} are identical to those specified in Theorem 3.1. Furthermore, the equality holds with $|H_G|(A, A_{\text{aux}}) \le 1$ for some $J_{i,j}, r_i, K_0 \in \mathbb{R}$.

Table 1: Reported probability RMSE averaged over total 17 values of different coupling strengths $\kappa \in \{0, 0.25, 0.5, \ldots, 4\}$, and 50 independent repetitions for each sample size n.

Sparse third-order RMSE (4.1) (↓)					
Method	$n = 100$	$n = 300$	$n = 1{,}000$	$n = 3{,}000$	$n = 10{,}000$
Universal Ising	0.3580	0.2908	**0.1561**	**0.0858**	**0.0468**
BELIEF [4]	0.3547	0.2867	0.1717	0.0916	0.0498
Pairwise logistic	0.3865	0.3767	0.3685	0.3622	0.3590
Sparse fourth-order RMSE (4.2) (↓)					
Method	$n = 100$	$n = 300$	$n = 1{,}000$	$n = 3{,}000$	$n = 10{,}000$
Universal Ising	0.4079	0.3065	**0.1179**	**0.0545**	**0.0299**
BELIEF [4]	0.3754	0.2784	0.1265	0.0602	0.0327
Pairwise logistic	0.4311	0.4204	0.4115	0.4064	0.4039

Proof. Please see Section E.2 for a detailed proof. □

3.3 Conditional Independence Encoded on Universal Ising Graphs

Here, we highlight that our result using universal Ising representation is constructive: we specify the augmented graph and give explicit bounds on its numbers of vertices and edges. This graph determines the unary and pairwise features through which we represent the binary conditional mean.

Definition 3.4 (Universal Ising Vector)**.** Fix the augmented graph $G = (V, E)$ from Theorem 3.1 and a valid auxiliary-state selector α. Write its $N = |V|$ binary variables as $A_1, \ldots, A_N$, where the first p variables form A and the remaining variables form $\alpha(A)$. We define *universal Ising vector* by

$$A^\dagger(A) = \big(1, A_1, \ldots, A_N, (A_i A_j)_{(i,j)\in E}\big)^\top \in \{-1, 1\}^k, \qquad k = 1 + N + |E|.$$

With Definition 3.4, we highlight that the conditional mean determines the entire conditional distribution under the universal Ising representation and its corresponding graph structure:

Theorem 3.2 (Conditional Sufficiency of Universal Ising Representation)**.** Let $B \in \{-1, 1\}$, and suppose that $\mathbb{E}[B \mid A] = \theta_0^\top A^\dagger(A)$ for some $\theta_0 \in \mathbb{R}^k$. Then, $B \perp A \mid \theta_0^\top A^\dagger(A)$.

Proof. Please see Section E.3 for a detailed proof. □

Remark 3.1. Theorem 3.2 shows how two-body interaction terms govern the conditional independence of binary variables. While both universal Ising and binary expansion-based framework seek to analyze linearity in binary regression, Theorem 3.2 highlights the key difference between them.

To complement these structural results, we next establish the consistency of least-squares estimation.

Theorem 3.3 (Consistency for Universal Ising Representation)**.** Let $\{A_i, B_i\}_{i=1}^n$ be i.i.d. binary observations. Define a fixed universal Ising vector $A^\dagger(A) \in \{-1, 1\}^k$ with a fixed k such that

$$\mathbb{E}[B|A] = \theta_0^\top A^\dagger$$

for some $\theta_0 \in \mathbb{R}^k$. Further, define the squared-error loss and its set of minimizers:

$$\mathcal{L}(\theta) = \mathbb{E}\big[(B - \theta^\top A^\dagger)^2\big], \quad \Theta_0 = \operatorname*{argmin}_{\theta\in\mathbb{R}^k} \mathcal{L}(\theta)$$

Then, any least-squares estimator $\widehat{\theta}_n$ satisfies $\inf_{\theta^*\in\Theta_0} \|\widehat{\theta}_n - \theta^*\|_2 \xrightarrow{a.s.} 0$ as $n \to \infty$, where

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^n \big(B_i - \theta^\top A^\dagger(A_i)\big)^2.$$

Proof. Please see Section E.4 for a detailed proof. □

Building on the exact universality, Theorem 3.3 guarantees that least-squares estimation asymptotically reproduces the true conditional mean for a fixed universal Ising representation. By establishing convergence to the set of coefficients realizing this mean, the result allows nonunique coefficients.

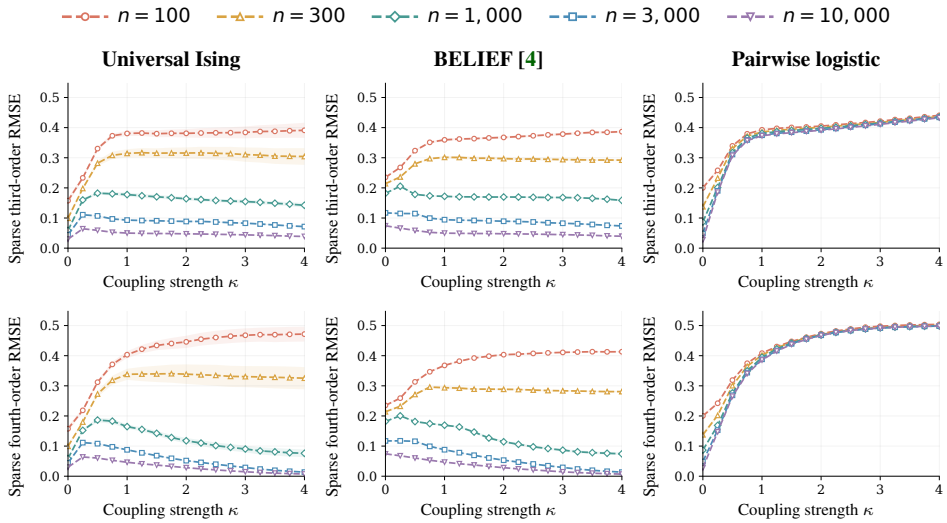


Figure 2: Probability RMSE across coupling strengths κ. Colors indicate sample size n. Curves show means over 50 repetitions, and the shaded bands show pointwise 95% confidence intervals.

4 Proof-of-Concept Experiments

In this section, we conduct synthetic experiments to evaluate the universal Ising representation for binary regression in practice. Previously, Corollary 3.1.1 establishes the existence of a two-body Hamiltonian on an augmented Ising graph that expresses the conditional mean $\mathbb{E}[B|A]$ as a linear combination of spin variables and their pairwise products, while Theorem 3.3 further guarantees that the least-squares estimators converge in probability to the set of population squared-error minimizers for a fixed universal Ising representation $A^\dagger(A)$. Specifically, a given input A may admit multiple valid configurations of A_{aux}. We therefore specify a deterministic selection rule before fitting to choose one valid auxiliary configuration for each input and define the fixed feature map $A^\dagger(A)$[3].

Setup. We consider two synthetic binary regression models with sparse third- and fourth-order logit interactions. Their conditional probabilities $\Pr(B = 1 \mid A)$ take the forms in (4.1) and (4.2), respectively, where $\sigma(t) = (1 + e^{-t})^{-1}$. We independently sample $p = 8$ covariates uniformly from $\{-1, 1\}$. These settings test whether the fixed universal Ising construction can learn higher-order relationships without prior knowledge of the active interaction terms. Specifically, we vary the coupling $\kappa \in \{0, 0.25, \ldots, 4\}$ to control interaction strength, and report $n \in \{100, 300, 1{,}000, 3{,}000, 10{,}000\}$ to examine the effect of sample size on estimation accuracy. We conduct 50 repetitions for each configuration and randomly permute the covariate indices in each repetition. We evaluate and report the probability root mean squared error (RMSE) exactly over all 256 possible inputs in Table 1.

$$\sigma\big(A_7 + \kappa(3A_1A_2A_3 - 2.5A_4A_5A_6)\big). \quad (4.1) \qquad \sigma\big(A_7 + \kappa(3A_1A_2A_3A_4 - 1.5A_5A_6)\big). \quad (4.2)$$

Interpretations. Figure 2 compares universal Ising, binary expansion linear effect (BELIEF) [4], and pairwise logistic regression across interaction strengths κ and sample sizes n. As κ increases, pairwise logistic regression exhibits increasing RMSE, and its curves for different sample sizes remain close under strong interactions. In comparison, universal Ising exhibits a clear separation across sample sizes, where increasing n reduces RMSE at each κ. Notably, larger κ amplifies the higher-order terms, making accurate prediction more dependent on capturing these interactions. Universal Ising accommodates these relationships through its auxiliary representation and attains low RMSE with sufficiently many observations, even under strong interactions. It achieves its clearest improvements over BELIEF at weak interaction strengths, while the two methods attain similar accuracy under strong interactions at larger sample sizes. Averaged over the examined coupling strengths, universal Ising achieves the lowest RMSE among the three methods in both settings for $n \geq 1{,}000$ (Table 1).

We further study the approximation of random Boolean functions in Section H. We defer **concluding remarks**, **related work**, and **additional numerical validations** to Sections A, B and I, respectively.

[3]We fix a literal ordering before fitting and select the first satisfied literal in each clause during both training and prediction.

References

[1] Gemma De las Cuevas and Toby S Cubitt. Simple universal models capture all classical spin physics. *Science*, 351(6278):1180–1183, 2016.

[2] Andrew Lucas. Ising formulations of many np problems. *Frontiers in physics*, 2:5, 2014.

[3] Peter McCullagh. *Generalized linear models*. Routledge, 2019.

[4] Benjamin Brown, Kai Zhang, and Xiao-Li Meng. Belief in dependence: leveraging atomic linearity in data bits for rethinking generalized linear models. *arXiv preprint arXiv:2210.10852*, 2022.

[5] Kai Zhang. Bet on independence. *Journal of the American Statistical Association*, 2019.

[6] I. G. Rosenberg. Reduction of bivalent maximization to the quadratic case. 1975.

[7] Hiroshi Ishikawa. Higher-order clique reduction in binary graph cut. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2993–3000. IEEE, 2009.

[8] Hiroshi Ishikawa. Transformation of general binary mrf minimization to the first-order case. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1234–1249, 2010.

[9] Fred Glover, Gary Kochenberger, and Yu Du. A tutorial on formulating and using qubo models. *arXiv preprint arXiv:1811.11538*, 2018.

[10] Martin Anthony, Endre Boros, Yves Crama, and Aritanan Gruber. Quadratic reformulations of nonlinear binary optimization problems. *Mathematical Programming*, 162:115–144, 2017.

[11] Michael R Garey and David S Johnson. *Computers and intractability*, volume 174. freeman San Francisco, 1979.

[12] Francisco Barahona. On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241, 1982.

[13] Saburou Muroga. *Threshold Logic and Its Applications*. John Wiley & Sons, 1971.

[14] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[15] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[16] Guido Montúfar. Deep narrow boltzmann machines are universal approximators. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1411.3784.

[17] Guido Montúfar and Nihat Ay. Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *Neural Computation*, 23(5):1306–1319, 2011. doi: 10.1162/NECO_a_00113. URL https://direct.mit.edu/neco/article/23/5/1306/7653/Refinements-of-Universal-Approximation-Results-for.

[18] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

[19] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[20] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 194–281. MIT Press, 1986.

[21] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[22] Ernst Ising. Contribution to the theory of ferromagnetism. *Z. Phys*, 31(1):253–258, 1925.

[23] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

[24] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[25] Sridevi Parise and Max Welling. Structure learning in markov random fields. *Advances in Neural Information Processing Systems*, 29:54, 2006.

[26] Jialei Wang and Mladen Kolar. Inference for sparse conditional precision matrices. *arXiv preprint arXiv:1412.7638*, 2014.

[27] Abhijith Jayakumar, Andrey Lokhov, Sidhant Misra, and Marc Vuffray. Learning of discrete graphical models with neural networks. *Advances in Neural Information Processing Systems*, 33:5610–5620, 2020.

[28] Marijtje AJ Van Duijn, Krista J Gile, and Mark S Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social networks*, 31(1):52–62, 2009.

[29] Yair Weiss and Judea Pearl. Belief propagation: technical perspective. *Communications of the ACM*, 53(10):94–94, 2010.

[30] Jonathan S Yedidia, William T Freeman, Yair Weiss, et al. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236–239): 0018–9448, 2003.

[31] Gary Kochenberger, Jin-Kao Hao, Fred Glover, Mark Lewis, Zhipeng Lü, Haibo Wang, and Yang Wang. The unconstrained binary quadratic programming problem: a survey. *Journal of combinatorial optimization*, 28:58–81, 2014.

[32] Masataka Sao, Hiroyuki Watanabe, Yuuichi Musha, and Akihiro Utsunomiya. Application of digital annealer for faster combinatorial optimization. *Fujitsu Scientific and Technical Journal*, 55(2):45–51, 2019.

[33] JUNLONG ZHAO and CHENLEI LENG. An analysis of penalized interaction models. *Bernoulli*, 22(3):1937–1961, 2016.

[34] Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

[35] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(1):3468–3497, 2009.

[36] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 91(434):576–583, 1996.

[37] Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[38] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, 2002.

[39] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[40] Christos H Papadimitriou. Computational complexity. In *Encyclopedia of computer science*, pages 260–265. 2003.

[41] Xinran Li and Xiao-Li Meng. A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533):353–367, 2021.

[42] Claude E Shannon. The synthesis of two-terminal switching circuits. *The Bell System Technical Journal*, 28(1):59–98, 1949.

Appendix

A Discussion and Conclusion **14**

B Related Work **14**

B.1 Limitation 15

C Supplementary Theoretical Backgrounds **15**

C.1 Faithful Reduction from SAT to Vertex Cover Problem 15

C.2 Faithful Reduction from Vertex Cover Problem to Ground State Energy Problem . . 16

D Proofs in Section 2 **18**

D.1 Proof of Theorem 2.1 18

D.2 Proof of Corollary 2.1.1 20

D.3 Proof of Corollary 2.1.2 20

E Proof of Section 3 **21**

E.1 Proof of Theorem 3.1 21

E.2 Proof of Corollary 3.1.1 23

E.3 Proof of Theorem 3.2 24

E.4 Proof of Theorem 3.3 24

F Generalization of Universal Ising to Non-Binary Regression **26**

F.1 Universal Ising for Non-Binary Regression 26

F.2 Example of Universal Ising for Discrete Case. 26

G Binary Expansion Linear Effect **28**

G.1 Preliminaries for Binary Linear Model 28

G.2 Binary Expansion Linear Effect 28

H Experimental Details **30**

I Additional Numerical Results **33**

Impact Statement

By the formal nature of this work, we do not expect immediate negative social impact.

LLM Usage Disclosure

We used large language models (LLMs) to aid and polish writing, such as improving clarity, grammar, and conciseness. We also used LLMs for retrieval and discovery, for example exhausting literature to identify potential missing related work. All technical content, proofs, experiments, and results are original contributions by the authors.

A Discussion and Conclusion

We introduce a general framework for modeling in binary regression that is distinct from common approaches such as GLMs. We demonstrate that a class of models, termed universal models, is capable of replicating any relationship between binary covariates and response expressible by existing models. Our main contribution is the constructive proof for such replication by encoding universality construction as a SAT problem and leveraging the structural equivalence among NP-complete problems (Theorem 2.1). For application, we showcase the universal model representation for GLMs and we address the issue of perfect separation (Corollary 2.1.1 and Corollary 2.1.2). Furthermore, we propose Universal Ising as a concrete example of universality framework, showcasing the existence of a linear model class with pairwise interactions in binary regression (Theorem 3.1 and Corollary 3.1.1). In addition, we characterize the conditional independence structure for Universal Ising on the graph constructed through the establishment of universality (Theorem 3.2).

Extension to Non-Binary Case. Please refer to Section F for a detailed discussion.

B Related Work

Universal Binary Models. Universal Ising builds on established paradigms of universal binary function models while preserving interpretability. Classic threshold circuits — networks of McCulloch-Pitts neurons [13, 14] — can represent any Boolean function given enough depth or hidden units [15]. Similarly, energy-based models such as Boltzmann machines can capture intricate binary patterns using latent variables, with restricted Boltzmann machines (RBMs) proven as universal approximators for binary data distributions [16–20]. While powerful, threshold circuits and RBMs often lack transparency. In contrast, Universal Ising achieves universality through pairwise (two-body) interactions, where each learned parameter connects only two variables. This design mirrors the Ising model from statistical physics [2, 21, 22], but Universal Ising extends the Ising framework with auxiliary variables, retaining two-body interpretability while matching the expressive capacity of deeper models.

Graph Recovery. Graph recovery aims to infer the underlying structure of a graphical model from observed data. Universal Ising provides a Ising-like graph structure that opens a new avenue to incorporate existing results in this field. Classical approaches, such as graphical lasso [23, 24] and structure learning for Markov random fields [25], estimate sparse precision matrices and conditional dependencies between variables [26]. More recent advancements incorporate neural networks and variational inference to enhance scalability and robustness [27]. In the context of Universal Ising with Ising graph, graph recovery involves estimating interaction strengths between binary variables, often using maximum likelihood estimation, pseudolikelihood methods [28], and belief propagation [29, 30].

Optimization Theory. Universal Ising extends to combinatorial-optimization methods that convert higher-order binary objectives into equivalent pairwise models [2]. In particular, the Quadratic Unconstrained Binary Optimization (QUBO) formulation applies the similar principle by seeking the binary vector that optimizes a quadratic cost composed of pairwise products [9, 10, 31]. Because of this structural similarity, Universal Ising engages branch-and-bound, roof-duality, semidefinite relaxation, large-neighborhood search, and specialized annealing hardware built for QUBO instances [32]. In sum, Universal Ising is not a standalone theory; integration with traditional algorithms and cutting-edge hardware places it within the broader combinatorial-optimization ecosystem.

Interpretable Statistical Models. In classical statistics, generalized linear models (GLMs) such as logistic regression remain popular for their interpretability but typically include only linear terms [3]. Complex interactions require manually engineered product features, leading to a combinatorial explosion in the number of terms . Hierarchical interaction approaches address this by adding penalty/priors so that a higher-order interaction is included only if lower-order terms are present [33–36]. While such methods improve interpretability, they are limited by a fixed set of possible interactions. By contrast, Universal Ising endogenously learns auxiliary binary variables that effectively capture higher-order dependencies through pairwise factors, achieving universal approximation without enumerating all interactions and preserving a clear "factor graph" structure.

Perfect Separation. A secondary consideration arises around perfect separation in logistic regression: if a linear boundary perfectly splits the classes, the unpenalized logistic MLE diverges [37].

Although this issue is well-known and addressed via penalization or Bayesian priors [38], Universal Ising inherently sidesteps it by not relying on an unbounded linear coefficient. However, we emphasize that Universal Ising's principal advantage is its universal approximation combined with a pairwise-interpretable formulation, bridging the gap between classic interpretable models and more flexible—but often opaque—function approximators.

Linearity via Binary Expansion. Prior works addressing direct linearity in binary regression predominantly rely on the *binary expansion* approach [4, 5]. Binary expansion maps $A \in \{-1, 1\}^p$ to the power vector $A^{\otimes}$ containing all interaction terms among observed covariates. Since these features form a basis for functions on $\{-1, 1\}^p$, any binary-input conditional mean can be written as

$$\mathbb{E}[B|A] = \beta^\top A^{\otimes}$$

for a suitable β. This representation is exact and avoids the GLM link, but it introduces explicit high-order product features up to order p. Universal Ising takes a different route: it trades explicit high-order observed features for auxiliary variables and pairwise graph factors. Please see Section G for details.

Auxiliary-Variable Pairwise Reductions. Auxiliary-variable reductions from higher-order binary objectives to pairwise forms are classical in pseudo-Boolean optimization, MRF inference, QUBO, and Ising formulations [6–8, 2, 9, 10]. We do not claim that such reductions are new in isolation. Our contribution is to use faithful-reduction constructions to formulate a binary-regression representation theorem and compare it with GLM and binary-expansion viewpoints.

B.1 Limitation

This paper is primarily about representation. Theorems in this work establish that binary conditional-mean functions admit constructive Hamiltonian representations with pairwise interactions after augmenting the binary state space. They do not imply that the resulting model can be learned efficiently from finite samples, that the representation is identifiable, that the auxiliary variables have direct semantic meaning, or that uncertainty quantification is immediately available. The proof-of-concept experiments are designed to validate expressivity in small truth-table settings. Scalable estimation, sample complexity, parameter recovery, model selection, and uncertainty quantification are important directions for future work.

Although Universal Ising achieves interpretability via linear effects and reduces interaction complexity, scaling to large models demands substantial optimization because the number of auxiliary variables grows with problem size. A direction for future work is to sharpen the reduction between computational problems: devising a more efficient, faithful reduction from SAT to GSE would lower the upper bound on auxiliary covariates, thereby improving scalability without compromising interpretability.

C Supplementary Theoretical Backgrounds

In this section, we detail the faithful reduction applied in the proof of Theorem 3.1 for special case $p = 3$. Section C.1 revisits the one-to-one correspondence between solutions in SAT and vertex cover problem. Section C.2 extends it to vertex cover problem and the ground state energy problem.

C.1 Faithful Reduction from SAT to Vertex Cover Problem

Recall the Boolean formula $\phi_\ell(x_1, x_2, x_3; F_\ell, u_\ell)$ constructed in the proof of Theorem 3.1. It has five Boolean variables and eight clauses. Every input $x \in \{0, 1\}^3$ admits a satisfying extension, and every satisfying assignment has $F_\ell = 1$ if and only if $x = \ell$.

We construct the graph $G_\ell = (V_\ell, E_\ell)$ with $V_\ell = V_B \cup V_C$ and $E_\ell = E_B \cup E_C \cup E_{B,C}$ as follows.

Construction of the Vertex Set. We introduce two vertices per Boolean variable:

$$V_B = \bigcup_{i=1}^{5} \{v_i, \bar{v}_i\}.$$

The mapping h_1 sends x_i to v_i and $\neg x_i$ to $\overline{v}_i$ for $i = 1, 2, 3$, sends F_ℓ and $\neg F_\ell$ to v_4 and $\overline{v}_4$, and sends u_ℓ and $\neg u_\ell$ to v_5 and $\overline{v}_5$, respectively. We also introduce three vertices per clause:

$$V_C = \bigcup_{j=1}^{8} \{c_j^1, c_j^2, c_j^3\}.$$

Here, $h_2^{-1}(c_j^k)$ denotes the literal appearing at position k in Clause j. Thus, $|V_\ell| = 10 + 24 = 34$.

Construction of the Edge Set. We connect each variable vertex to its negation, form a triangle for each clause, and connect each clause vertex to the variable vertex representing the same literal:

$$E_B = \bigcup_{i=1}^{5} \Big\{ \{v_i, \overline{v}_i\} \Big\},$$

$$E_C = \bigcup_{j=1}^{8} \Big\{ \{c_j^1, c_j^2\}, \{c_j^1, c_j^3\}, \{c_j^2, c_j^3\} \Big\},$$

$$E_{B,C} = \bigcup_{j=1}^{8} \bigcup_{k=1}^{3} \Big\{ \{c_j^k, a\} \mid a \in V_B,\ h_1^{-1}(a) = h_2^{-1}(c_j^k) \Big\}.$$

This gives $|E_\ell| = 5 + 24 + 24 = 53$, as illustrated in Figure 1.

Correspondence between Satisfying Assignments and Minimum Vertex Covers. An independent set contains at most one vertex from each of the five variable pairs and at most one vertex from each of the eight clause triangles. Its cardinality is therefore at most $5 + 8 = 13$.

Given any satisfying assignment to ϕ_ℓ, select the vertex representing the false literal from each variable pair and one vertex representing a true literal from each clause triangle. These 13 vertices form an independent set: each selected clause vertex is connected in V_B only to the vertex representing its true literal, which was not selected. Hence the maximum independent-set size is 13, and the minimum vertex-cover size is $34 - 13 = 21$.

Conversely, every independent set of size 13 selects exactly one vertex from each variable pair and each clause triangle. Assign each Boolean variable so that its selected variable vertex represents the false literal. Each selected clause vertex then represents a true literal, since its corresponding variable vertex cannot also be selected. Thus every clause is satisfied.

Taking complements gives the correspondence with minimum vertex covers. Every satisfying assignment extends to a minimum vertex cover, and every minimum vertex cover determines a satisfying assignment through its restriction to V_B. The restriction contains exactly the variable vertices representing true literals, so it preserves the input and flag values. The full vertex cover need not be unique, since a clause may contain more than one true literal.

C.2 Faithful Reduction from Vertex Cover Problem to Ground State Energy Problem

Recall (3.7):

$$H_\ell(s) = -\sum_{i \in V_\ell} s_i + 2 \sum_{(i,j) \in E_\ell} s_i s_j + K_0,$$

where $s_i \in \{0, 1\}$ indicates whether vertex i is selected. Suppose the selected vertices contain an edge. Choose a selected vertex with $d \geq 1$ selected neighbors. Changing its value from 1 to 0 changes the energy by $1 - 2d < 0$. Therefore, every ground state represents an independent set. On an independent set I, the energy equals $K_0 - |I|$, so the ground states correspond exactly to maximum independent sets. Taking complements gives a bijection with minimum vertex covers. By Section C.1, the maximum independent-set size is 13. Setting $K_0 = 13$ therefore makes H_ℓ nonnegative and

integer-valued, with minimum zero. The change of variables $A_i = 1 - 2s_i$ is bijective and gives

$$\begin{aligned} H_\ell(A) = & -\frac{|V_\ell|}{2} + \frac{1}{2}\sum_{i\in V_\ell} A_i + \frac{|E_\ell|}{2} + \frac{1}{2}\sum_{(i,j)\in E_\ell} A_i A_j + K_0 \\ & - \frac{1}{2}\sum_{(i,j)\in E_\ell} (A_i + A_j), \end{aligned}$$

which is precisely (3.8). This transformation preserves all energy values and the correspondence between ground states and minimum vertex covers. Under the indexing $v_i \mapsto s_{2i-1}$ and $\overline{v}_i \mapsto s_{2i}$ for $i = 1, \ldots, 5$, every zero-energy configuration satisfies $x_i = (A_{2i-1} + 1)/2$ for $i = 1, 2, 3$ and $A_7 = 2F_\ell - 1$. Every input admits an extension to the remaining spins with $H_\ell(A) = 0$. In every such extension, $A_7 = 1$ if and only if the input matches ℓ.

D Proofs in Section 2

In this section, we present a detailed construction of universality in statistical modeling and its application. Specifically, we begin with the proof of the universality-based binary regression in Section D.1. Then, we prove its application regarding the reformulation of GLMs in Section D.2. Finally, we show the universality-based model address the perfect separation issue in Section D.3.

D.1 Proof of Theorem 2.1

This section presents the formal proof of Theorem 2.1.

Theorem D.1 (Theorem 2.1 Restated: Binary Regression via Universality)**.** Let $N \in \mathbb{N}$. Let $\mathcal{M}_{\text{uni}} = \{H_\theta\}_\theta$ be a family of models parametrized by θ. Suppose $\mathcal{M}_{\text{uni}}$ is closed and its GSE problem admits a polynomial-time faithful reduction from SAT. Then, for any $A = (A_1, \ldots, A_p)^\top \in \{-1, 1\}^p$ and $B \in \{-1, 1\}$, there exists a Hamiltonian $H_\theta \in \mathcal{M}_{\text{uni}}$ and a binary vector $A_{\text{aux}} \in \{-1, 1\}^N$ such that:

$$\mathbb{E}[B|A] = H_\theta(A, A_{\text{aux}}), \tag{D.1}$$

where H_θ and A_{aux} are dependent on A, and N is determined by the structure of the faithful reduction specific to the model class $\mathcal{M}_{\text{uni}}$. Furthermore, (D.1) holds with $|H_\theta(A, A_{\text{aux}})| \leq 1$ for all A.

Proof. Our proof builds upon [1]. We prove the theorem with the following three steps.

Step 1: Initial Decomposition with Probit Model. We introduce a function as an intermediary.

Without loss of generality, we use the probit model:

$$\mathbb{E}[B|A] := f(A) = 2\Phi(\beta^\top A) - 1,$$

where $\beta \in \mathbb{R}^p$ is the model parameters and Φ is the cumulative distribution function of the standard normal distribution.

Let $\mathcal{F}_{\text{energy}}$ denote the energy spectrum of $f(A)$, and let $\ell = [2^p]$ be an index set that denotes 2^p possible assignments to A. Next, we decompose $f(A)$ by:

$$f(A) = \sum_{\ell=1}^{2^p} \left(f^\ell \cdot e_\ell(A)\right),$$

where $f^\ell \in \mathcal{F}_{\text{energy}}$ takes the ℓ-th configuration of A as input. For simplicity, we denote "the ℓ-th configuration of A" by ℓ, with the understanding that it is an index set for configurations of A.

Step 2: Encoding Desired Solutions in SAT

We encode the one-to-one correspondence between covariates A in the target model and Boolean variables $x = (x_1, \ldots, x_p)$ with $x_i = (A_i + 1)/2 \in \{0, 1\}$ for $i \in [p]$.

- **Step 2.1: Construction of Flag Covariate SAT**

  We introduce an auxiliary Boolean variable $F^\ell_{\text{flag}} \in \{0, 1\}$ termed *flag covariate*. We leverage the states of F^ℓ_{flag} as an indication of whether the input A has the same configuration as ℓ by constructing the following SAT:

  $$\phi^1_\ell(x; F^\ell_{\text{flag}}) = e_\ell(A) \odot F^\ell_{\text{flag}}.$$

  In other words, a solution satisfies ϕ_ℓ with $A = \ell$ if and only if $F^\ell_{\text{flag}} = 1$.

  By definition, there exists a faithful reduction from $\phi^1_\ell(x; F^\ell_{\text{flag}})$ to a GSE problem of a model $M_{\text{uni}} \in \mathcal{M}_{\text{uni}}$, where elements in the solution sets are identified via a bijective mapping $\psi_1 : x_i \to$

$A_i \in \{-1, 1\}$; $F^\ell_{\text{flag}} \to A^\ell_{\text{flag}} \in \{-1, 1\}$, with additional covariates constructed during the faithful reduction, denoted by $A^1_{\text{aux}} \in \{-1, 1\}^{N_1}$.

The GSE problem is characterized by Hamiltonian H^ℓ_1 acting on binary covariates $A \cup A^\ell_{\text{flag}} \cup A^1_{\text{aux}}$:

$$\begin{cases} H^\ell_1 = 0 & e_\ell(A) = \psi_1^{-1}(A^\ell_{\text{flag}}), A^1_{\text{aux}} \in \Sigma_1 \\ H^\ell_1 \geq 1 & \text{otherwise,} \end{cases} \tag{D.2}$$

where Σ_1 is a non-empty set of configurations of A^1_{aux}.

- **Step 2.2: Formulating States of the Flag Covariate**

  We construct two SAT with Boolean formulas ϕ^2_ℓ and ϕ^3_ℓ to further control the states of F^ℓ_{falg}.

  SAT with Boolean Formula ϕ^2_ℓ.

  We define $\phi^2_\ell = F_\ell \odot 0$. We perform another faithful reduction from ϕ^2_ℓ to a GSE problem of M_{uni} via bijective mapping $\psi_2 : F^\ell_{\text{flag}} \to B^\ell_{\text{flag}} \in \{-1, 1\}$ with additional covariates $A^2_{\text{aux}} \in \{-1, 1\}^{N_2}$ constructed during the faithful reduction.

  Then, the GSE problem is characterized by Hamiltonian H^ℓ_2 acting on $B^\ell_{\text{flag}} \cup A^2_{\text{aux}}$ such that:

$$\begin{cases} H^\ell_2 = 0 & \psi_2^{-1}(B^\ell_{\text{flag}}) = 0, A^2_{\text{aux}} \in \Sigma_2 \\ H^\ell_2 \geq 1 & \text{otherwise,} \end{cases} \tag{D.3}$$

  where Σ_2 is a non-empty set of configurations of A^2_{aux}. Further, we define $a^\star$ as an arbitrary configuration of A^2_{aux}, and let $\kappa_\ell \coloneqq H^\ell_2(\psi_2^{-1}(B^\ell_{\text{flag}}) = 1, A^2_{\text{aux}} = a^\star) \geq 1$.

  SAT with Boolean Formula ϕ^3_ℓ.

  Building upon ϕ^2_ℓ, we construct the last SAT with $\phi^3_\ell = \big((F_\ell \odot 1) \wedge (A^2_{\text{aux}} = a^\star)\big) \vee (F_\ell \odot 0)$ to further control the states of A^2_{aux} when $F_\ell = 1$. We perform the other faithful reduction from ϕ^3_ℓ to a GSE problem of M_{uni} via bijective mapping $\psi_3 : F^\ell_{\text{flag}} \to C^\ell_{\text{flag}} \in \{-1, 1\}$ with additional covariates $A^3_{\text{aux}} \in \{-1, 1\}^{N_3}$.

  Then, the GSE problem is characterized by Hamiltonian H^ℓ_3 acting on $C^\ell_{\text{flag}} \cup A^3_{\text{aux}}$:

$$\begin{cases} H^\ell_3 = 0 & \psi_3^{-1}(C^\ell_{\text{flag}}) = 1, A^2_{\text{aux}} = a^\star, A^3_{\text{aux}} \in \Sigma_3 \\ H^\ell_3 = 0 & \psi_3^{-1}(C^\ell_{\text{flag}}) = 0, A^3_{\text{aux}} \in \Sigma'_3 \\ H^\ell_3 \geq 1 & \text{otherwise,} \end{cases} \tag{D.4}$$

  where $\Sigma_3 \cup \Sigma'_3$ is a non-empty set of configurations of A^3_{aux}.

Step 3: Leveraging Solutions to GSE Problem

Recall that the closure property (Definition 2.7) ensures that for any two models in the class, their sum belongs to the class as well. This property allow us to construct a Hamiltonian by the linear combination of H^ℓ_1, H^ℓ_2 and H^ℓ_3. Specifically, we define a Hamiltonian H_θ for model $\mathcal{M}_{\text{uni}} \in \mathcal{M}_{\text{uni}}$:

$$\begin{aligned} & H_\theta(A, A^\ell_{\text{flag}}, B^\ell_{\text{flag}}, C^\ell_{\text{flag}}, A^1_{\text{aux}}, A^2_{\text{aux}}, A^3_{\text{aux}}) \\ \coloneqq & \underbrace{\left(\sum_{\ell=1}^{2^p} H^\ell_1(A, A^\ell_{\text{flag}}, A^1_{\text{aux}})\right) \cdot \Delta}_{(T_1)} && \text{(From (D.2))} \\ & + \underbrace{\sum_{\ell=1}^{2^p} \left(\frac{1}{\kappa_\ell} H^\ell_2(B^\ell_{\text{flag}}, A^2_{\text{aux}})\right) \cdot f^\ell}_{(T_2)} && \text{(From (D.3))} \end{aligned}$$

$$+ \underbrace{\left(\sum_{\ell=1}^{2^p} H_3^\ell(C_{\text{flag}}^\ell, A_{\text{aux}}^2, A_{\text{aux}}^3) \right) \cdot \Delta}_{(T_3)}, \qquad \text{(From (D.4))}$$

where $\Delta \coloneqq \max_A \mathbb{E}[B|A]$.

Given any configuration ℓ with output value f^ℓ, we discuss four scenarios for terms in the summation:

- **Scenario 1:** $e_\ell(A) \neq F_{\text{flag}}^\ell$**.**

  In this case, the flag covariate A_{flag}^ℓ incorrectly signifies whether $A = \ell$ or $A \neq \ell$.

  Therefore, every such term outputs $H_1^\ell > \Delta$ in (T_1).

- **Scenario 2:** $e_\ell(A) = F_{\text{flag}}^\ell$ **and** $F_{\text{flag}}^\ell = 0$**.**

  In this case, A_{flag}^ℓ correctly signifies that $A \neq \ell$, and we have $\psi_2^{-1}(B_{\text{flag}}^\ell) = 0$. Therefore, every such term outputs $H_1^\ell = 0$ in (T_1), outputs $H_2^\ell = 0$ in (T_2) and outputs $H_3^\ell = 0$ in (T_3).

- **Scenario 3:** $e_\ell(A) = F_{\text{flag}}^\ell$**,** $F_{\text{flag}}^\ell = 1$ **and** $A_{\text{aux}}^2 \neq a^\star$**.**

  In this case, A_{flag}^ℓ correctly signifies that $A = \ell$, giving $\psi_2^{-1}(B_{\text{flag}}^\ell) = 1$ and $\psi_3^{-1}(C_{\text{flag}}^\ell) = 1$.

  Since $A_{\text{aux}}^2 \neq a^\star$, every such term outputs $H_1^\ell = 0$ in (T_1), outputs $H_2^\ell = 0$ in (T_2) and outputs $H_3^\ell > \Delta$ in (T_3).

- **Scenario 4:** $e_\ell(A) = F_{\text{flag}}^\ell$**,** $F_{\text{flag}}^\ell = 1$ **and** $A_{\text{aux}}^2 = a^\star$**.**

  In this case, A_{flag}^ℓ correctly signifies that $A \neq \ell$, and we have both $\psi_2^{-1}(B_{\text{flag}}^\ell) = 1$ and $\psi_3^{-1}(C_{\text{flag}}^\ell) = 1$. Since $A_{\text{aux}}^2 = a^\star$, every such term outputs $H_1^\ell = 0$ in (T_1), outputs $H_2^\ell = \kappa_\ell$ in (T_2) and outputs $H_3^\ell = 0$ in (T_3).

We set $A_{\text{aux}} \coloneqq A_{\text{flag}}^\ell \cup B_{\text{flag}}^\ell \cup C_{\text{flag}}^\ell \cup A_{\text{aux}}^1 \cup A_{\text{aux}}^2 \cup A_{\text{aux}}^3 \in \{-1, 1\}^N$ with $N = N_1 + N_2 + N_3 + 3$. Combining all four scenarios, we conclude that for all $H(A, A_{\text{aux}}) < \Delta$, $H(A, A_{\text{aux}}) = f(A)$.

This completes the proof. □

D.2 Proof of Corollary 2.1.1

We now present the main proof of Corollary 2.1.1.

Corollary D.1.1 (Corollary 2.1.1 Restated: Rethinking GLMs through Universality)**.** Let A_{aux} and $\mathcal{M}_{\text{uni}} = \{H_\theta\}_\theta$ be a binary vector and a family of models defined in Theorem 2.1. Let $A \in \{-1, 1\}^p$ and $B \in \{-1, 1\}$. Then, for any GLMs with link function $g : (-1, 1) \to \mathbb{R}$ and parameter $\gamma \in \mathbb{R}^p$, there exist a H_θ such that:

$$\mathbb{E}[B|A] = g^{-1}(\gamma^\top A) = H_\theta(A, A_{\text{aux}}),$$

where the dimension of θ and A_{aux} are determined by the faithful reduction specific to $\mathcal{M}_{\text{uni}}$.

Proof. By considering $f = g^{-1}(\gamma^\top A^\otimes)$ in the proof of Theorem 2.1, this completes the proof. □

D.3 Proof of Corollary 2.1.2

We now present the main proof of Corollary 2.1.2.

Corollary D.1.2 (Corollary 2.1.2 Restated: Addressing Perfect Separation via Universality)**.** Let A_{aux} and $\mathcal{M}_{\text{uni}} = \{H_\theta\}_\theta$ be a binary vector and a family of models defined in Theorem 2.1. Let $B \in \{-1, 1\}$ be the response and $A \in \{-1, 1\}^p$ be covariates. Let $\{B_i, A_i\}_{i=1}^n$ be a set of sample.

Then, there exists a H_θ such that:

$$B_i = H_\theta(A_i, A_{\text{aux}}), \quad \text{for all} \quad i \in [n],$$

where the dimension of θ and A_{aux} are determined by the faithful reduction specific to $\mathcal{M}_{\text{uni}}$.

Proof. Given a set of sample $\{B_i, A_i\}_{i=1}^n$, we define a function $B_i = \Psi(A_i)$ for all $i \in [n]$, where Ψ is the mapping determined by the sample. Next, we extend the domain of Ψ to all configurations $A \in \{-1, 1\}^p$ by setting $\Psi(A) = 0$ for any A not observed in the sample. Then, we consider the target function $f = \Psi(A)$ in Step 1 of the proof of Theorem 2.1. This completes the proof. □

E Proof of Section 3

We prove the UI universality result in Section E.1 and its binary regression corollary in Section E.2. We then establish conditional sufficiency and least-squares consistency in Section E.3 and Section E.4.

E.1 Proof of Theorem 3.1

Here, we extend the proof from special case $p = 3$ shown in the main text to arbitrary $p \in \mathbb{N}$.

Theorem E.1 (Theorem 3.1 Restated, Universal Ising Representation)**.** Let $A = (A_1, \ldots, A_p)^\top \in \{-1, 1\}^p$. Then, for any function $f : A \to \mathbb{R}$, there exists a $A_{\text{aux}} = (A_{p+1}, \ldots, A_{|V|})^\top \in \{-1, 1\}^{|V|-p}$ dependent on A, and Hamiltonian $H_G \in \mathcal{M}_{\text{UI}}$ fixed across all inputs A defined on a graph $G = (V, E)$ such that

$$f(A) = H_G(A, A_{\text{aux}}),$$

where $V = \{A_1, \ldots, A_p, A_{p+1}, \ldots A_{|V|}\}$. Furthermore, the cardinality of V and E are bounded by

$$|V| \le 2^p \left(7p - 5 + 9\lceil p/2 \rceil\right), \quad |E| \le 2^p \left(8p - 7 + 18\lceil p/2 \rceil\right).$$

Proof. For $p = 1$, any $f : \{-1, 1\} \to \mathbb{R}$ is represented by the following without auxiliary variables:

$$H_G(A) = \frac{f(1) + f(-1)}{2} + \left(\frac{f(1) - f(-1)}{2}\right) \cdot A,$$

where $A \in \{-1, 1\}$. We then extend the construction for any $p \ge 2$ with the following two steps.

Step 1: Boolean Formula for general $p \ge 2$. For each target configuration ℓ, we construct a formula ϕ_ℓ in 3CNF with flag covariate $F_\ell \in \{0, 1\}$ and additional Boolean variables as needed. Every input x admits a satisfying extension, and every satisfying assignment has $F_\ell = 1$ if and only if $x = \ell$.

- **Scenario 1: p is an even number** Let $M = p/2$ and fix a target configuration $\ell = (x_1, \ldots, x_p)$, we construct the following SAT with auxiliary Boolean variables $u_{\ell,1}, \ldots, u_{\ell,p-2}$ for $p \ge 4$:

$$\begin{aligned}
&\phi_\ell^{\text{even}}(x_1, x_2, \ldots, x_p; F_\ell; u_{\ell,1}, u_{\ell,2}, \ldots, u_{\ell,p-2}) \\
&= \underbrace{(\xi_1 \vee \eta_1 \vee u_{\ell,1})}_{\text{Clause } 1} \\
&\quad \wedge \bigwedge_{j=2}^{M-1} \Big[\underbrace{(\neg u_{\ell,2j-3} \vee \xi_j \vee u_{\ell,2j-2})}_{\text{Clause } (2j-2)} \wedge \underbrace{(\neg u_{\ell,2j-2} \vee \eta_j \vee u_{\ell,2j-1})}_{\text{Clause } (2j-1)} \Big] \\
&\quad \wedge \underbrace{(\neg u_{\ell,p-3} \vee \xi_M \vee u_{\ell,p-2})}_{\text{Clause } (p-2)} \wedge \underbrace{(\neg u_{\ell,p-2} \vee \eta_M \vee F_\ell)}_{\text{Clause } (p-1)} \\
&\quad \wedge \bigwedge_{j=1}^{M} \Big[\underbrace{(\xi_{j.1} \vee \eta_{j.1} \vee \neg F_\ell)}_{\text{Clause } j.1} \wedge \underbrace{(\xi_{j.2} \vee \eta_{j.2} \vee \neg F_\ell)}_{\text{Clause } j.2} \wedge \underbrace{(\xi_{j.3} \vee \eta_{j.3} \vee \neg F_\ell)}_{\text{Clause } j.3} \Big],
\end{aligned}$$

where an empty conjunction is taken to be true. For each $j = 1, \ldots, M$, ξ_j is a literal of x_{2j-1} and η_j is a literal of x_{2j}, each chosen to be false exactly when its input bit matches the corresponding bit of ℓ. The remaining literals are defined by

$$\underbrace{(\xi_{j.1}, \eta_{j.1}) = (\neg\xi_j, \eta_j)}_{\text{Clause } j.1}; \quad \underbrace{(\xi_{j.2}, \eta_{j.2}) = (\neg\xi_j, \neg\eta_j)}_{\text{Clause } j.2}; \quad \underbrace{(\xi_{j.3}, \eta_{j.3}) = (\xi_j, \neg\eta_j)}_{\text{Clause } j.3}.$$

For example, when $p = 4$ and $\ell = (1, 0, 0, 1)$, we choose $(\xi_1, \eta_1) = (\neg x_1, x_2)$ and $(\xi_2, \eta_2) = (x_3, \neg x_4)$. For $p = 2$, we use $(\xi_1 \vee \eta_1 \vee F_\ell)$ together with Clauses 1.1, 1.2, and 1.3, without introducing any u variables. Clauses 1 to $p - 1$ force $F_\ell = 1$ when $x = \ell$, while Clauses $j.1$, $j.2$, and $j.3$ for $j = 1, \ldots, M$ force $x = \ell$ when $F_\ell = 1$. For any input x, choose $F_\ell = 1$ if $x = \ell$ and 0 otherwise; for $p \geq 4$, choose $u_{\ell,i} = 1$ if the first $i + 1$ bits of x match those of ℓ and 0 otherwise, for $i = 1, \ldots, p - 2$. These choices satisfy all clauses, so every input admits a satisfying extension.

- **Scenario 2: p is a odd number Scenario 2: p is an odd number.** Let $K = (p - 1)/2$ and $S = (p + 1)/2$. For odd $p \geq 3$, fix a target configuration $\ell \in \{0, 1\}^p$ and introduce auxiliary Boolean variables $u_{\ell,1}, \ldots, u_{\ell,p-2}$. We construct the following Boolean formula:

$$\begin{aligned}
&\phi_\ell^{\text{odd}}(x_1, x_2, \ldots, x_p; F_\ell; u_{\ell,1}, u_{\ell,2}, \ldots, u_{\ell,p-2}) \\
&= \underbrace{(\xi_1 \vee \eta_1 \vee u_{\ell,1})}_{\text{Clause } 1} \\
&\quad \wedge \bigwedge_{j=2}^{K} \Big[\underbrace{(\neg u_{\ell,2j-3} \vee \xi_j \vee u_{\ell,2j-2})}_{\text{Clause } (2j-2)} \wedge \underbrace{(\neg u_{\ell,2j-2} \vee \eta_j \vee u_{\ell,2j-1})}_{\text{Clause } (2j-1)} \Big] \\
&\quad \wedge \underbrace{(\neg u_{\ell,p-2} \vee \eta_S \vee F_\ell)}_{\text{Clause } (p-1)} \\
&\quad \wedge \bigwedge_{j=1}^{S} \Big[\underbrace{(\xi_{j.1} \vee \eta_{j.1} \vee \neg F_\ell)}_{\text{Clause } j.1} \wedge \underbrace{(\xi_{j.2} \vee \eta_{j.2} \vee \neg F_\ell)}_{\text{Clause } j.2} \wedge \underbrace{(\xi_{j.3} \vee \eta_{j.3} \vee \neg F_\ell)}_{\text{Clause } j.3} \Big],
\end{aligned}$$

where an empty conjunction is taken to be true. For each $j = 1, \ldots, K$, ξ_j is a literal of x_{2j-1} and η_j is a literal of x_{2j}, each chosen to be false exactly when its input bit matches the corresponding bit of ℓ. For the last pair, set $\xi_S = \xi_1$ and choose η_S as a literal of x_p that is false exactly when x_p matches the last bit of ℓ. For $j = 1, \ldots, S$, the remaining literals are defined by

$$\underbrace{(\xi_{j.1}, \eta_{j.1}) = (\neg\xi_j, \eta_j)}_{\text{Clause } j.1}; \quad \underbrace{(\xi_{j.2}, \eta_{j.2}) = (\neg\xi_j, \neg\eta_j)}_{\text{Clause } j.2}; \quad \underbrace{(\xi_{j.3}, \eta_{j.3}) = (\xi_j, \neg\eta_j)}_{\text{Clause } j.3}.$$

Clauses 1 to $p - 1$ force $F_\ell = 1$ when $x = \ell$, while Clauses $j.1$, $j.2$, and $j.3$ for $j = 1, \ldots, S$ force $x = \ell$ when $F_\ell = 1$. For any input x, choose $F_\ell = 1$ if $x = \ell$ and 0 otherwise, and choose $u_{\ell,i} = 1$ if the first $i + 1$ bits of x match those of ℓ and 0 otherwise, for $i = 1, \ldots, p - 2$. These choices satisfy all clauses, so every input admits a satisfying extension.

- **Altogether: Boolean Formula for all $p \geq 2$.** Combining the two scenarios, for any $p \geq 2$ and target configuration $\ell \in \{0, 1\}^p$, we define:

$$\phi_\ell = \left((p \bmod 2) \wedge \phi_\ell^{\text{odd}}\right) \vee \left(\neg(p \bmod 2) \wedge \phi_\ell^{\text{even}}\right), \tag{E.1}$$

where, for each fixed p, the parity of p selects the corresponding 3CNF formula constructed above. Every input x admits a satisfying extension, and every satisfying assignment has $F_\ell = 1$ if and only if $x = \ell$.

Step 2. Faithful Reduction from SAT to GSE Problem

Given any $\ell \in \{x | x \in \{0, 1\}^p\}$, we construct the explicit form of SAT in 3CNF with Boolean formula ϕ_ℓ (E.1), and we faifully reduce it to a vertex cover problem.

- **SAT to Vertex Cover Problem** We follow the same construction as in the case $p = 3$. The map h_1 assigns each Boolean variable and its negation to a pair of vertices in V_B, while h_2 assigns each literal occurrence in a clause to its corresponding vertex in V_C. In particular, $h_1 : x_i \to v_i$ for $i = 1, \ldots, p$, and $h_1 : F_\ell \to v_{p+1}, h_1 : u_{\ell,i} \to v_{p+1+i}$ for $i = 1, \ldots, p-2$. Each negated literal is mapped to the corresponding barred vertex. With the same edge construction, every satisfying assignment extends to a minimum vertex cover of G_ℓ, and every minimum vertex cover determines a satisfying assignment through its choices in V_B.

- **Vertex Cover Problem to GSE Problem.** We transform G_ℓ into the Hamiltonian in (3.8), using $h_3 : v_i \to s_{2i-1}$ and $\overline{v}_i \to s_{2i}$ for $i = 1, \ldots, 2p-1$, with the remaining indices assigned to clause vertices. Set $A_i = 1 - 2s_i$ for every vertex. In particular, the flag vertex v_{p+1} maps to s_{2p+1}.

  The formula has $2p-1$ Boolean variables and $p - 1 + 3\lceil p/2 \rceil$ clauses. Since it is satisfiable, a maximum independent set contains one vertex for each Boolean variable and one for each clause. We therefore set $K_0 = 3p - 2 + 3\lceil p/2 \rceil$, so H_ℓ takes nonnegative integer values with minimum zero. Every input admits an extension with $H_\ell = 0$, and in every such extension $A_{2p+1} = 2F_\ell - 1$, which equals 1 if and only if $x = \ell$.

$$\mathcal{H}_\ell = \left(\Delta + \sum_{\ell'} |f^{\ell'}| \right) H_\ell + f^\ell \frac{A_{2p+1} + 1}{2},$$

  where $\Delta > \max_{\ell'} f^{\ell'}$ and the sum ranges over all 2^p target configurations. When summing $\mathcal{H}_\ell$ over all ℓ, the flag terms contribute at least $-\sum_{\ell'} |f^{\ell'}|$. Since each H_ℓ is a nonnegative integer, any configuration with at least one $H_\ell > 0$ therefore has total energy at least Δ.

  Repeating this construction for all 2^p target configurations ℓ, we define:

$$H_G(A, A_{\text{aux}}) = \sum_\ell \mathcal{H}_\ell,$$

  where A denotes the p input spins and each construction uses a separate set of auxiliary spins, collected in A_{aux}. Every input admits an auxiliary assignment with all $H_\ell = 0$. Exactly one F_ℓ then equals 1, giving $H_G(A, A_{\text{aux}}) = f(A)$, while any configuration with some $H_\ell > 0$ has total energy at least Δ. Thus, H_G reproduces the values of f below Δ. Since each $\mathcal{H}_\ell$ contains only constant, one-body, and two-body terms, their sum retains the same form. Hence, H_G on $G = \bigcup_\ell G_\ell$ belongs to $\mathcal{M}_{\text{UI}}$.

- **Determining the Graph Size** For any $p \geq 2$, the Boolean formula ϕ_ℓ contains $2p-1$ Boolean variables and $p - 1 + 3\lceil p/2 \rceil$ clauses. This gives $|V_B| = 2(2p-1)$, $|V_C| = 3(p - 1 + 3\lceil p/2 \rceil)$, $|E_B| = 2p - 1$, and $|E_C| = |E_{B,C}| = 3(p - 1 + 3\lceil p/2 \rceil)$. Thus, for every target configuration ℓ, $|V_\ell| = 7p - 5 + 9\lceil p/2 \rceil$ and $|E_\ell| = 8p - 7 + 18\lceil p/2 \rceil$. Taking the union over all 2^p target configurations gives $|V| \leq 2^p(7p - 5 + 9\lceil p/2 \rceil)$ and $|E| \leq 2^p(8p - 7 + 18\lceil p/2 \rceil)$.

This completes the proof. □

E.2 Proof of Corollary 3.1.1

We now present the proof of Corollary 3.1.1.

Corollary E.1.1 (Corollary 3.1.1 Restated, Universal Ising for Binary Regression)**.** There exists a Hamiltonian $H_G \in \mathcal{M}_{\text{UI}}$ defined on a graph $G = (V, E)$ specified in Theorem 3.1 such that

$$\mathbb{E}[B|A] = H_G(A, A_{\text{aux}})$$

where the vertex set, the edge set and A_{aux} are identical to those specified in Theorem 3.1. Furthermore, the equality holds with $|H_G|(A, A_{\text{aux}}) \leq 1$ for some $J_{i,j}, r_i, K_0 \in \mathbb{R}$.

Proof. Since Theorem 3.1 holds for any $f : A \to \mathbb{R}$, it suffice to set $f(A) = 2\Phi(\beta^\top A) - 1$, where $\beta \in \mathbb{R}^p$ is the model parameters and Φ is the cumulative distribution function of the standard normal distribution. This completes the proof. □

E.3 Proof of Theorem 3.2

We now present the proof of Theorem 3.2.

Theorem E.2 (Theorem 3.2 Restated, Conditional Sufficiency of Universal Ising Representation)**.** Let $B \in \{-1, 1\}$, and suppose that $\mathbb{E}[B \mid A] = \theta_0^\top A^\dagger(A)$ for some $\theta_0 \in \mathbb{R}^k$. Then, $B \perp A \mid \theta_0^\top A^\dagger(A)$.

Proof. Let $S = \theta_0^\top A^\dagger(A)$. From the universal Ising representation, we have $\mathbb{E}[B|A] = S$. Then, it follows:

$$\begin{aligned}
\mathbb{P}(B = 1|A) &= \frac{1}{2}(1 + \mathbb{E}[B|A]) \\
&= \frac{1}{2}(1 + \mathbb{E}[\mathbb{E}[B|A]|S]) && \text{(By the definition of } S\text{)} \\
&= \frac{1}{2}(1 + \mathbb{E}[B|S]) && \text{(By the tower property)} \\
&= \mathbb{P}(B = 1|S).
\end{aligned}$$

Since B is binary and S is a function of A, this implies $B \perp A \mid S$. This completes the proof. □

E.4 Proof of Theorem 3.3

Theorem E.3 (Theorem 3.3 Restated, Consistency for Universal Ising Representation)**.** Let $\{A_i, B_i\}_{i=1}^n$ be i.i.d. binary observations. Define a fixed universal Ising vector $A^\dagger(A) \in \{-1, 1\}^k$ with a fixed k such that

$$\mathbb{E}[B|A] = \theta_0^\top A^\dagger$$

for some $\theta_0 \in \mathbb{R}^k$. Further, define the squared-error loss and its set of minimizers:

$$\mathcal{L}(\theta) = \mathbb{E}\big[(B - \theta^\top A^\dagger)^2\big], \quad \Theta_0 = \operatorname*{argmin}_{\theta \in \mathbb{R}^k} \mathcal{L}(\theta)$$

Then, any least-squares estimator $\widehat{\theta}_n$ satisfies $\inf_{\theta^* \in \Theta_0} \|\widehat{\theta}_n - \theta^*\|_2 \xrightarrow{a.s.} 0$ as $n \to \infty$, where

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \big(B_i - \theta^\top A^\dagger(A_i)\big)^2.$$

Proof. Recall $A \in \{-1, 1\}^p$ and $B \in \{-1, 1\}$. By the universality construction of the Ising vector, we have an exact expression: $\mathbb{E}[B|A] = \theta_0^\top A^\dagger$, where $A^\dagger : A \to \{-1, 1\}^k$ with a fixed $k \geq p$. Let

$$\mathcal{A} := \{A^\dagger \in \{-1, 1\}^k \mid \mathbb{P}(A^\dagger = a^\dagger) > 0\} \quad \text{and} \quad p_{a^\dagger} := \mathbb{P}(A^\dagger = a^\dagger).$$

We then show Theorem 3.3 with the following two steps.

Step 1: Characterize the population minimizers. Expanding the squared-error loss gives

$$\mathcal{L}(\theta) - \mathcal{L}(\theta_0) = \mathbb{E}\big[\big((\theta - \theta_0)^\top A^\dagger\big)^2\big] = \sum_{a^\dagger \in \mathcal{A}} p_{a^\dagger} \big(\theta^\top a^\dagger - \theta_0^\top a^\dagger\big)^2.$$

Since the right-hand side is nonnegative and every $p_{a^\dagger}$ is positive, we have

$$\Theta_0 = \big\{\theta \in \mathbb{R}^k \mid \theta^\top a^\dagger = \theta_0^\top a^\dagger \text{ for every } a^\dagger \in \mathcal{A}\big\}.$$

Thus, Θ_0 is the nonempty solution set of a fixed finite system of linear equations.

Step 2: Converge to the true conditional means. The tower property gives

$$\mathbb{E}[B|A^\dagger] = \mathbb{E}[\mathbb{E}[B|A]|A^\dagger] = \theta_0^\top A^\dagger.$$

For each $a^\dagger \in \mathcal{A}$, we define its sample count and its sample response mean:

$$N_n(a^\dagger) := \sum_{i=1}^{n} \mathbf{1}\{A^\dagger(A_i) = a^\dagger\} \quad \text{and} \quad \overline{B}_n(a^\dagger) := \frac{\sum_{i=1}^{n} B_i \mathbf{1}\{A^\dagger(A_i) = a^\dagger\}}{N_n(a^\dagger)}.$$

By the strong law of large numbers,

$$\frac{N_n(a^\dagger)}{n} \xrightarrow{a.s.} p_{a^\dagger} > 0 \quad \text{and} \quad \overline{B}_n(a^\dagger) \xrightarrow{a.s.} \theta_0^\top a^\dagger.$$

Since $\mathcal{A}$ is finite, these convergences hold simultaneously for all $a^\dagger \in \mathcal{A}$, and all sample counts are eventually positive almost surely. For such n, the within-state squared-error decomposition gives

$$\sum_{i:A^\dagger(A_i)=a^\dagger} \big(B_i - \theta^\top a^\dagger\big)^2 = \sum_{i:A^\dagger(A_i)=a^\dagger} \big(B_i - \overline{B}_n(a^\dagger)\big)^2 + N_n(a^\dagger)\big(\overline{B}_n(a^\dagger) - \theta^\top a^\dagger\big)^2.$$

The first term does not depend on θ. Comparing the empirical loss at its minimizer $\widehat{\theta}_n$ with that at θ_0:

$$\sum_{a^\dagger \in \mathcal{A}} \frac{N_n(a^\dagger)}{n} \big(\widehat{\theta}_n^\top a^\dagger - \overline{B}_n(a^\dagger)\big)^2 \le \sum_{a^\dagger \in \mathcal{A}} \frac{N_n(a^\dagger)}{n} \big(\theta_0^\top a^\dagger - \overline{B}_n(a^\dagger)\big)^2 \xrightarrow{a.s.} 0.$$

Each summand on the left is nonnegative, and each weight converges to a strictly positive value. Thus, $\widehat{\theta}_n^\top a^\dagger - \overline{B}_n(a^\dagger) \xrightarrow{a.s.} 0$ for every $a^\dagger \in \mathcal{A}$. Combining this with the convergence of $\overline{B}_n(a^\dagger)$ gives

$$\widehat{\theta}_n^\top a^\dagger \xrightarrow{a.s.} \theta_0^\top a^\dagger \quad \text{for every} \quad a^\dagger \in \mathcal{A}.$$

That is,

$$\sum_{a^\dagger \in \mathcal{A}} \big(\widehat{\theta}_n^\top a^\dagger - \theta_0^\top a^\dagger\big)^2 \xrightarrow{a.s.} 0.$$

This completes the proof. □

F Generalization of Universal Ising to Non-Binary Regression

Universal Ising offers a structured, binary representation of pairwise interactions for statistical models in binary regression. In this section, we show that this representation extends beyond binary covariates. First, we introduce the generalization technique in Section F.1, followed by a concrete example in Section F.2 to illustrate its application.

F.1 Universal Ising for Non-Binary Regression

For a regression with covariates taking T distinct values, generalizing the two-body representation requires replicating Step 1 of the proof of Theorem 3.1, where the process encodes target states into a SAT problem. For a regression with continuous covariates, discretization becomes essential since no model with discrete variables replicates a continuous one. After discretization, the construction follows the general discrete case. Next, we elaborate on these two non-binary cases.

- **Universal Ising for Discrete Covariate.** Consider Steps 1 and Step 2 from the proofs of Theorem 2.1 and Theorem 3.1. The construction of both the universality principle and the Universal Ising framework begins by formulating a SAT in 3CNF that encodes the solutions of interest. Subsequently, the computational equivalence between SAT and the GSE problem facilitates a faithful reduction of these solutions into regression models. Therefore, it suffices to replicate the first-step SAT formulation for general discrete covariates. Specifically, since any Boolean formula can be transformed into an equivalent 3CNF representation [39, 40] in polynomial time, we encode states of every discrete covariates using $\log_2 T$ Boolean variables. A concrete example of this extension is provided in Section F.2.
- **Universal Ising for Continuous Covariate.** For continuous covariates, on the other hand, exact reconstruction of the energy spectrum is infeasible with Hamiltonian over discrete variables. By incorporating existing discretization techniques [41, 5] that approximate continuous variables, the universality construction and Universal Ising framework generalize to continuous case while preserving its fundamental properties. In particular , the following lemmas guarantee arbitrary accuracy for binary discretization.

Lemma F.1 (Discretization of Continuous Variable, [4, 5]). Let $U = (U_1, \ldots, U_p)^\top \in [-1, 1]^p$. Then, there exist a sequence of biary variables $\{A_{j,d}\}$ with $j \in [p]$ and $d \in [D]$ that only takes values -1 and 1, such that $\max_{1\le j\le p}\{|U_j - U_{j,D}|\} \to 0$ uniformly as $D \to \infty$, where $U_{j,D} = \sum_{d=1}^{D}(A_{j,d})/2^d$.

Lemma F.2 (Discretization Error, Theorem 4.7 of [4]). Let $U_1, \ldots, U_p \sim \mathrm{Unif}(-1, 1)$ and define approximation $U_{j,D} = \sum_{d=1}^{D}(A_{j,d})/2^d$. Then, for a differentiable function $g : (-1, 1)^p \to \mathbb{R}$, we have $|g(U_1, \ldots, U_p) - g(U_{1,D}, \ldots, U_{p,D})| = \mathcal{O}_p(2^{-D})$.

By Lemma F.1, the first D data bits $\{A_d\}_{d=1}^{D}$ approximates the distribution of U, where each $A_{j,d}$ takes values -1 and 1, Taking elements in the sequence $\{A_{j,d}\}$ with $j \in [p]$ and $d \in [D]$ as a set of binary covariates, Universal Ising framework (Theorem 3.1) applies to continuous covariates with error specified by Lemma F.2, where we achieve arbitrary accuracy at the expense of increasing the number of covaraites. The subsequent derivations are identical through Step 1 to Step 3.

F.2 Example of Universal Ising for Discrete Case.

We present a concrete example of using Universal Ising to replicate logistic regression with covariates taking four possible values.

Definition F.1 (Logistics Regression with Discrete Covariate). Given binary response $B \in \{-1, 1\}$ and covariates $A = (A_1, \ldots, A_p)^\top \in \{1, 2, \ldots, T\}^p$ that takes T distinct values, we have logistic

regression with the form:

$$\mathbb{P}(B = 1|A) = g^{-1}\left(\gamma_0 + A^\top \gamma\right), \tag{F.1}$$

where $g : (0, 1) \to \mathbb{R}$ is the logit link function, $\gamma_0 \in \mathbb{R}$ and $\gamma \in \mathbb{R}^p$ are model parameters.

We demonstrate that the first part of Theorem 3.1 holds for Definition F.1 as well, while the upper-bound of the graph size provided in the second part of Theorem 3.1 changes due to the increase in degrees of freedom for the given target model (F.1). To begin with, we denote the output of the conditional expectation by $H_{\text{logit}}(A)$ (recall Hamiltonian, Definition 2.5).

Example 1 (Modeling over Discrete Covariates)**.** Let $T_1, T_2, T_3, T_4 \in \mathbb{R}$ be distinct, and consider logistic regression with covariates $A = (A_1, A_2)^\top \in \{T_1, T_2, T_3, T_4\}^2$. Encode each covariate using two binary spins, and denote the combined encoding by $X(A) \in \{-1, 1\}^4$. For any $\gamma_0 \in \mathbb{R}$ and $\gamma \in \mathbb{R}^2$, there exist a finite integer N and a fixed Hamiltonian $H_G \in \mathcal{M}_{\text{UI}}$ such that every input A admits an auxiliary assignment $A_{\text{aux}} \in \{-1, 1\}^N$ satisfying

$$g^{-1}(\gamma_0 + A^\top \gamma) = H_G(X(A), A_{\text{aux}}),$$

where g is the logit link function. The size of $G = (V, E)$ is bounded by $|V| \leq 656$ and $|E| \leq 976$.

Proof. The encoding is a bijection between $\{T_1, T_2, T_3, T_4\}^2$ and $\{-1, 1\}^4$. Define $f(X(A)) = g^{-1}(\gamma_0 + A^\top \gamma)$ and apply Theorem 3.1 with $p = 4$. This completes the proof. □

G Binary Expansion Linear Effect

To highlight the benefits of Universal Ising and potential application of the universality-based framework, we quote relevant results from [4]. Section G.1 introduces preliminaries to binary linear model. Section G.2 presents existing results of modeling linearity for binary regression.

G.1 Preliminaries for Binary Linear Model

We introduce essential definitions for the formulation of linear model in binary regression based on binary expansion.

Definition G.1 (Kronecker Product)**.** The Kronecker product, denoted by $\otimes$, is a binary operator that takes two matrices and produces a block matrix. Specifically, for two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, the Kronecker product $A \otimes B \in \mathbb{R}^{mp\times nq}$ is defined as:

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}$$

Given binary predictors represented by vector, we define its *power vector* through Kronecker product:

Definition G.2 (Power Vector)**.** Consider a set of binary variables represented by random vector: $A = (A_1, \ldots, A_p)^\top \in \{-1, 1\}^p$. We denote the power vector of A by $A^\otimes$ with the following form:

$$A^\otimes = \begin{pmatrix} 1 \\ A_1 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ A_2 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 \\ A_p \end{pmatrix}$$

We construct the power vector $A^\otimes$ through $p-1$ sequential Kronecker products, embedding each component A_i in a 2×1 matrix with unity. This construction maps the original p-dimensional binary space to a 2^p-dimensional space. We refer to this as *binary expansion* of a binary vector.

Crucially, the set with all possible realizations of $A^\otimes$ constitutes a Hadamard matrix of order 2^p by Sylvester's construction:

Definition G.3 (Hadamard Matrix)**.** An order n Hadamard matrix H_n, is an $n \times n$ matrix with entries ± 1 such that $H_n H_n^T = nI_n$. Moreover, Sylvester's construction generates Hadamard matrices of order 2^k for $k \in \mathbb{N}$ with $H_1 = [1]$ by:

$$H_{2^k} = \begin{pmatrix} H_{2^{k-1}} & H_{2^{k-1}} \\ H_{2^{k-1}} & -H_{2^{k-1}} \end{pmatrix}.$$

Brown et al. [4] propose the binary expansion linear effect (BELIEF) framework by the binary expansion: $A \to A^\otimes$.

G.2 Binary Expansion Linear Effect

We begin with the motivation of BELIEF.

- **Motivation: Linearity in Binary Variable**

  The following lemma is the key motivation for the establishment of binary expansion based model.

Lemma G.1 (Lemma 2.2 of Brown et al. [4])**.** Let $A = (A_1, \ldots, A_p)^\top \in \{-1, 1\}^p$. For any function $f : \{-1, 1\}^p \to \mathbb{R}$, there exists $\beta \in \mathbb{R}^{2^p}$ such that $f(A) = \beta^\top A^\otimes$.

Proof. Let $c_v \in \mathbb{R}$. For any function $f : A \to \mathbb{R}$, we decompose $f(A)$ by summing over $v \in \{-1,1\}^p$:

$$\begin{aligned} f(A) &= \sum_{v\in\{-1,1\}^p} c_v \mathbb{1}(A_1 = v_1)\cdots\mathbb{1}(A_p = v_p) \\ &= \sum_{v\in\{-1,1\}^p} c_v \left(\frac{1+A_1v_1}{2}\right)\cdots\left(\frac{1+A_pv_p}{2}\right). \end{aligned}$$

We rearrange the last line into the form $\beta^\top A^\otimes$ for some $\beta \in \mathbb{R}^{2^p}$. This completes the proof. □

Lemma G.1 asserts that any function over the power vector of binary variables ($A^\otimes$) is linear. This motivates the idea of constructing a statistical model among binary variables through a linear model in contrast to classical GLMs.

- **Formal Result for BELIEF Framework** Theorem G.1 characterizes the binary expansion-based approach for modeling in binary regression problem:

Theorem G.1 (Binary Expansion Linear Effect, Theorem 2.1 of [4]). Let $A = (A_1,\ldots,A_p)^\top \in \{-1,1\}^p$, $B \in \{-1,1\}$ and $C = (C_1,\ldots,C_q)^\top \in \{-1,1\}^q$. Then,

1. For all i, j, A and C are independent if and only if power vectors $A^\otimes$, $C^\otimes$ satisfy
$$\mathrm{Cov}(A_i^\otimes, C_j^\otimes) = 0.$$
2. There exists a constant vector $\beta \in \mathbb{R}^{2^p}$ such that:
$$\mathbb{E}[B|A] = \beta^\top A^\otimes \tag{G.1}$$
3. Whenever (G.1) holds, we have $\beta = \mathbb{E}[A^\otimes A^{\otimes\top}]^{-1}\mathbb{E}[A^\otimes B]$ whenever $\mathbb{E}[A^\otimes A^{\otimes\top}]$ is invertible. Moreover, β satisfies (1): $|\beta^\top A^\otimes| \le 1$, (2): $\|H_{2^p}\beta\|_\infty \le 1$ and (3): $\|\beta\|_2 \le 1$.

Remark G.1. Part 1 of BELIEF main theorem establishes an analogue to a fundamental property of Gaussian world. Specifically, the pairwise correlations between components of $A^\otimes$ and $C^\otimes$ determine the independence of A and C. Parts 2 and Part 3 of BELIEF main theorem constitute the theoretical foundation of BELIEF through Lemma G.1, where the conditional expectation $\mathbb{E}[B|A]$ maps from the σ-algebra generated by $\{A_1, ..., A_p\}$.

In particular, (G.1) extends to estimation with BELIEF. Formally, let $\{B_i, A_i\}_{i=1}^n$ be a set of i.i.d samples, where each A_i is a p-dimensional binary predictors and B_i is binary response. Further, we define $\mathcal{B} = (B_1,\ldots,B_n)^\top \in \mathbb{R}^n$ and $\mathcal{A} = (A_1^\otimes,\ldots,A_n^\otimes)^\top \in \mathbb{R}^{n\times 2^p}$. BELIEF framework guarantees that, whenever the least square estimator (LSE) $\widehat{\beta} = (\mathcal{A}^\top\mathcal{A})^{-1}\mathcal{A}^\top\mathcal{B}$ exists, it satisfies $(1) : \|H_{2^p}\widehat{\beta}\|_\infty \le 1$, $(2) : \|\mathcal{A}\widehat{\beta}\|_\infty \le 1$ and $(3) : \|\widehat{\beta}\|_2 \le 1$, where H_{2^p} is the order 2^p Hadamard matrix by Sylvester's construction.

Modeling Linearity through BELIEF. These showcase how BELIEF address the limitations of GLMs. In terms of interpretability, components of β resemble those of classical linear models: their signs and magnitudes quantify the direction and strength of binary interactions with the response. Plus, properties of $\widehat{\beta}$ ensure that every output across all 2^p configurations lies within $[-1, 1]$.

Addressing Perfect Separation. For perfect separation, $\widehat{\beta}$ reveals that the parameter space forms a 2^p-dimensional hypercube. When perfect separation occurs, vertices of such hypercube correspond to a model where A completely determines B. These vertices fully characterize the phenomenon, providing a geometric perspective on the structure of the parameter space under such conditions.

H Experimental Details

We present proof-of-concept numerical results to support our theoretical claims in Theorem 3.1. We evaluate Universal Ising on randomly generated Boolean functions $f : \{-1, 1\}^p \to \{-1, 1\}$ with up to 6 auxiliary variables. We increase p and measure both accuracy and interpretability.

Objective: Control Expressivity and Interpretability. We study how many auxiliary variables are needed for both expressivity and interpretability.

- **Expressive Capacity.** We measure Universal Ising's ability to approximate random Boolean functions under a fixed upper bound on auxiliary variables.

- **Interpretability.** We also track the stability of the two-body parameters J_{ij}. By studying how J_{ij} changes as we add more auxiliary bits, we see if Universal Ising can discover reliable structural interactions.

Algorithm. Algorithm 1 shows our EM-like procedure for training Universal Ising. For each input $A \in \{-1, 1\}^p$, the algorithm introduces k auxiliary bits $A_{\text{aux}} \in \{-1, 1\}^k$. We then optimize these bits and the model parameters to approximate $f(A)$.

- **E-step (Expectation).** For each A, we test all 2^k configurations of auxiliary bits. We pick the configuration that minimizes $(f(A) - H_G(\widetilde{A}))^2$, where $\widetilde{A} = (A, A_{\text{aux}})$ and

$$H_G(\widetilde{A}) \;=\; \sum_{i<j} J_{ij} A_i A_j \;+\; \sum_i r_i A_i \;+\; K_0. \tag{H.1}$$

- **M-step (Maximization).** With these chosen auxiliary bits, we form a training set of pairwise and linear features. We update $\{J_{ij}, r_i, K_0\}$ via linear regression to minimize squared error.

We repeat these steps until convergence. Although the worst-case theory may require exponentially many auxiliary bits, most functions in practice need only a few. This EM-like approach avoids searching the full model class.

Data. We sample 500 random Boolean functions for each $p \in \{3, 4, 5, 6, 7, 8, 9\}$. Each function $f : \{-1, 1\}^p \to \{-1, 1\}$ is a ground-truth target. We test Universal Ising with up to 6 auxiliary variables.

Computational Resource. All experiments run on a single CPU core (Intel Xeon, 2.6 GHz). Our code uses Python (`scikit-learn`, `NumPy`, `pandas`) without GPU acceleration. We can efficiently enumerate k auxiliary bits for $p \leq 9$.

Results and Discussion. We summarize our key findings on the expressive power and interpretability of the Universal Ising model using two figures:

- **Expressivity.** Figure 3(a) shows the fraction of randomly generated Boolean functions that Universal Ising can approximate with squared loss below 10^{-4}. When $p = 3$ or $p = 4$, nearly all functions require only one or two auxiliary variables. For $p = 5$, around 50% of functions can be fitted using at most three auxiliary bits—far below the exponential upper bound from our universality result (Theorem 3.1).

  These results echo Shannon's classic finding [42] that most Boolean functions are hard in the worst case, but demonstrate that many random functions are in fact "compressible" through low-order pairwise interactions and a small number of latent variables. Universal Ising provides a practical way to realize this compression.

Algorithm 1 EM-like Training Procedure for Universal Ising Model with Auxiliary Variables

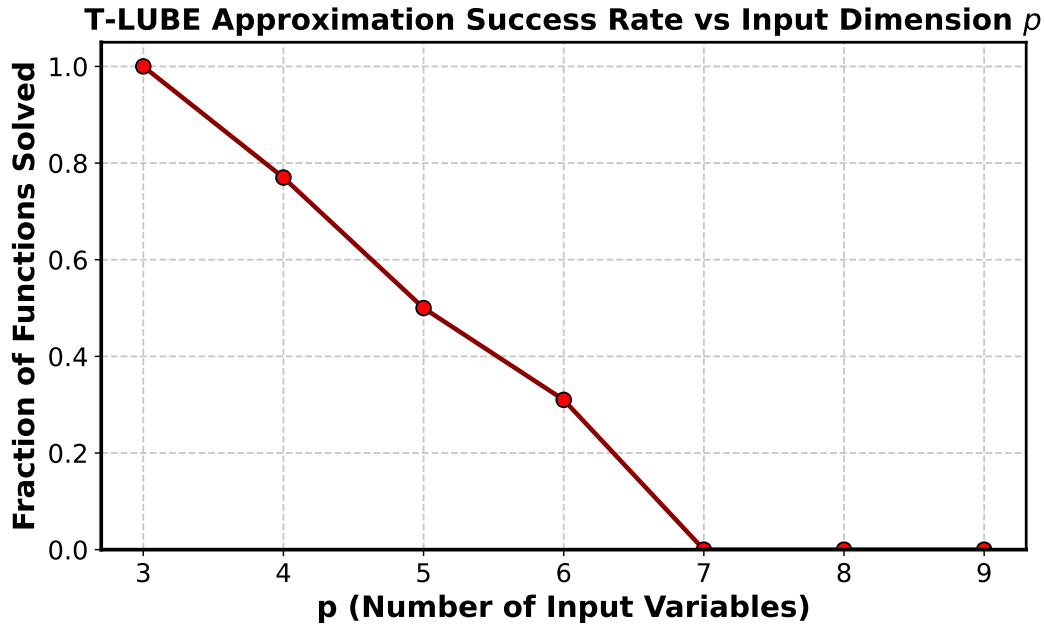
Require: Binary function $f : \{-1, 1\}^p \to \mathbb{R}$, number of auxiliary bits k, regularization strength λ, maximum iterations T

Ensure: Approximate two-body model parameters $\{J_{ij}, r_i, K_0\}$

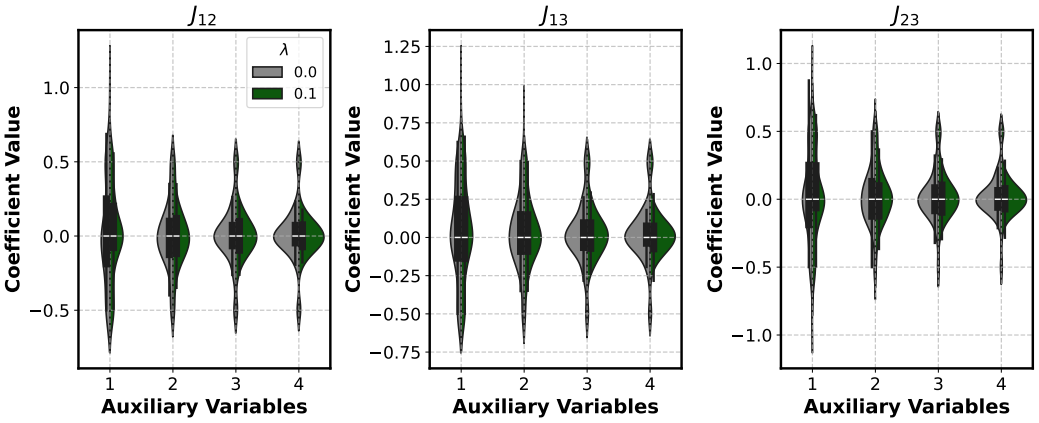
Initialize model parameters $J_{ij}, r_i, K_0 \leftarrow 0$
for $t = 1$ to T **do**
  $\mathcal{D} \leftarrow \emptyset$
  for all $A \in \{-1, 1\}^p$ **do**
    $f_A \leftarrow f(A)$
    min_error $\leftarrow \infty$, best_aux $\leftarrow \emptyset$
    for all $A_{\text{aux}} \in \{-1, 1\}^k$ **do**
      $\widetilde{A} \leftarrow (A, A_{\text{aux}})$
      Construct all pairwise and linear features from $\widetilde{A}$
      $x \leftarrow$ feature vector of $\widetilde{A}$
      $H_G(\widetilde{A}) \leftarrow \sum_{i<j} J_{ij} A_i A_j + \sum_i r_i A_i + K_0$
      if $(f_A - H_G(\widetilde{A}))^2 <$ min_error **then**
        min_error $\leftarrow (f_A - H_G(\widetilde{A}))^2$
        best_aux $\leftarrow A_{\text{aux}}$
      end if
    end for
    $\widetilde{A}^* \leftarrow (A, \text{best_aux})$
    $x^* \leftarrow$ features from $\widetilde{A}^*$
    Add (x^*, f_A) to $\mathcal{D}$
  end for
  Update $\{J_{ij}, r_i, K_0\}$ via ridge regression on $\mathcal{D}$ with strength λ
  optional: Check convergence and break if necessary
end for
return final model parameters $\{J_{ij}, r_i, K_0\}$

- **Interpretability.** Figure 3(b) shows the distribution of two-body interaction coefficients J_{ij} across 100 random Boolean functions when $p = 3$. As we increase the number of auxiliary variables from 1 to 4, the coefficient distributions become more concentrated, and their variances decrease.

  This convergence behavior suggests that the model discovers consistent structural interactions even with increasing auxiliary capacity. In low dimensions, the learned coefficients tend to stabilize, indicating that Universal Ising not only approximates functions accurately but also learns interpretable and coherent representations.



(a) Universal Ising success rate across dimensions p.



(b) Effect of L2 regularization on J_{ij} distributions for $p = 3$

Figure 3: (a) Proportion of Boolean functions that can be fitted by Universal Ising with loss below 10^{-4} using at most 6 auxiliary variables. Here, we use Two-Body, linear, universal binary effect (T-LUBE) to denote the Universal Ising (UI) representation. (b) Comparison of the two-body coefficient distributions J_{12}, J_{13}, J_{23} across 100 random Boolean functions for $p = 3$, under two regularization strengths $\lambda = 0$ (black) and $\lambda = 0.1$ (red). The left and right halves of each violin represent the two settings, respectively. Regularization leads to more concentrated distributions, indicating improved stability and interpretability.

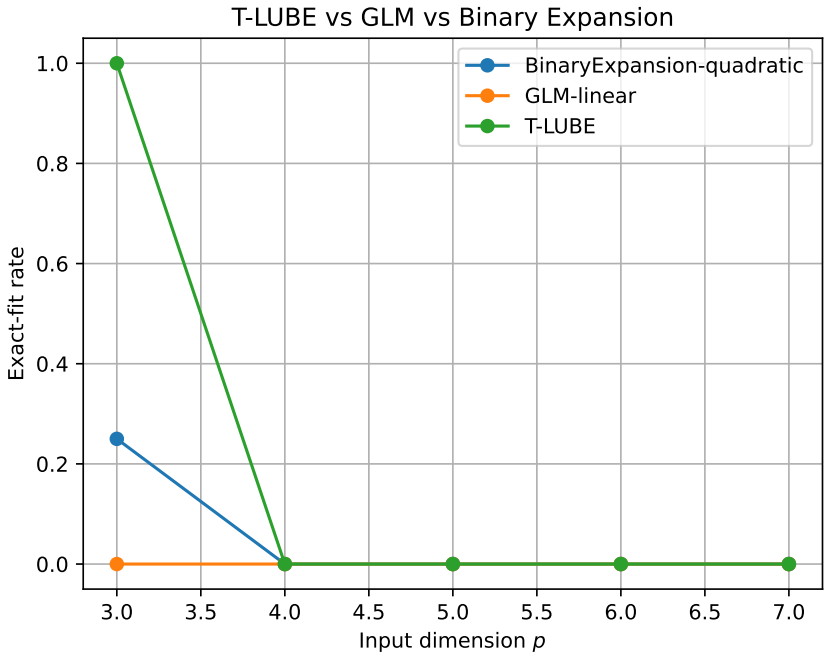


Figure 4: Comparison of exact-fit rate for Universal Ising, generalized linear models (GLM), and quadratic binary expansion as a function of input dimension p. Here, we use Two-Body, linear, universal binary effect (T-LUBE) to denote the Universal Ising (UI) representation. UI achieves perfect reconstruction in the low-dimensional regime ($p = 3$), while both baselines fail. As p increases, all methods degrade under fixed model capacity, illustrating the increasing difficulty of representing random Boolean functions in higher dimensions.

Table 2: Average runtime (in seconds) of Universal Ising, GLM, and quadratic binary expansion, averaged over $p = 3, \ldots, 7$. GLM and quadratic expansion are faster due to their closed-form regression updates, whereas Universal Ising incurs additional cost from latent-variable search in the E-step.

Method	Mean Runtime (s)	Main Computational Step
GLM-linear	0.00023	linear regression
BinaryExpansion-quadratic	0.00023	quadratic regression
Universal Ising	0.01829	latent search + regression

I Additional Numerical Results

Comparison with GLM and Binary Expansion. We compare Universal Ising against two standard alternatives on the same random Boolean-function fitting task: (i) a generalized linear model (GLM) with linear features only, and (ii) a quadratic binary expansion model without auxiliary variables. For each input dimension $p \in \{3, 4, 5, 6, 7\}$, we sample random Boolean functions $f : \{-1, 1\}^p \to \{-1, 1\}$ and evaluate all methods on the full truth table. We report two quantities: *exact-fit rate*, defined as the fraction of functions with final mean-squared error below 10^{-8}, and *runtime*, measured in seconds on the same CPU implementation.

Results. The comparison is summarized in Fig. 4 and Table 2. For $p = 3$, Universal Ising reaches an exact-fit rate of 1.00, while the quadratic binary expansion baseline reaches 0.25 and GLM fails to exactly fit any sampled function. This demonstrates that the auxiliary-variable formulation improves representational capacity in the low-dimensional regime. For larger p, all methods degrade under fixed model capacity, reflecting the increasing difficulty of representing random Boolean functions.

Table 3: **Sensitivity to Initialization on the Random Boolean-Function Fitting Task.** For each input dimension $p \in \{3, 4, 5\}$, we fit Universal Ising to multiple randomly generated targets $f : \{-1, 1\}^p \to \{-1, 1\}$ and report the mean final loss, exact-fit rate, and average number of EM iterations. We compare two parameter initializations: zero initialization and small random Gaussian initialization with scale 10^{-3}. Small random initialization consistently attains near-zero loss and exact reconstruction for $p = 3, 4, 5$, while zero initialization works for $p = 3$ but fails for $p = 4, 5$. This shows that the EM-like optimization is sensitive to symmetry-preserving initialization, and that a simple random initialization is sufficient to obtain stable high-quality solutions.

p	Initialization	Mean Loss	Exact-Fit Rate	Iterations
3	Random (10^{-3})	$\sim 10^{-30}$	1.00	2.00
3	Zero	$\sim 10^{-30}$	1.00	2.75
4	Random (10^{-3})	$\sim 10^{-30}$	1.00	2.00
4	Zero	0.36	0.00	2.00
5	Random (10^{-3})	$\sim 10^{-30}$	1.00	2.00
5	Zero	0.79	0.00	2.00

Computational complexity. We briefly compare the computational complexity of the methods. GLM and quadratic binary expansion reduce to a single regression problem, leading to polynomial-time complexity in the number of features. In contrast, Universal Ising employs an EM-like procedure, where the E-step enumerates all auxiliary configurations for each input, introducing an additional exponential dependence on the number of auxiliary variables. This difference explains the higher runtime observed for Universal Ising, and reflects a fundamental trade-off between computational efficiency and representational flexibility.

From the efficiency perspective, GLM and quadratic binary expansion are significantly faster than Universal Ising, since they only require a single regression solve. By contrast, Universal Ising performs an additional latent-variable search over all auxiliary assignments for each input in every EM iteration. Nevertheless, the runtime remains practical in the regime considered here. This reflects a trade-off between computational cost and representational flexibility.

These results show that GLM and quadratic binary expansion provide fast but capacity-limited baselines, whereas Universal Ising offers improved representational flexibility through its auxiliary-variable formulation, making the trade-off between expressivity and efficiency explicit.

Sensitivity analysis. We evaluate the stability of Universal Ising with respect to initialization, tuning choices, stopping criteria, and data variation.

Initialization. We compare zero initialization and small random initializations with different scales. For each input dimension $p \in \{3, 4, 5\}$, we generate multiple random Boolean functions and evaluate performance averaged over these instances. Table 3 shows that the optimization outcome depends on initialization. For $p \geq 4$, zero initialization leads to suboptimal solutions, while small random initialization consistently yields near-zero loss and perfect exact-fit rates. This behavior is expected due to the non-convex nature of the EM-like optimization. In particular, zero initialization preserves symmetry in the model and can lead to degenerate assignments in the E-step, while small random initialization breaks this symmetry and enables convergence to better solutions. Importantly, stable and accurate solutions can be reliably obtained using simple random initialization.

Stopping criteria and tuning choices. We vary the maximum number of iterations and convergence tolerance (max_iter $\in \{5, 10, 20\}$, tol $\in \{10^{-4}, 10^{-6}\}$) while keeping all other settings fixed. Across these configurations, we observe identical losses, exact-fit rates, and iteration counts. This indicates that the algorithm converges quickly (typically within 2–3 iterations) and is stable with respect to reasonable tuning choices.

Auxiliary variables. We further examine the effect of the number of auxiliary variables under random initialization. For each input dimension $p \in \{3, \ldots, 7\}$, we generate 20 independent random Boolean functions $f : \{-1, 1\}^p \to \{-1, 1\}$, where each function value is sampled uniformly from $\{-1, 1\}$. All reported results are averaged over these instances.

We vary the number of auxiliary variables in the range $k \in \{1, \ldots, 9\}$. For each configuration, we run the EM-like optimization procedure with a fixed maximum of 10 iterations and convergence tolerance 10^{-6}. Model parameters are initialized using small random Gaussian values with scale 10^{-3}.

Figure 5 shows that the exact-fit rate improves as the number of auxiliary variables increases. For lower dimensions ($p \leq 5$), there exists a small threshold value of k beyond which the model achieves perfect reconstruction across all instances. For example, perfect reconstruction is consistently achieved for $k \geq 3$ when $p = 5$.

For higher dimensions ($p = 6, 7$), larger values of k are required to achieve perfect reconstruction. Within the tested range $k \leq 6$, performance continues to improve but does not yet reach 100% for $p = 7$, indicating that higher-dimensional problems require larger auxiliary capacity. Due to the non-convex nature of the optimization, the empirical performance is not strictly monotonic in k, but the overall trend is clearly increasing.

These results indicate that the number of auxiliary variables directly controls the representational capacity of the model. Overall, auxiliary variables provide a controllable mechanism for increasing representational capacity, with a clear dependence on problem dimensionality.

Data sensitivity. All results are averaged over multiple independently generated random Boolean functions. We observe consistent trends across different instances, indicating that the fitted model is not sensitive to small variations in the data in this setting.

Overall, these results demonstrate that Universal Ising exhibits stable and predictable behavior in practice, with sensitivity primarily arising from initialization, which is a common characteristic of latent-variable models.

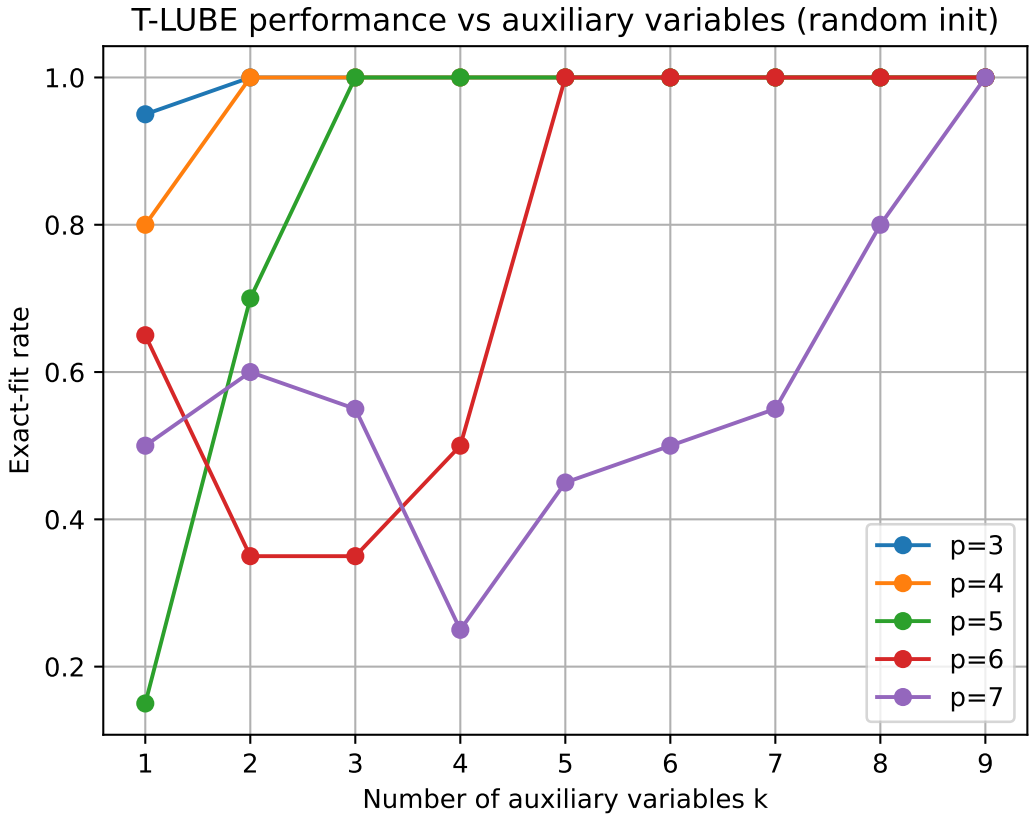


Figure 5: **Exact-Fit Rate versus the Number of Auxiliary Variables k on the Random Boolean-Function Fitting Task.** For each input dimension $p \in \{3, \ldots, 7\}$, we generate 20 independent random targets $f : \{-1, 1\}^p \to \{-1, 1\}$ with outputs sampled uniformly from $\{-1, 1\}$, fit Universal Ising with the EM-like procedure, and report the fraction of instances that are exactly reconstructed. Across all runs, model parameters are initialized from a Gaussian distribution with scale 10^{-3}, the maximum number of iterations is 10, and the convergence tolerance is 10^{-6}. The exact-fit rate increases with k for every p. For lower dimensions ($p \leq 5$), only a few auxiliary variables are enough to reach 100% exact recovery. For higher dimensions ($p = 6, 7$), performance still improves with k but requires larger auxiliary capacity, and within the tested range perfect recovery is not always reached. The small non-monotonic fluctuations reflect the non-convexity of the optimization rather than a reversal of the overall trend. Here, we use Two-Body, linear, universal binary effect (T-LUBE) to denote the Universal Ising (UI) representation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We are very precise about our main results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please see our limitation discussions in Section B.1.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see our Sections H and I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submitted our code with the draft.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see our Sections H and I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see our plots and tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see our Sections H and I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: All authors read through carefully.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see the beginning of our appendix for impact statement.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Given the formal nature of this work, we don't see such concerns.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: Given the formal nature of this work, we don't see such concerns.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Given the formal nature of this work, we don't see such concerns.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Given the formal nature of this work, we don't see such concerns.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Given the formal nature of this work, we don't see such concerns.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Please see our LLM usage statement at the beginning of the appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.