# Certified Adaptive Discovery of Algebraic Laws

Luca Blanchi

June 2026

**Abstract**

We develop a finite-sample inference framework for algebraic laws discovered after adaptive search. The basic input is a volume certificate: a subset $A$ of a finite sample space $X$, an upper bound $V \geq |A|$, and a prefix-free description length. Such certificates generate valid e-values under the uniform null. We prove exact-fit, partial-fit, likelihood-ratio, sequential, replicated, and change-point variants. We then specialize the framework to rational points of bounded height in projective space, where algebraic laws are closed subvarieties and volume certificates are supplied by height-counting bounds. The resulting tests apply to symbolic regression over rational data, algebraic subspace clustering, finite-field congruence laws, and related algebraic-discovery problems. We also prove a universal dominance statement: relative to a computably enumerable certificate language, a universal mixture dominates every effective volume-certified algebraic e-test, up to a multiplicative constant. The paper is not a new rational-point-counting theorem; rather, it supplies a rigorous post-selection and optional-stopping-valid inference layer for algebraic law discovery.

## 1 Introduction

Many scientific-discovery and symbolic-regression procedures search through large families of polynomial relations. Given data, an algorithm may report an equation

$$F(x) = 0$$

or a more general algebraic condition

$$x \in Y \subsetneq \mathbb{P}^n$$

only after inspecting the data. This creates an immediate post-selection problem. A relation may look striking simply because the search space was large, the degree was high, or the algorithm was allowed to adapt.

This paper gives a finite-sample significance theory for such discoveries. The central idea is simple. If a candidate law cuts down the sample space from $X$ to a subset $A$, and one has a certificate

$$|A| \leq V,$$

then an exact fit to $A$ is unlikely under a uniform null when $V/|X|$ is small. If many possible laws are searched, the search cost is paid through a prefix-free description length. This produces e-values and e-processes that remain valid after arbitrary adaptive search and optional stopping.

The framework is deliberately modular. The statistical component is based on e-values, Kraft weights, likelihood-ratio mixtures, and nonnegative supermartingales. The geometric component enters only through volume certificates. In arithmetic applications, these certificates may come from elementary height-counting bounds, determinant-method estimates, finite-field Schwartz-Zippel bounds, or problem-specific geometry.

The contribution is therefore not a new height-counting theorem. Instead, it is a certified inference layer for algebraic discovery: once an algebraic candidate comes with a valid volume certificate and a code length fixed in advance as part of a certificate language, it receives a post-selection-valid evidence score.

The main results are:

- exact-fit post-selection validity;
- partial and noisy-fit validity via binary KL scores;
- likelihood-ratio mixtures and oracle inequalities;
- conditional and sequential e-processes for adaptive data streams;
- a corrected finite-class model-selection consistency theorem;
- replicated-law aggregation under explicit independence or conditional-validity assumptions;
- a change-point mixture for emergent laws;
- universal dominance for effective volume-certified algebraic e-tests;
- concrete algebraic certificates for hypersurfaces, linear subspaces, projective varieties, and finite fields;
- applications to symbolic regression, algebraic subspace clustering, finite-field congruence discovery, low-rank structure, and conservation laws.

## 2 Background and terminology

An e-value for a null hypothesis $H_0$ is a nonnegative random variable $E$ satisfying

$$\mathbb{E}_{H_0} E \leq 1.$$

By Markov's inequality,

$$\mathbb{P}_{H_0}(E \geq 2^s) \leq 2^{-s}.$$

Thus a large e-value gives evidence against $H_0$, with $\min\{1, 1/E\}$ acting as a valid p-value.

An e-process is a nonnegative process $(E_t)_{t\geq 0}$ that is a supermartingale, or at least satisfies $\mathbb{E}_{H_0} E_t \leq 1$ for all $t$, under the null. If $(E_t)$ is a nonnegative supermartingale with $E_0 \leq 1$, Ville's inequality gives

$$\mathbb{P}_{H_0}\left(\sup_{t\geq 0} E_t \geq 2^s\right) \leq 2^{-s}.$$

This gives optional-stopping validity.

We use prefix-free code lengths $\kappa(c)$ satisfying Kraft's inequality:

$$\sum_c 2^{-\kappa(c)} \leq 1.$$

The certificate language and code lengths are fixed before the data are analyzed. The data-dependent discovery procedure may choose a certificate adaptively after seeing the data, but it must choose from this pre-specified language.

## 3 Abstract volume certificates

Let $X$ be a finite set. Under the basic null,

$$H_0 : \quad X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(X).$$

**Definition 3.1** (Volume certificate)**.** *A volume certificate is a triple*

$$c = (A_c, V_c, \kappa_c),$$

*where*

$$A_c \subseteq X, \qquad V_c > 0, \qquad |A_c| \leq V_c,$$

*and $\kappa_c$ is a prefix-free code length. Set*

$$p_c = \min\left\{1, \frac{V_c}{|X|}\right\}.$$

Then

$$\mathbb{P}_{H_0}(X_i \in A_c) \leq p_c.$$

The interpretation is that $A_c$ is the set satisfying a discovered law, $V_c$ is an upper bound on its null volume, and $\kappa_c$ is the search or description cost of the law.

Certificates with $p_c = 1$ are valid but noninformative. Likelihood-ratio formulae below are stated for $0 < p_c < 1$; endpoint cases are handled by omission or by the corresponding limiting convention.

## 4 Exact-fit adaptive discovery

For fixed $c$, define

$$S_N(c) = N \log_2 \frac{1}{p_c} - \kappa_c$$

if

$$X_1, \ldots, X_N \in A_c,$$

and define $S_N(c) = -\infty$ otherwise. Let

$$S_N^* = \sup_c S_N(c).$$

**Theorem 4.1** (Exact-fit post-selection significance)**.** *Under $H_0$, for every $s \geq 0$,*

$$\mathbb{P}_{H_0}(S_N^* \geq s) \leq 2^{-s}.$$

*Consequently, if an arbitrary adaptive discovery procedure reports a certificate $c$ with $S_N(c) = s$, then*

$$p_{\text{post}} \leq 2^{-s}$$

*is a valid post-selection p-value.*

*Proof.* For fixed $c$,

$$\mathbb{P}_{H_0}(X_1, \ldots, X_N \in A_c) \leq p_c^N.$$

If $S_N(c) \geq s$, then

$$N \log_2 \frac{1}{p_c} - \kappa_c \geq s,$$

or equivalently,

$$p_c^N \leq 2^{-s} 2^{-\kappa_c}.$$

Therefore

$$\mathbb{P}_{H_0}(S_N^* \geq s) \leq \sum_c \mathbb{P}_{H_0}(X_1, \ldots, X_N \in A_c) \, \mathbf{1}_{\{N \log_2(1/p_c) - \kappa_c \geq s\}}.$$

Using the preceding bound,

$$\mathbb{P}_{H_0}(S_N^* \geq s) \leq \sum_c 2^{-s} 2^{-\kappa_c} \leq 2^{-s}.$$

This proves the theorem. □

Equivalently,

$$E_N = \sum_c 2^{-\kappa_c} p_c^{-N} \mathbf{1}_{\{X_1, \ldots, X_N \in A_c\}}$$

is an e-value, and $\log_2 E_N \geq S_N^*$.

# 5 Partial and noisy algebraic laws

Exact fit is often too strict. Define

$$H_c = \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in A_c\}}.$$

For $q, p \in [0, 1]$, let

$$d_2(q \mid p) = q \log_2 \frac{q}{p} + (1 - q) \log_2 \frac{1 - q}{1 - p}$$

be binary KL divergence in bits, with the usual boundary conventions.

For $0 < p_c < 1$, define

$$K_N(c) = N d_2(H_c/N \mid p_c) - \kappa_c - \log_2(N + 1)$$

if $H_c/N > p_c$, and define $K_N(c) = -\infty$ otherwise. Let

$$K_N^* = \sup_c K_N(c).$$

**Theorem 5.1** (Partial-fit post-selection significance)**.** *Under $H_0$, for every $s \geq 0$,*

$$\mathbb{P}_{H_0}(K_N^* \geq s) \leq 2^{-s}.$$

*Thus, if a reported law $c$ contains $h$ of $N$ data points, the score*

$$N d_2(h/N \mid p_c) - \kappa_c - \log_2(N + 1)$$

*is a valid post-selection evidence score whenever $h/N > p_c$.*

*Proof.* For fixed $c$, the random variable $H_c$ is stochastically dominated by

$$Z_c \sim \mathrm{Bin}(N, p_c).$$

For $q > p_c$, the type-counting Chernoff bound gives

$$\mathbb{P}(Z_c/N \geq q) \leq (N+1)2^{-Nd_2(q|p_c)}.$$

If $K_N(c) \geq s$, then

$$Nd_2(H_c/N \mid p_c) \geq s + \kappa_c + \log_2(N+1).$$

Hence

$$\mathbb{P}_{H_0}(K_N(c) \geq s) \leq 2^{-s}2^{-\kappa_c}.$$

A union bound over $c$ gives

$$\mathbb{P}_{H_0}(K_N^* \geq s) \leq \sum_c 2^{-s}2^{-\kappa_c} \leq 2^{-s}.$$

□

# 6 Likelihood-ratio mixtures and oracle bounds

For fixed $c$ with $0 < p_c < 1$ and for $\theta \in [p_c, 1]$, define

$$L_{c,N}(\theta) = \left(\frac{\theta}{p_c}\right)^{H_c} \left(\frac{1-\theta}{1-p_c}\right)^{N-H_c}.$$

For every $\theta \in [p_c, 1]$, $L_{c,N}(\theta)$ is an e-value under $H_0$.

Let

$$\theta_j = \frac{j}{N}, \qquad j = \lceil Np_c \rceil, \dots, N.$$

Define

$$M_{c,N} = \frac{1}{N+1} \sum_{j=\lceil Np_c \rceil}^{N} L_{c,N}(\theta_j).$$

**Proposition 6.1** (Robust GLR lower bound)**.** *The quantity $M_{c,N}$ is an e-value. If $H_c/N > p_c$, then*

$$\log_2 M_{c,N} \geq Nd_2(H_c/N \mid p_c) - \log_2(N+1).$$

*Proof.* A subprobability average of e-values is an e-value. If $H_c/N > p_c$, the grid contains $\theta = H_c/N$. Therefore

$$M_{c,N} \geq \frac{1}{N+1} L_{c,N}(H_c/N).$$

Taking logarithms,

$$\log_2 M_{c,N} \geq Nd_2(H_c/N \mid p_c) - \log_2(N+1).$$

□

Mixing over certificates yields the oracle inequality

$$\log_2 \sum_c 2^{-\kappa_c} M_{c,N} \geq \sup_c \left[Nd_2(H_c/N \mid p_c) - \kappa_c - \log_2(N+1)\right].$$

## 7 Conditional and sequential certificates

Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration and let $X_t$ be the observation at time $t$.

A predictable certificate $c$ specifies events

$$A_{c,t} \subseteq \mathcal{X}_t,$$

where $A_{c,t}$ is $\mathcal{F}_{t-1}$-measurable, and predictable numbers

$$0 < p_{c,t} < 1.$$

Assume the null satisfies

$$\mathbb{P}_{H_0}(X_t \in A_{c,t} \mid \mathcal{F}_{t-1}) \leq p_{c,t}$$

almost surely for every $c, t$.

Define

$$Z_{c,t} = \mathbf{1}_{\{X_t \in A_{c,t}\}}.$$

Let

$$\theta_{c,t} \in [p_{c,t}, 1]$$

be predictable. Define

$$L_{c,t} = \prod_{i=1}^{t} \left(\frac{\theta_{c,i}}{p_{c,i}}\right)^{Z_{c,i}} \left(\frac{1-\theta_{c,i}}{1-p_{c,i}}\right)^{1-Z_{c,i}}.$$

**Theorem 7.1** (Conditional-volume e-process)**.** *For every $c$ and every predictable strategy $(\theta_{c,t})$, the process $(L_{c,t})_{t\geq 0}$ is a nonnegative supermartingale under the null. Consequently,*

$$\mathbb{P}_{H_0}\left(\sup_{t\geq 0} L_{c,t} \geq 2^s\right) \leq 2^{-s}.$$

*Proof.* Let

$$q_t = \mathbb{P}_{H_0}(Z_{c,t} = 1 \mid \mathcal{F}_{t-1}) \leq p_{c,t}.$$

The conditional mean of the $t$-th factor is

$$q_t \frac{\theta_{c,t}}{p_{c,t}} + (1-q_t)\frac{1-\theta_{c,t}}{1-p_{c,t}}.$$

For $q \leq p \leq \theta$,

$$q\frac{\theta}{p} + (1-q)\frac{1-\theta}{1-p} = 1 - \frac{(p-q)(\theta-p)}{p(1-p)} \leq 1.$$

Thus

$$\mathbb{E}[L_{c,t} \mid \mathcal{F}_{t-1}] \leq L_{c,t-1}.$$

Ville's inequality gives the claimed optional-stopping bound. □

Now suppose $\mathcal{C}$ is prefix-free and, for each $c$, $\Pi_c$ is a countable family of predictable strategies with weights $w(\pi \mid c) \geq 0$ satisfying

$$\sum_{\pi\in\Pi_c} w(\pi \mid c) \leq 1.$$

Define

$$U_t = \sum_{c\in\mathcal{C}} 2^{-\kappa_c} \sum_{\pi\in\Pi_c} w(\pi \mid c) L_{c,\pi,t}.$$

**Corollary 7.2** (Universal anytime-valid search)**.** *The process* $(U_t)$ *is a nonnegative supermartingale with initial expectation at most* 1*. Hence*

$$\mathbb{P}_{H_0}\left(\sup_{t\geq 0}\log_2 U_t \geq s\right) \leq 2^{-s}.$$

# 8 Change-point discovery

We now make the change-point extension explicit. Suppose a certificate $c$ may become active after an unknown time $\tau$. For $t < \tau$, define

$$L_{c,\tau,t}(\theta) = 1.$$

For $t \geq \tau$, define

$$L_{c,\tau,t}(\theta) = \prod_{i=\tau}^{t}\left(\frac{\theta}{p_c}\right)^{Z_{c,i}}\left(\frac{1-\theta}{1-p_c}\right)^{1-Z_{c,i}},$$

where

$$Z_{c,i} = \mathbf{1}_{\{X_i\in A_c\}}.$$

Under the no-law null, this is an e-process that is inactive before $\tau$ and starts from value 1 at time $\tau - 1$.

Let $\Theta_c \subset [p_c, 1]$ be countable with weights $a_{\theta|c} \geq 0$, $\sum_{\theta\in\Theta_c} a_{\theta|c} \leq 1$. Let

$$w_\tau = \frac{1}{\tau(\tau+1)}, \qquad \sum_{\tau\geq 1} w_\tau = 1.$$

Define the full mixture over all start times by

$$U_t^{\mathrm{cp}} = \sum_c 2^{-\kappa_c}\sum_{\tau=1}^{\infty} w_\tau \sum_{\theta\in\Theta_c} a_{\theta|c} L_{c,\tau,t}(\theta).$$

**Theorem 8.1** (Change-point algebraic discovery)**.** *Under the null,* $(U_t^{\mathrm{cp}})_{t\geq 0}$ *is a nonnegative supermartingale with initial expectation at most* 1*. In particular,*

$$\mathbb{P}_{H_0}\left(\sup_{t}\log_2 U_t^{\mathrm{cp}} \geq s\right) \leq 2^{-s}.$$

*Proof.* For each fixed $c, \tau, \theta$, the process $L_{c,\tau,t}(\theta)$ is identically 1 for $t < \tau$, has a conditionally valid first likelihood factor at $t = \tau$, and then evolves by conditionally valid likelihood factors. Hence it is a nonnegative supermartingale. The mixture weights satisfy

$$\sum_c 2^{-\kappa_c} \leq 1, \qquad \sum_{\tau\geq 1} w_\tau = 1, \qquad \sum_{\theta\in\Theta_c} a_{\theta|c} \leq 1.$$

Therefore their weighted sum is a nonnegative supermartingale with initial expectation at most 1. Ville's inequality gives the result. □

For computation at time $t$, the infinite tail $\tau > t$ contributes only its inactive value. It may therefore be evaluated as a deterministic tail mass, or truncated with the remaining tail kept as an inactive reserve.

## 9 Power under planted algebraic incidence

Let $P$ be an alternative distribution on $X$. For certificate $c$, set

$$\theta_c = P(X \in A_c).$$

Assume

$$\theta_c > p_c.$$

**Theorem 9.1** (Linear power)**.** *Under i.i.d. sampling from $P$,*

$$\frac{1}{N} K_N(c) \to d_2(\theta_c \mid p_c)$$

*almost surely. Consequently, for every sequence $s_N = o(N)$,*

$$\mathbb{P}_P(K_N^* \geq s_N) \to 1.$$

*Proof.* By the strong law,

$$H_c/N \to \theta_c$$

almost surely. Since $q \mapsto d_2(q \mid p_c)$ is continuous for $q > p_c$,

$$d_2(H_c/N \mid p_c) \to d_2(\theta_c \mid p_c) > 0.$$

The penalty terms satisfy

$$\kappa_c/N \to 0, \qquad \log_2(N+1)/N \to 0.$$

Therefore

$$K_N(c)/N \to d_2(\theta_c \mid p_c).$$

Since $K_N^* \geq K_N(c)$, the result follows. □

## 10 Finite-class model-selection consistency

The score $K_N(c)$ equals $-\infty$ whenever $H_c/N \leq p_c$. Therefore, for consistency statements, it is cleaner to use the nonnegative score

$$K_N^+(c) = \max\left\{0,\, N d_2(H_c/N \mid p_c) - \kappa_c - \log_2(N+1)\right\},$$

where the KL term is interpreted as 0 when $H_c/N \leq p_c$.

Let $\mathcal{C}_0 \subset \mathcal{C}$ be finite. Under an alternative distribution $P$, define

$$\theta_c = P(X \in A_c)$$

and

$$J(c) = \begin{cases} d_2(\theta_c \mid p_c), & \theta_c > p_c, \\ 0, & \theta_c \leq p_c. \end{cases}$$

Assume $c_*$ is the unique maximizer of $J(c)$ over $\mathcal{C}_0$, and assume

$$J(c_*) > 0.$$

Let

$$\widehat{c}_N \in \arg\max_{c \in \mathcal{C}_0} K_N^+(c).$$

**Theorem 10.1** (Finite-class consistency)**.** *Under $P$,*

$$\mathbb{P}(\widehat{c}_N = c_*) \to 1.$$

*Proof.* For each fixed $c$,

$$K_N^+(c)/N \to J(c)$$

almost surely. Indeed, if $\theta_c > p_c$, this follows from the same argument as in Theorem 9.1. If $\theta_c \le p_c$, then positive deviations above $p_c$ vanish asymptotically in normalized KL score, and $K_N^+(c)/N \to 0$.

Since $\mathcal{C}_0$ is finite and $c_*$ uniquely maximizes $J$, there is a margin

$$\gamma = J(c_*) - \max_{c \neq c_*} J(c) > 0.$$

Almost surely, for all sufficiently large $N$,

$$K_N^+(c_*)/N > J(c_*) - \gamma/3,$$

while

$$K_N^+(c)/N < J(c) + \gamma/3 \le J(c_*) - 2\gamma/3$$

for every $c \neq c_*$. Thus $c_*$ eventually uniquely maximizes $K_N^+$. □

# 11 Replicated laws

Suppose there are environments $e = 1, \ldots, E$. In environment $e$, observations satisfy a null incidence certificate

$$\mathbb{P}(X_{e,t} \in A_c \mid \mathcal{F}_{e,t-1}) \le p_{c,e,t}.$$

There are two valid settings.

## 11.1 Independent environments

If the environments are independent under the null, then the product of environment-specific e-values is an e-value.

## 11.2 Ordered conditional validity

More generally, order all observations in a single global filtration. If each environment-specific factor is conditionally valid given the past, then the product process is a supermartingale.

Under either condition, if environment $e$ supplies $N_e$ observations, hit fraction $\widehat{q}_{c,e}$, and constant certificate $p_{c,e}$, the replicated score

$$R(c) = \sum_{e=1}^{E} N_e d_2(\widehat{q}_{c,e} \mid p_{c,e}) - \kappa_c - \sum_{e=1}^{E} \log_2(N_e + 1)$$

is post-selection valid after mixing over $c$. The model cost $\kappa_c$ is paid once, while evidence accumulates across environments.

# 12 Simultaneous effect-size confidence sets

Let $q_c = P(X \in A_c)$ be the true incidence probability. For $u \in [0,1]$, let $M_{c,N}(u)$ be an e-value valid under the null

$$H_{0,c}(u) : q_c \le u.$$

Define the confidence set

$$\mathcal{I}_{c,N}(\alpha) = \left\{ u \in [0,1] : M_{c,N}(u) < \frac{2^{\kappa_c}}{\alpha} \right\}.$$

**Theorem 12.1** (Simultaneous post-selection confidence sets)**.** *With probability at least* $1-\alpha$*,*

$$q_c \in \mathcal{I}_{c,N}(\alpha)$$

*for every* $c$*.*

*Proof.* For the true value $q_c$,

$$\mathbb{P}\left( M_{c,N}(q_c) \ge \frac{2^{\kappa_c}}{\alpha} \right) \le \alpha 2^{-\kappa_c}.$$

Union-bound over $c$ and use Kraft:

$$\sum_c \alpha 2^{-\kappa_c} \le \alpha.$$

□

We call these confidence sets, not intervals; they need not be intervals without further monotonicity assumptions.

# 13 Information-theoretic optimality

Let $P_0$ be a reference distribution and let $A \subseteq X$ have exact null mass $P_0(A) = p$. For $\theta \in (0,1)$, define the tilted alternative

$$\frac{dP_\theta}{dP_0}(x) = \begin{cases} \theta/p, & x \in A, \\ (1-\theta)/(1-p), & x \notin A. \end{cases}$$

Then $P_\theta(A) = \theta$.

For $N$ observations, the likelihood ratio is

$$L_N(\theta) = \left(\frac{\theta}{p}\right)^H \left(\frac{1-\theta}{1-p}\right)^{N-H}.$$

Also,

$$D_2(P_\theta \mid P_0) = d_2(\theta \mid p).$$

If a test has type-I error at most $\alpha$ under $P_0^N$ and type-II error at most $\beta$ under $P_\theta^N$, data processing gives

$$N d_2(\theta \mid p) \ge d_2(1-\beta \mid \alpha).$$

Thus

$$N \ge \frac{d_2(1-\beta \mid \alpha)}{d_2(\theta \mid p)}.$$

The KL scores above detect at the same information rate, up to description length and universal coding regret.

## 14 Necessity of search penalties

Let $A_1, \dots, A_M$ be disjoint subsets of $X$, each with null mass $p$. Assign each model code length

$$\kappa = \log_2 M.$$

Kraft is tight:

$$\sum_{j=1}^{M} 2^{-\kappa} = 1.$$

The event that all $N$ samples fall into one of the $A_j$ has probability

$$Mp^N.$$

For each such model, the exact-fit score is

$$S = N \log_2(1/p) - \log_2 M.$$

Thus

$$Mp^N = 2^{-S}.$$

Hence the bound in Theorem 4.1 is sharp in the abstract volume-certificate model, and the search penalty cannot be uniformly removed.

## 15 Universal dominance

Let $\mathcal{E}$ be a computably enumerable class of elementary certified e-values or e-processes. Let $\mathbf{m}(e)$ be a universal lower-semicomputable semimeasure on $\mathcal{E}$. Define

$$U = \sum_{e \in \mathcal{E}} \mathbf{m}(e) e.$$

A certified e-test is any lower-semicomputable subprobability mixture

$$E = \sum_{e \in \mathcal{E}} w(e) e,$$

where $w(e) \geq 0$, $\sum_e w(e) \leq 1$, and $w$ is lower semicomputable.

**Theorem 15.1** (Universal dominance)**.** *For every certified e-test $E$, there exists a constant $C_E$, depending only on the effective description of $E$, such that*

$$E \leq C_E U$$

*pointwise. Equivalently,*

$$\log_2 E \leq \log_2 U + O(K(E)).$$

*Proof.* By universality of $\mathbf{m}$, every lower-semicomputable semimeasure $w$ satisfies

$$w(e) \leq C_E \mathbf{m}(e)$$

for all $e$, for some constant $C_E$. Hence

$$E = \sum_e w(e) e \leq C_E \sum_e \mathbf{m}(e) e = C_E U.$$

□

This theorem characterizes the precise class being dominated: effective tests built as mixtures of certified geometric evidence. It does not claim dominance over all possible statistical tests.

# 16 Algebraic certificates over height balls

Let

$$X_B = \mathbb{P}^n(\mathbb{Q})_{\le B}$$

be a finite height ball in projective space. An algebraic certificate is a proper closed algebraic subset

$$Y \subsetneq \mathbb{P}^n_{\mathbb{Q}}$$

together with a bound

$$|Y(\mathbb{Q}) \cap X_B| \le V_B(Y).$$

Then

$$p_Y = \min\left\{1, \frac{V_B(Y)}{|X_B|}\right\}.$$

The statistical machinery is independent of how $V_B(Y)$ is obtained.

## 16.1 Hypersurfaces

Let $F \in \mathbb{Z}[X_0, \ldots, X_n]$ be nonzero homogeneous of degree $D$, and let $Y = Z(F)$. A crude box-counting Schwartz-Zippel argument gives

$$|Y(\mathbb{Q}) \cap X_B| \le C_n D B^n.$$

Since

$$|X_B| \gg_n B^{n+1},$$

one obtains

$$p_F \le \min\{1, C_n D/B\}.$$

Thus an exact degree-$D$ homogeneous polynomial law has leading score

$$N \log_2(B/D) - \kappa(F) - O_n(N).$$

## 16.2 Linear subspaces

If $L \subseteq \mathbb{P}^n$ is a rational $m$-plane, then

$$|L(\mathbb{Q}) \cap X_B| \ll_m B^{m+1}.$$

Thus

$$p_L \le C_{n,m} B^{-(n-m)}.$$

For a union of $K$ rational $m$-planes,

$$p \le C_{n,m} K B^{-(n-m)}.$$

### 16.3 General projective varieties

Let $Y \subsetneq \mathbb{P}^n_{\mathbb{Q}}$ be a reduced closed algebraic subset of dimension $m \geq 1$ and total degree $D$. A finite-projection argument, or sharper determinant-method estimates when available, can supply bounds of the schematic form

$$|Y(\mathbb{Q}) \cap X_B| \leq C_n D^{O_n(1)} B^{m+1}.$$

Consequently,

$$p_Y \leq C_n D^{O_n(1)} B^{-(n-m)}.$$

For $m = 0$,

$$|Y(\mathbb{Q}) \cap X_B| \leq D,$$

and

$$p_Y \leq C_n D B^{-(n+1)}.$$

These bounds are intentionally crude and may be replaced by sharper arithmetic estimates.

### 16.4 Finite fields

Let $X = \mathbb{F}_q^d$. If $F \in \mathbb{F}_q[x_1, \ldots, x_d]$ is nonzero of degree $D < q$, then Schwartz-Zippel gives

$$|Z(F)| \leq D q^{d-1},$$

so

$$p_F \leq D/q.$$

## 17 Application I: adaptive symbolic regression over rational data

Let a symbolic-regression procedure search through homogeneous polynomials and report

$$F \in \mathbb{Z}[X_0, \ldots, X_n],$$

of degree $D$ and code length $\kappa(F)$, vanishing on $h$ of $N$ rational points in $X_B$.

Using

$$p_F \leq C_n D/B,$$

the valid post-selection score is

$$N d_2 \left( \frac{h}{N} \,\middle|\, \min\{1, C_n D/B\} \right) - \kappa(F) - \log_2(N+1).$$

For exact fit, this becomes

$$N \log_2 \frac{B}{C_n D} - \kappa(F).$$

Therefore a discovered degree-$D$ polynomial law is significant only if

$$N \log_2(B/D) \gg \kappa(F) + s.$$

This is a finite-sample anti-overfitting threshold.

## 18 Application II: algebraic subspace clustering

Suppose an algorithm discovers

$$Y = L_1 \cup \cdots \cup L_K,$$

a union of rational $m$-planes in $\mathbb{P}^n$. Then

$$p_Y \le C_{n,m} K B^{-(n-m)}.$$

If all $N$ observations lie in the discovered union, the exact-fit score is

$$N \left[(n-m)\log_2 B - \log_2 K - O_{n,m}(1)\right] - \kappa(Y).$$

Thus increasing the number of fitted subspaces is penalized both by increased null volume and by model description length. Partial-fit versions use the KL score and allow outliers.

## 19 Application III: finite-field congruence laws

Let

$$X = \mathbb{F}_q^d$$

under the uniform null. Let $F \in \mathbb{F}_q[x_1, \ldots, x_d]$ be nonzero of degree $D < q$. Since

$$p_F \le D/q,$$

an exact discovered congruence law

$$F(X_i) = 0 \qquad i = 1, \ldots, N$$

has score

$$N \log_2(q/D) - \kappa(F).$$

If the same integer polynomial law is tested across fields

$$\mathbb{F}_{q_1}, \ldots, \mathbb{F}_{q_E}$$

and remains nonzero modulo each $q_e$, then the replicated exact score is

$$\sum_e N_e \log_2(q_e/D) - \kappa(F),$$

under the independence or conditional-validity assumptions of Section 11.

## 20 Additional examples

The same framework applies whenever a valid volume certificate is available.

### 20.1 Low-rank matrices

Nonzero $a \times b$ rational matrices form a projective space $\mathbb{P}^{ab-1}$. The rank-$\le r$ determinantal variety has codimension

$$(a-r)(b-r).$$

For fixed $a, b, r$, its degree is constant, hence

$$p_r \le C_{a,b,r} B^{-(a-r)(b-r)}.$$

Exact low-rank structure across $N$ observations has leading evidence

$$N(a-r)(b-r)\log_2 B - \kappa(r) - O_{a,b,r}(N).$$

### 20.2 Conservation laws

Let $X_t$ be a state process on a bounded integer box. A candidate invariant $I$ gives events

$$A_{I,t} = \{x : I(x) = I(X_1)\}.$$

If one can certify

$$\mathbb{P}(I(X_t) = I(X_1) \mid \mathcal{F}_{t-1}) \leq p_I,$$

then the conditional e-process theorem applies. The scientific content lies in justifying this null certificate.

### 20.3 Missing data

If only coordinates indexed by a predictable pattern $\Omega_t$ are observed, replace $Y$ by its projection

$$\pi_{\Omega_t}(Y).$$

The framework applies once one certifies

$$\mathbb{P}_0(\pi_{\Omega_t}(X_t) \in \pi_{\Omega_t}(Y) \mid \mathcal{F}_{t-1}) \leq p_{Y,t}.$$

## 21 Multiple discoveries

If several independent or dependent discovery projects produce valid e-values

$$E_1, \ldots, E_J,$$

then e-value multiple-testing procedures such as e-BH may be used to control false discovery rate under their stated assumptions. This allows many algebraic-discovery pipelines to be run in parallel, provided each reported law is accompanied by a valid e-value.

## 22 Scope and limitations

This framework does not produce volume certificates automatically. It converts valid certificates into post-selection-valid inference.

The framework is not a new theorem in rational-point counting. Any sharper arithmetic or geometric estimate can be substituted for the crude certificates stated here.

The universal dominance theorem is relative to the chosen effective certificate language. It is not a dominance theorem over all possible statistical tests.

The algebraic null model must be meaningful for the application. Uniform height balls and finite fields are convenient reference cases, but applications may require conditional or nonuniform volume certificates.

The term “model-selection consistency” in this paper is finite-class and incidence-based. It does not assert recovery of an algebraic variety in a metric or scheme-theoretic sense.

## 23 Conclusion

We have developed a certified inference framework for algebraic laws discovered after adaptive search. The core mechanism is a volume certificate combined with a prefix-free model code. From this we obtain finite-sample post-selection validity, noisy and partial-fit validity, optional-stopping validity, finite-class consistency, replicated-law aggregation under explicit assumptions, change-point mixtures, and universal dominance over effective certified e-tests.

The main practical implication is an anti-overfitting law for algebraic discovery. A polynomial equation, subspace union, congruence law, or algebraic constraint discovered after inspecting the data is significant only when its certified volume reduction beats both its description length and its geometric complexity.

For symbolic regression over rational points, a degree-$D$ exact polynomial law receives leading evidence

$$N \log_2(B/D) - \kappa(F) - O_n(N).$$

Thus high-degree interpolation is not significant merely because it fits; the equation must be simple and its null volume must be small.

The framework is best viewed as a rigorous MDL/e-value layer for algebraic and arithmetic discovery. Its strength depends on the quality of the supplied volume certificates.

## References


[SSVV11] G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk, *Test martingales, Bayes factors and p-values*, Statistical Science 26 (2011), no. 1, 84–101.

[VW21] V. Vovk and R. Wang, *E-values: calibration, combination and applications*, Annals of Statistics 49 (2021), no. 3, 1736–1754.

[WR22] R. Wang and A. Ramdas, *False discovery rate control with e-values*, Journal of the Royal Statistical Society, Series B 84 (2022), no. 3, 822–852.

[RSSS23] A. Ramdas, P. Grunwald, V. Vovk, and G. Shafer, *Game-theoretic statistics and safe anytime-valid inference*, Statistical Science 38 (2023), no. 4, 576–601.

[GSTV01] P. Gacs, J. Tromp, and P. Vitanyi, *Algorithmic statistics*, IEEE Transactions on Information Theory 47 (2001), no. 6, 2443–2463.

[VV04] N. Vereshchagin and P. Vitanyi, *Kolmogorov's structure functions and model selection*, IEEE Transactions on Information Theory 50 (2004), no. 12, 3265–3290.

[Grunwald07] P. Grunwald, *The Minimum Description Length Principle*, MIT Press, 2007.

[LiV] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer.

[Poonen17] B. Poonen, *Rational Points on Varieties*, Graduate Studies in Mathematics 186, American Mathematical Society, 2017.

[BHBS06] T. D. Browning, D. R. Heath-Brown, and P. Salberger, *Counting rational points on algebraic varieties*, Duke Mathematical Journal 132 (2006), no. 3, 545–578.

[HB02] D. R. Heath-Brown, *The density of rational points on curves and surfaces*, Annals of Mathematics 155 (2002), no. 2, 553–595.

[Salberger12] P. Salberger, *On the density of rational and integral points on algebraic varieties*, Journal fur die reine und angewandte Mathematik 606 (2007), 123–147.

[Hartshorne77] R. Hartshorne, *Algebraic Geometry*, Springer, 1977.

[Harris92] J. Harris, *Algebraic Geometry: A First Course*, Springer, 1992.

[DSS09] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on Algebraic Statistics*, Birkhauser, 2009.

[Watanabe09] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, 2009.

[VMS16] R. Vidal, Y. Ma, and S. Sastry, *Generalized Principal Component Analysis*, Springer, 2016.

[Landsberg12] J. M. Landsberg, *Tensors: Geometry and Applications*, American Mathematical Society, 2012.

[KTT15] A. Kiraly, L. Theran, and R. Tomioka, *The algebraic combinatorial approach for low-rank matrix completion*, Journal of Machine Learning Research 16 (2015), 1391–1436.

[PBM16] J. Peters, P. Buhlmann, and N. Meinshausen, *Causal inference by using invariant prediction: identification and confidence intervals*, Journal of the Royal Statistical Society, Series B 78 (2016), no. 5, 947–1012.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, ChatGPT, by OpenAI, was used to assist with mathematical drafting, formalization, review, and editing. This work is shared as a preliminary AI-assisted mathematical note. The mathematical content may have been only partially reviewed and may contain errors; it should not be treated as peer-reviewed or as a fully verified manuscript.