



Paper Type: Original Article

Interpretable Population Dynamics: Causal Inference for Engineering Optimization Decisions

Rania A. Mahmoud^{1,*} , Kwame A. Mensah²

¹ Department of Computer Science, Alexandria University, Alexandria, Egypt; r.mahmoud@alexu.edu.eg.

² Department of Computer Science, University of Cape Coast, Cape Coast, Ghana.

Citation:

Received: 08 January 2026

Revised: 11 March 2026

Accepted: 28 June 2026

Mahmoud, R. A., & Mensah, K. A. (2026). Interpretable population dynamics: Causal inference for engineering optimization decisions. *Metaheuristic Algorithms with Applications*, 3(2), 168-194.

Abstract

Engineering teams increasingly ask of an optimization algorithm more than a good design: They ask for a justification of it. This paper presents Interpretable Population Dynamics (IPD), a metaheuristic for optimization over Riemannian manifolds in which population evolution is governed by an explicitly modeled causal system, so that the algorithm can audit and explain its own search behavior. Within IPD, geodesic variation operators act as interventions on a causal graph of search dynamics, scheduling selects the action with maximal predicted counterfactual progress, and inverse-probability-weighted estimation on logged trajectories quantifies the Average Treatment Effect (ATE) of every algorithmic factor. The experimental program evaluates the method on ten problems spanning four manifold species, the sphere, the Stiefel manifold, the Symmetric Positive Definite (SPD) cone, and $SO(3)$, against twelve racing competitors, including newly admitted Riemannian trust-region, adaptive stochastic, and quasi-Newton rivals, over 21 independent runs under a 12,000-evaluation budget. IPD attains the smallest median optimality gap on seven of the ten problems, with a median reduction of 26.2% against the strongest rival on those problems; exact paired permutation tests with Benjamini-Hochberg control confirm 94 of 100 contrasts, while Riemannian CMA-ES retains the two smooth SPD instances and a Riemannian trust-region method retains the best-conditioned Stiefel instance. In the audit leg, an honest Causal Forest reproduces the audit's effect ordering with Spearman correlation 0.90, whereas naive correlational attribution overstates the dominant operator's effect by 43.7%. Grouping by curvature regime locates the scheduling gain chiefly on positive-curvature multimodal instances, two one-sided equivalence tests certify wall-clock parity with the fastest baselines wherever objective evaluation dominates, and a regime surrogate with 85.9% cross-validated accuracy converts logged trajectories into decision-ready evidence.

Keywords: Interpretable optimization, Causal inference, Population dynamics, Riemannian manifolds, Metaheuristics, Explainable artificial intelligence.

1 | Introduction

When engineering teams entrust hard design choices to an optimization algorithm, responsibility travels with the design, because the people who sign off on the result answer for outcomes they did not directly compute.

 Corresponding Author: r.mahmoud@alexu.edu.eg

 <https://doi.org/10.48313/maa.vi.77>



Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

Whenever a metaheuristic proposes a sensor layout or a learned metric, the decision-maker reasonably asks why the search settled where it did, which mechanisms produced the observed progress, and whether more budget would change the answer. Provided that the objective admits few assumptions and the application demands graceful degradation on multimodal or noisy landscapes, population-based methods remain attractive in practice [1], [2]. Their internal decision processes, however, resist inspection: Credit for improvement spreads implicitly across operators, parameters, and chance, and no standard metaheuristic documented to date produces evidence about its own search that would survive design review.

A growing share of engineering optimization problems is most naturally stated on Riemannian manifolds, because direction vectors live on spheres, orthogonal and low-rank factors on the Stiefel manifold, and covariance and diffusion structures in the cone of Symmetric Positive Definite (SPD) matrices [3], [4]. Treating such a problem in ambient Euclidean space distorts geometry, forces repair or projection steps, and can alter the landscape being searched. For single-trajectory methods, the literature offers mature machinery, namely tangent spaces, retractions, and geodesic steps; population-based search on manifolds has received far less attention, and, so far as the record establishes, no method of either kind explains its own dynamics.

The explainable-AI toolkit, from Shapley-value regressors [5] and local surrogates [6] to the broader attribution literature [7], takes predictors rather than iterative search processes as its object. Pointed at optimizer logs, such tools return correlational statements: An operator co-occurred with progress, a parameter tracked the fitness curve. Within a feedback loop, such associations are treacherous, because the algorithm itself decides, as a function of the current state, when each operator is applied. It follows that the observed association between an operator and subsequent progress mixes the operator's true contribution with the circumstances that triggered its use.

The position defended here is that the search process itself should be the object of causal analysis [8]. Under this design, a small set of sufficient statistics summarizes the evolving population, every operator application acts as an intervention carrying an explicit dose on the resulting stochastic process, and a structural causal model, the causal graph of search dynamics, links doses to state transitions. Inverse-probability weighting on logged trajectories estimates the effects of algorithmic factors, after which the same model serves prospectively: The scheduler chooses each iteration's action by predicting counterfactual progress under candidate interventions. Explanation thereby ceases to be an afterthought and becomes the control signal of the optimizer.

The empirical portion of this work is organized around three questions, each matched to one falsifiable hypothesis. The questions separate the optimization claim, which concerns what Intervention-Aware Scheduling (IAS) buys over operator-blind search at equal budget, the identification claim, which concerns whether the population-dynamics effects are estimable under the stated assumptions, and the transfer claim, which concerns whether the estimates survive changes of manifold and dimension. The three claims are logically independent, so none can stand on the strength of another. Sections 6.4 through 6.8 supply the evidence for each hypothesis.

RQ1: Does intervention-aware causal operator scheduling improve manifold optimization relative to operator-blind competitors under a shared evaluation budget?

RQ2: Are the population-dynamics causal effects identifiable and consistently estimable under the stated assumptions, as the causal audit, now contested by an independent estimator, verifies?

RQ3: Do the estimated intervention effects transfer across curvature regimes and dimensions, as the grouping and ladder studies of Section 6.8 test?

H1. At equal budget, Interpretable Population Dynamics (IPD) attains a positive median gap reduction against the best-performing of the ten optimizer rivals on the majority of the ten problems and holds the smallest median gap on at least seven of them.

H2. The audit reports positive Average Treatment Effects (ATEs) whose Confidence Intervals (CIs) exclude zero, an independent causal estimator reproduces their ordering, and naive correlational attribution departs materially from both.

H3. Neither curvature regime nor dimension moderates the intervention-scheduling effect, so the grouped and scaled estimates agree with the pooled estimate within their confidence-interval widths.

The contributions of this work are as follows:

- I. A causal formulation of population-based search on Riemannian manifolds, with operator doses entering as interventions on a structural model of population-dynamics state transitions (Section 3).
- II. The IPD algorithm itself, which couples geodesic variation operators, namely covariance-adapted tangent mutation, Geodesic Crossover (GC), and retraction-based local search, with IAS driven by counterfactual simulation over a five-iteration horizon (Section 4).
- III. An estimation methodology converting logged trajectories into ATEs with bootstrap CIs, together with a quantitative contrast against naive correlational attribution that indicates confounding-induced overestimation of 43.7% for the dominant operator, cross-checked by an independent causal estimator (Sections 4.5 and 6.5).
- IV. A theoretical treatment covering convergence in probability to the global optimum on manifolds via a geodesic-distance potential function, consistency for the causal estimator under state-sufficient logging, and per-iteration complexity including estimation overhead (Section 5).
- V. An empirical program on ten manifold problems spanning four manifold species, against twelve racing rivals, including Riemannian trust-region, adaptive stochastic, and quasi-Newton newcomers, and one rival causal estimator in the audit leg, over 21 runs per pairing under a 12,000-evaluation budget, with exact permutation inference, false-discovery-rate control, and equivalence testing of the overhead claim, and with estimator-agreement, curvature-sensitivity, and cross-dimension analyses (Section 6).

The remainder of the paper proceeds as follows. Section 2 reviews optimization on manifolds, causal inference, and explainability for search, and states the research gap; Section 3 formalizes manifold-constrained problems, the population state space, and the causal graph of search dynamics; and Section 4 presents the IPD algorithm together with its explainability pipeline, culminating in *Algorithm 1*. Section 5 develops the convergence, consistency, and complexity results, while Section 6 reports the experimental program: optimization results under exact permutation inference, the causal audit contested by an independent estimator, an intervention-attribution case study, ablations, and curvature and cross-dimension analyses. Section 7 discusses how decision-makers should interpret the outputs and when the estimates can fail; Section 8 lists limitations; and Section 9 concludes.

2 | Related Work

Three literatures bear on the present work: Riemannian optimization supplies the geometry, causal inference supplies the identification and estimation machinery, and interpretable machine learning supplies the interface between numbers and decisions. The discussion reviews each literature in turn and then states the gap that IPD addresses.

2.1 | Optimization on Riemannian Manifolds

The recognition that orthogonality and rank constraints induce a nontrivial search geometry dates to early numerical linear algebra, and the view of constraint sets as manifolds equipped with tangent-space calculus was systematized by Edelman et al. [9]. Earlier still, Smith had introduced Riemannian optimization techniques for eigenvalue-type problems [10], while Nishimori and Akaho [11] developed learning algorithms following quasi-geodesic flows on the Stiefel manifold. Around retractions, which act as first-order surrogates for the Riemannian exponential, Absil et al. [4] consolidated the field in a book-length treatment of line-search and trust-region methods on embedded manifolds. Three later contributions sharpened the picture: Bonnabel

[12] analyzed stochastic gradient descent on manifolds, Absil et al. [13] introduced Riemannian trust-region methods, and Boumal et al. [14] established global convergence rates for nonconvex manifold optimization.

On the SPD cone, an affine-invariant geometry whose distance is induced by matrix logarithms arises naturally [15], [16]; Forstner and Moonen [17] studied a comparable metric structure for covariance matrices; and the cone underlies applications from information-theoretic metric learning [18] to covariance estimation. Sensor-network localization, a representative nonconvex least-squares problem over products of spheres whose semidefinite relaxations are well understood [19], illustrates the engineering setting in which such products arise. By comparison, population-based search on manifolds remains underdeveloped: Geodesic variants of particle swarm optimization built on the classical dynamics of Kennedy and Eberhart [20] exist, as do manifold adaptations of evolution strategies, including the tangent-space realization of Covariance Matrix Adaptation (CMA) [21] that serves as the strongest baseline here, yet both inherit the black-box character of their Euclidean ancestors. None of these methods maintains a model of why its own population moves as it does, and that vacancy is precisely what this paper addresses.

2.2 | Causal Modeling and Inference in Computational Intelligence

Two interchangeable languages underpin modern causal inference. In the potential-outcomes tradition, effects are contrasts between hypothetical worlds in which the same unit receives different treatments [22], [23]; in the structural tradition, data-generating processes are graphs manipulated with the do-operator [8], [24]; and modern textbook treatments consolidate both languages while stating their assumptions explicitly [25]. Graphical model discovery addresses the harder problem of learning causal structure from observational data [26]. For time series, Granger's [27] notion of predictive causation is known to confuse influence with shared history, a caution reinforced in observational network settings [28] and one that motivates modern time-series causal discovery [29], [30].

Classically, estimation from observational data proceeds by balancing treatment assignment through propensity scores [31], whose inverse-probability form descends from the Horvitz-Thompson [32] estimator of survey sampling; difference-in-differences instead exploits policy changes over time [33], and the assumption audits these estimators demand are codified in dedicated treatises [34]. Within computational intelligence, causal vocabulary appears almost exclusively for external systems, for example feature selection, spatio-temporal forecasting, or neural architecture choice, whereas the algorithms doing the selecting are themselves analyzed, if at all, with purely descriptive statistics. No metaheuristic in this literature maintains a causal model of its own dynamics, estimates the effects of its own decisions, and schedules itself by interventions; filling that vacancy is the purpose of Sections 3 and 4.

2.3 | Explainability for Search and Optimization

Interpretable machine learning favors models whose structure can be read directly, a design philosophy surveyed by Rudin [35], with classification and regression trees [36] and rule induction [37] as the workhorses. That intelligibility matters in high-stakes deployment was made concrete by the pneumonia-risk study of Caruana et al. [38]. Saliency methods such as Grad-CAM [39] explain visual classifiers but admit no natural analogue for a cloud of points on a manifold. Applied to optimization logs, post-hoc attribution tools such as Shapley regressors and local surrogates [5], [6] inherit the correlational character of their estimands, yet are routinely presented as if they established contribution rather than co-occurrence.

An older tradition of credit assignment exists within evolutionary computation itself. Holland's [40] adaptive plans, CMA in evolution strategies [21], and the self-adapting control parameters of differential evolution [41] all redistribute credit for observed improvement; adaptive operator selection formalizes the task as a multi-armed bandit over operators [42], and parameter control is surveyed by Karafotias et al. [43]. These mechanisms learn from reward rather than from interventions: Because they condition on the very state that also caused the assignment, their credit estimates carry confounding by construction, and users rarely see

them as auditable evidence. IPD retains the adaptivity while replacing observed-reward credit assignment with intervention-based effect estimation, and it elevates the resulting estimates to first-class outputs of the search.

2.4 | Research Gap and Positioning

Since the no-free-lunch theorems render algorithm performance instance-specific [44], the case strengthens for methods that can report what happened in a particular run rather than what happens on average; sound statistical comparison practice likewise demands per-instance, distribution-aware evidence [45]. Read jointly, the three reviews above expose a vacancy: No existing method is simultaneously: 1) manifold-native, 2) equipped with a causal model of its own dynamics, 3) capable of built-in attribution with uncertainty quantification, 4) scheduled by interventions, and 5) covered by convergence analysis. *Table 1* records this positioning against representative neighbors.

Table 1. Positioning of IPD relative to representative prior approaches. Entries indicate that the property is native to the method rather than achievable only through external post-hoc analysis.

Method	Manifold-Native	Causal model of Own Dynamics	Built-In Attribution	Intervention-Based Scheduling	Convergence Analysis
R-CMA-ES [21]	Yes	No	No	No	Partial (inherited from Euclidean ancestor)
Geodesic PSO [20]	Yes	No	No	No	No
Tangent-space DE [41]	Yes (tangent embedding)	No	No	No	No
SHAP-on-EC, post-hoc [5], [7]	No	No	Correlational only	No	n/a
IPD (this work)	Yes	Yes	Yes (ATEs with CIs)	Yes	Yes (global, in probability)

*ATE, CI, n/a: Not applicable to a post-hoc analysis tool.

Each closest neighbor forgoes at least two of the five properties. R-CMA-ES is manifold-native and, as Section 6 confirms, the strongest baseline in practice; its Covariance Adaptation (CA), however, is correlational credit assignment learned from observed reward, and it offers no account of its own dynamics. Geodesic PSO and tangent-space differential evolution move correctly on the manifold while inheriting the opacity of their Euclidean ancestors. Post-hoc attribution applied to optimizer logs, the SHAP-on-EC row of *Table 1*, yields correlational statements without identification, a distortion quantified in Section 6.5. Provided that this reading of the literature is correct, IPD is the first population-based optimizer to hold all five properties at once; whether the design costs or pays for itself in optimization quality is tested in Section 6.

3 | Problem Formulation on Manifolds

This section sets up the formal objects: The manifold-constrained problem, the population state space with its sufficient statistics, and the causal graph that turns operator scheduling into intervention design. Notation is kept minimal, so that the estimation methodology of Section 4.5 and the consistency argument of Section 5.2 can be stated without auxiliary machinery.

3.1 | Manifold-Constrained Optimization

Writing M for a smooth embedded submanifold of $\mathbb{R}^n$, and $T_X M$ for the tangent space at each point X in M , the problem takes the following form. Throughout, $X_i(t)$ denotes the i -th element of a population on M , and F is the objective to be minimized over the manifold:

$$\min_{X \in M} F(X). \quad (1)$$

A retraction $R_X: T_X M \rightarrow M$ organizes moves on M ; it sends a tangent vector G to a feasible point while agreeing with the Riemannian exponential to first order, so that $R_X(0) = X$ and the differential of R_X at 0

is the identity [4]. Because retractions cost far less than exact exponentials while preserving descent structure, they serve as the workhorse of manifold optimization. Four retraction families cover the benchmark suite: On the sphere $S^{(d-1)}$ and on products of spheres such as the oblique manifold, the normalization retraction $R_X(G) = (X + G) / \|X + G\|$, applied factor by factor; on the Stiefel manifold $St(p, n)$ of orthonormal p -frames, the Q factor of the QR decomposition of $X + G$; on the SPD cone, the matrix-exponential retraction given explicitly in Eq. (7) below; and on $SO(3)$, the Rodrigues exponential map, which is exact and inexpensive in three dimensions. F is assumed continuously differentiable on M , with exact evaluations; Section 9 discusses noisy extensions.

Because single-trajectory manifold methods rely on Riemannian gradients, they can stall on multimodal or ill-conditioned landscapes. A population $X_1, \dots, X_N$, by contrast, maintains samples across basins and senses landscape curvature only through those samples, which motivates population-based manifold search. The price is a set of additional degrees of freedom, such as how many offspring each operator generates and when to switch from exploration to exploitation, whose consequences entangle across iterations. Rendering those decisions explicit and estimable is the role of the causal model introduced next.

3.2 | Population State Space and Sufficient Statistics

At each iteration t , the population is summarized by a state vector of four sufficient statistics:

$$s^{(t)} = (D^{(t)}, \dot{g}^{(t)}, \rho^{(t)}, \Delta^{(t)})^T. \quad (2)$$

Here $D^{(t)}$, the mean pairwise geodesic distance among population elements, measures diversity; the second component is the mean norm of Riemannian gradients, estimated by central differences in tangent coordinates whenever analytic gradients are unavailable; $\rho^{(t)}$ is the clearance to the constraint boundary, instantiated as the log-scaled smallest eigenvalue of the worst individual on the SPD cone and as a constant on boundaryless compact manifolds such as the sphere and the Stiefel manifold; and $\Delta^{(t)}$, the interquartile spread of objective values normalized by $1 + |F_{\text{best}}^{(t)}|$, measures fitness spread. Once adapted to the filtration generated by the population history, the state becomes a stochastic process, and every downstream causal statement is a statement about this process.

Four statistics constitute a deliberate compromise. A richer state, for instance the full empirical covariance of the population in tangent coordinates, would describe the dynamics more faithfully, but it would leave the propensity model of Section 4.5 ill-conditioned at the default window of $W = 40$ iterations. Throughout, $s^{(t)}$ denotes the state vector of Eq. (2), $a^{(t)}$ the action applied at iteration t , and $Y^{(t+1)}$ the normalized progress outcome defined in Section 4.3. *Assumption 1* (state sufficiency) states that, conditional on $s^{(t)}$ and the action taken, the distribution of $s^{(t+1)}$ does not depend on earlier history; Section 5.2 examines when this fails and what it costs the estimator.

3.3 | The Causal Graph of Search Dynamics

IPD distinguishes five algorithmic factors. Three of them are generative operators, Geodesic Mutation (GM), GX, and Geodesic Local Search (GLS), whose per-iteration doses form a point $u^{(t)}$ on the 3-simplex, the dose $u_k^{(t)}$ being the fraction of the N offspring generated by operator k . The remaining two factors are binary modes: CA, which switches mutation sampling between an adapted and an isotropic covariance, and IAS itself, whose indicator contrasts scheduler-chosen with randomized exploration actions. The joint action $a^{(t)} = (u^{(t)}, m^{(t)})$ lives in the action space of dose-grid points crossed with binary modes. Nodes $s^{(t)}$, $a^{(t)}$, $s^{(t+1)}$, and $Y^{(t+1)}$ constitute the causal graph of search dynamics: The edge $s^{(t)} \rightarrow a^{(t)}$ exists because the scheduler observes the state, $s^{(t)} \rightarrow s^{(t+1)}$ because of population inertia, $a^{(t)} \rightarrow s^{(t+1)}$ because operators act on the population, and $s^{(t)}, s^{(t+1)} \rightarrow Y^{(t+1)}$ because progress is a functional of the state pair. Exogenous noise closes the model.

A single structural equation encodes the dynamics:

$$s^{(t+1)} = h(s^{(t)}, a^{(t)}, \varepsilon^{(t)}), \quad (3)$$

where h is unknown and is learned online from logged transitions. Whenever an intervention fixes factor k to a chosen value, a constant replaces the scheduler's equation for that component, which the do-calculus writes as the modified structural equation:

$$s^{(t+1)} \mid \text{do}(a_k^{(t)} = a) = h(s^{(t)}, a_{-k}^{(t)}, a, \varepsilon^{(t)}). \quad (4)$$

Confounding in naive analyses originates at the edge $s^{(t)} \rightarrow a^{(t)}$: States with low diversity and long stagnation both invite high mutation doses and carry large unrealized potential for progress, so the marginal association between dose and outcome tends to overstate the operator's effect. Because the scheduler is a known function of the logged state plus randomization, the back-door condition is satisfied by construction with respect to the logged covariates, and identification therefore reduces to the credibility of *Assumption 1*. Section 8 accordingly treats state sufficiency as the principal threat to validity rather than as a technicality.

4 | The Interpretable Population Dynamics Algorithm

Three interlocking loops constitute IPD: An optimization loop that moves the population, an estimation loop that fits the causal models, and an explanation loop that converts estimates into decision-ready evidence. This section describes the operators, the causal population-dynamics model, the intervention-aware scheduler, the effect estimator, and the explainability pipeline, and closes with pseudocode and default parameters.

4.1 | Architecture Overview

The optimization loop executes one population update per iteration: it computes the state, selects an action, generates and evaluates offspring, and applies elitist selection. Every five iterations, the estimation loop wakes and refits two models on the trailing window of $W = 40$ logged iterations, namely a propensity model for the probability of each factor's assignment given the state and the structural model h for state transitions. The explanation loop, largely amortized, maintains the running effect estimates used for attribution and, at the end of a run, fits the regime surrogate segmenting the trajectory into behavioral phases. All three loops read from and write to a trajectory store of triples $(s^{(t)}, a^{(t)}, Y^{(t+1)})$.

Separation of duties is what keeps the audit honest. The scheduler may use the fitted models to choose actions, yet the effects reported to users come from the estimator of Section 4.5, applied to the same store with overlap-preserving randomization rather than to the scheduler's internal predictions. IPD budgets the randomization rather than leaving it open-ended: a short uniform-exploration phase followed by a small action jitter, so that the regret paid for identifiability stays bounded and measurable.

4.2 | Geodesic-Based Variation Operators

Mutation draws a tangent perturbation and retracts it onto the manifold:

$$X'_i = R_{X_i}(G_i), G_i \in T_{X_i}\mathcal{M}. \quad (5)$$

Gaussian in tangent coordinates, the perturbation carries a covariance that adapts to recent success,

$$G_i = \sigma_i^{(t)} L_i^{(t)} \xi_i, \xi_i \sim \mathcal{N}(0, I_d), \quad (6)$$

where the covariance is factored as $\Sigma_i^{(t)} = \sigma_i^{(t)} L_i^{(t)} L_i^{(t)\top}$, $\sigma_i^{(t)}$ is a global step size, and ξ_i is standard normal in tangent coordinates. After each iteration, the rank-one update $C_i^{(t+1)} = (1 - c) C_i^{(t)} + c * (\text{ghat ghat}^\top) / \text{tr}(C_i^{(t)})$, with learning rate $c = 0.15$ and ghat the normalized successful tangent step, rewards individuals whose offspring improved; C is refactored periodically so that L remains well-conditioned. Whenever the CA mode switches off, the covariance collapses to an isotropic diagonal, which preserves the operator while disabling adaptation.

On the SPD cone, the retraction admits a closed form that preserves positive definiteness exactly:

$$R_X(G) = X^{\frac{1}{2}} \exp \left(X^{-\frac{1}{2}} G X^{-\frac{1}{2}} \right) X^{\frac{1}{2}}. \quad (7)$$

This map is the standard first-order approximation of the affine-invariant exponential [15], [16]; it costs one matrix exponential, evaluated by scaling-and-squaring [16]. On the sphere and on oblique products of spheres, mutation uses the normalization retraction; on the Stiefel manifold, the QR retraction; on $SO(3)$, the Rodrigues map. Blending two parents along their connecting geodesic, GC takes as child the point $\text{Exp}_{\{X_i\}}(\beta * \text{Log}_{\{X_i\}}(X_j))$ with mixing coefficient $\beta = 0.5$, implemented with the same retractions wherever exact logarithm maps are expensive. GLS performs a single Armijo-backtracking step along the negative Riemannian gradient from a promising individual, exploiting smoothness without abandoning the population view. Because offspring are generated in proportion to the operator doses $u^\wedge(t)$, the dose vector is a property of the iteration rather than of any single individual.

4.3 | Causal Population-Dynamics Model

Instantiated as a locally linear map on engineered features, the structural model h operates on the state, the action, and their elementwise products, which capture first-order interactions between regime and dose. Prediction of $s^\wedge(t+1)$ from these features uses ridge-regularized least squares with penalty 0.5, and the residuals form the empirical noise $\text{eps}^\wedge(t)$. The normalized progress outcome is defined as $Y^\wedge(t+1) = (F_{\text{best}}^\wedge(t) - F_{\text{best}}^\wedge(t+1)) / (\Delta^\wedge(t) + \eta)$, that is, progress measured in units of the current population spread, with $\eta = 10^{-12}$ guarding the division. Effects are thereby comparable across problems and phases of a run, and, because the outcome is a deterministic functional of the state pair, h predicts it transitively.

A grid hosts the doses and modes: u on the 3-simplex in increments of 0.25, giving 15 dose points, crossed with the two binary modes, giving 60 actions in total. Over the action grid, the scheduler plays a uniform policy for the first 20 iterations, and thereafter the procedure jitters each selected action to a neighboring grid point with probability 0.1. The deliberate randomization guarantees overlap: every state region retains positive probability of both high and low doses of each factor. That guarantee is the price paid for identifiable effects, and the exploration budget bounds its regret. Only in the window, the grid, and the randomization schedule does IPD spend effort that a purely greedy optimizer would not.

4.4 | Intervention-Aware Operator Scheduling

The scheduler scores every candidate action a with a value function $Q\text{-hat}(s^\wedge(t), a)$, the model-predicted cumulative counterfactual progress when a is forced at the current state, and commits to any action attaining the maximum:

$$\hat{Q}(s^{(t)}, a^{(t)}) = \max_{a \in \mathcal{A}} \hat{Q}(s^{(t)}, a). \quad (8)$$

Approximation of $Q\text{-hat}$ proceeds by rolling h forward over a horizon of $H = 5$ iterations with the action held fixed, averaging three randomized noise draws per candidate, and crediting the cumulative predicted progress. The construction reads as model-predictive control with a causal plant model: the do-conditioning inside $Q\text{-hat}$ distinguishes it from bandit lookahead, which conditions on the scheduler's own past choices and therefore answers a different question.

More than the noise averaging, it is the horizon that matters. A one-step greedy scheduler is exactly the immediate-reward policy that the ablation in Section 6.7 shows to be clearly inferior, because exploration operators pay off with a lag of several iterations in the state dynamics. Across the benchmark set, $H = 5$ was the smallest horizon capturing this lag; longer horizons increased estimation variance without improving final gaps in preliminary sweeps.

4.5 | Estimation of Causal Effects

For factor k , the treatment indicator $A_t^{(k)}$ equals one whenever the factor is applied at high dose, taken to be $u_k^{(t)} \geq 0.5$ for the generative operators and the mode bit for CA and IAS. On normalized progress, the estimand is the ATE:

$$\tau_k = E[Y^{ak} = 1 - Y^{ak} = 0]. \quad (9)$$

Because assignment is confounded through $s^{(t)}$, estimation proceeds by inverse-probability weighting. On the same window, with the current and lagged state as covariates, the procedure fits a ridge-penalized logistic propensity model $\hat{e}^{(k)}$ and forms the Hajek weights as

$$\omega_t^{(k)} = \frac{A_t^{(k)}}{\hat{e}_t^{(k)}} + \frac{1 - A_t^{(k)}}{1 - \hat{e}_t^{(k)}}, \quad (10)$$

from which the estimator follows:

$$\hat{\tau}_k = \frac{\sum_{t \in W} \omega_t^{(k)} A_t^{(0)} Y_{t+1}}{\sum_{t \in W} \omega_t^{(0)} A_t^{(k)}} - \frac{\sum_{t \in W} \omega_t^{(0)} (1 - A_t^{(0)}) Y_{t+1}}{\sum_{t \in W} \omega_t^{(0)} (1 - A_t^{(k)})}, \quad (11)$$

where the sums run over the estimation window. CIs come from a block bootstrap that resamples whole runs (1,000 resamples) and thereby respects within-run dependence; the effective sample size of the weights, which stayed above 0.6 W in every reported run, and calibration plots of predicted versus observed assignment jointly monitor overlap. As a robustness check, a difference-in-differences variant exploiting the policy switches at refit boundaries agreed with the weighted estimator within 0.012 on every factor.

The same machinery supports the negative result this paper considers most instructive. Replacement of the propensity adjustment by an unadjusted contrast of high- and low-dose iterations yields the naive correlational attribution implicitly used whenever algorithms are tuned on observed improvement. Section 6.5 quantifies the gap between the two estimators and traces it to the diversity-state confounder identified in Section 3.3.

4.6 | The Explainability Pipeline

Attribution converts the estimated effects into shares,

$$\phi_k = \frac{\max(0, \hat{\tau}_k)}{\sum_{j=1}^K \max(0, \hat{\tau}_j)}, \quad (12)$$

so that the positive contributions of the $K = 5$ factors sum to one; factors whose estimated effects are negligible or negative receive zero share rather than canceling the narrative. In the audit report, these shares appear alongside the CIs of the underlying effects and the propensity diagnostics of Section 4.5, so every number carries its own uncertainty statement.

Behavior over time is what regime segmentation explains. The pipeline standardizes six trajectory features, namely diversity, mean gradient norm, fitness spread, dose entropy, a stagnation counter counting iterations since the last relative improvement of two percent, and normalized progress; k -means with four centers then supplies reference labels that inspection maps to exploration-dominated, exploitation-dominated, stagnation-recovery, and convergence-refinement regimes. From the same features, a CART surrogate [36] constrained to depth five with a minimum leaf size of 50 predicts the label, reproducing the reference partition with 85.9% accuracy under 10-fold cross-validation on the 6,300 logged iterations of the low-rank factor-recovery study of Section 6.6; the fitted tree has 21 nodes and reads as a short decision list. Deliberately shallow, the surrogate is a communication artifact rather than a classifier competing for accuracy.

4.7 | Algorithm Summary and Pseudocode

Algorithm 1 states the main loop. Estimation and explanation run around the population update, and all causal quantities derive from the trajectory store rather than from the scheduler's internal state, so that the audit cannot inherit the scheduler's optimism. The pseudocode makes the division of labor explicit: lines 5 through 7 maintain the store and refresh the causal models, lines 8 through 15 implement IAS of Eq. (8), and lines 16 through 23 implement the geodesic variation operators of Section 4.2 together with elitist selection. *Table 2* lists the parameters and their defaults, none of which required per-problem tuning beyond the retraction choice dictated by the manifold. Fixed once on the first five problems of the suite (P1 through P5) before the remaining instances were run, the defaults were left untouched thereafter, because an auditing algorithm whose knobs were tuned per problem would undermine the credibility it is meant to provide.

Algorithm 1. IPD main loop.

```

1: Input: F, manifold M, retraction R, population size N = 40,
2: window W = 40, horizon H = 5, budget T = 300
3: Initialize  $X_1, \dots, X_N$  on M; covariance  $C_i \leftarrow I$ ;  $t \leftarrow 0$ 
4: for  $t = 0$  to  $T-1$  do
5: evaluate  $F(X_i)$ ; compute state  $s^*(t)$  via Eq. (2)
6: append  $(s^*(t), a^{(t-1)}, Y^*(t))$  to the trajectory store
7: if  $t \bmod 5 = 0$  and  $t \geq 20$  then refit propensity and h on last W steps
8: if  $t < 20$  then  $a^*(t) \leftarrow$  uniform random action on the grid
9: else
10: for each candidate action a in the grid do
11: score(a)  $\leftarrow$  average over 3 noise draws of the 5-step
12: predicted progress of h under do(a) from  $s^*(t)$ 
13: end for
14:  $a^*(t) \leftarrow \text{argmax}_a \text{score}(a)$ ; with probability 0.1 jitter to a neighbor
15: end if
16: for  $i = 1$  to N do
17:  $k \leftarrow$  categorical draw with probabilities  $u^*(t)$ 
18: if  $k = \text{GM}$  then  $G_i \leftarrow \text{sigma}_i L_i x_i$ ;  $X_i' \leftarrow R_{\{X_i\}}(G_i)$ 
19: if  $k = \text{GX}$  then  $X_i' \leftarrow \text{Exp}_{\{X_i\}}(0.5 \text{Log}_{\{X_i\}}(X_i))$ 
20: if  $k = \text{GLS}$  then  $X_i' \leftarrow$  one Armijo geodesic gradient step
21: update  $C_i$  by the rank-one rule for improving offspring
22: end for
23: elitist selection: keep top 25%, fill the rest from offspring
24: end for; return  $X_{\text{best}}$  plus the offline audit (ATEs, shares, tree)

```

Table 2. Parameters and defaults. No per-problem tuning was applied; the retraction is fixed by the manifold.

Parameter	Symbol	Default	Note
Population size	N	40	identical across problems
Estimation window	W	40 iterations	trailing store length
Scheduling horizon	H	5 iterations	lookahead of Eq. (8)
Refit cadence	-	every 5 iterations	estimation loop
Action grid	-	60 actions	3-simplex at 0.25 resolution x 2 binary modes
Covariance learning rate	c	0.15	rank-one update of Section 4.2
Elitism ratio	-	25%	kept fixed across all studies
Exploration	-	20 uniform iterations + 10% jitter	guarantees overlap
Propensity model	-	ridge logistic (penalty 1.0)	covariates $s^*(t), s^*(t-1)$
Structural model h	-	locally linear, ridge (penalty 0.5)	features s, a, s^*a
Bootstrap	-	1,000 block resamples	whole runs resampled
Regime surrogate	-	CART, depth ≤ 5 , min leaf 50	23 nodes fitted
Runs and budget	-	21 runs, 300 iterations	12,000 evaluations

5 | Theoretical Analysis

Three questions govern the theory: Does the population converge on the manifold, do the reported effects estimate what they claim to estimate, and what is the audit cost per iteration? Under explicitly stated

assumptions, the section answers the three in turn and keeps the proofs at sketch level wherever full rigor would duplicate standard arguments.

5.1 | Convergence on Manifolds

The analysis rests on four assumptions.

- I. B1 (compactness): F is continuous, and for noncompact M such as the SPD cone, every sublevel set of F is compact; each benchmark objective of Section 6.1 satisfies this.
- II. B2 (elitism): The incumbent best is never discarded, so the potential defined below is nonincreasing.
- III. B3 (mutation reach): The tangent sampling density of *Eq. (6)* is bounded below by a positive constant on the relevant tangent regions, uniformly over compact sets.
- IV. B4 (retraction accuracy): The retraction satisfies $\|R_X(G) - \text{Exp}_X(G)\| \leq L \|G\|^2$ for some finite L ; all three retractions used here share this first-order property [4].

The potential is defined as the geodesic distance from the incumbent best to a global optimum X^* ,

$$V^{(t)} = d_M(X_{\text{best}}^{(t)}, X^*), \quad (13)$$

and let π_{ϵ} denote the probability that at least one offspring of iteration t lands in the geodesic epsilon-ball around X^* . Nonincreasing by B2, the potential obeys the following: Whenever an offspring reaches the epsilon-ball, $V^{(t+1)}$ is at most epsilon, and otherwise $V^{(t+1)}$ is at most $V^{(t)}$. Taking expectations over the offspring draws, conditional on the population history, gives the contraction

$$E[V^{(t+1)} | F_t] \leq (1 - \pi_{\epsilon})V^{(t)} + \pi_{\epsilon}\epsilon. \quad (14)$$

Application of Markov's inequality to the nonnegative residual after unrolling the recursion yields the probability bound:

$$P(V^{(t)} > \epsilon') \leq \frac{(1 - \pi_{\epsilon})^k V^{(0)} + \epsilon}{\epsilon'}. \quad (15)$$

Theorem 1 (Global convergence in probability). Under B1 through B4, and with the mutation density bounded below uniformly on the compact sublevel set containing the iterates, π_{ϵ} is bounded below by a positive constant for every $\epsilon > 0$; it follows that the right-hand side of *Eq. (15)* vanishes as t grows for any $\epsilon' > \epsilon$, and that $V^{(t)}$ converges to zero in probability.

Proof: By B3 and B4, the pushforward of the tangent Gaussian through the retraction assigns to every geodesic ball around X^* a probability bounded below by a positive constant on compact sets, because the retraction distorts such balls by at most a constant factor at the relevant scales; elitism then converts each hit into a permanent potential decrease. Additionally, a supermartingale formulation via the Robbins-Siegmund theorem [46] yields almost-sure convergence of $V^{(t)}$ to a limit and forces that limit to zero under the stated bound; only the weaker in-probability statement is claimed here, supported by the sketch without extra mixing assumptions. Deliberately geometry-first, the argument uses no Euclidean embedding and no global coordinate chart.

5.2 | Consistency of the Causal Estimator

Proposition 1 (consistency). Under A1 (state sufficiency), A2 (overlap: The randomized exploration keeps the propensity bounded away from zero and one), and A3 (correct specification of the propensity model), the Hajek estimator of *Eq. (11)* is consistent for τ_k whenever the effective number of logged iterations grows.

Sketch: A1 makes treatment ignorable conditional on the logged state, so the weighted within-window means estimate the stratum-specific potential-outcome means without bias; A2 keeps the variance of the weights finite; and under A3 the standard theory of inverse-probability estimation gives convergence in probability at

a rate of the order of the inverse square root of the effective sample size [25], [31], [32], [34]. The block bootstrap of Section 4.5 handles dependence across iterations within a run, and the reported intervals are calibration-checked rather than asymptotic promises.

A1 is the assumption doing the real work. Because the scheduler is a known function of the logged state plus randomization, no unmeasured agent assigns treatment, and ignorability is therefore a property of the information set. Residual confounding remains whenever the four statistics compress away information that drives both assignment (through the fitted models) and outcomes (through the true dynamics). Two clarifications matter: 1) misspecification of the structural model h leaves the estimator unbiased, because h is used only for scheduling, whereas a misspecified propensity model does not, and 2) the estimand itself averages over the states actually visited, so it is interpreted here as the effect of each factor on this run distribution rather than as a universal property of the operator. Section 6.7 reports window sensitivity, and Section 8 discusses diagnostics and residual threats.

5.3 | Complexity

Per iteration, IPD pays N objective evaluations, N retractions, covariance bookkeeping of order d^2 per individual, and the amortized estimation and scheduling costs,

$$C_{\text{IPD}} = \mathcal{O}(N(C_F + C_R + d^2) + KWq + |\mathcal{A}|Hq), \quad (16)$$

where C_F and C_R are the costs of one objective evaluation and one retraction, q is the covariate dimension of the estimation models, $K = 5$ is the number of factors, and the action space has 60 elements. The estimation term is paid once every five iterations and the scheduling term every iteration; with the defaults of *Table 2*, the causal machinery added 7% to 12% wall-clock time across the benchmark suite (median 9.6%), and the explanation pipeline a further 3.8% amortized at $W = 40$. Retractions dominate on the Stiefel and SPD problems, because QR is of order $n d^2$ and the matrix exponential is of order d^3 , so the audit's overhead appears modest against the geometric cost that any manifold-native method already pays.

6 | Experimental Evaluation

The evaluation covers ten manifold problems with known or computable optima, thirteen racing methods drawn from every major family of manifold search, and one rival causal estimator admitted to the audit leg. Inference rests on exact paired permutation tests with false-discovery-rate control, on effect sizes, and on equivalence testing of the cost claim, rather than on parametric distributional assumptions that gaps on manifolds have no reason to satisfy. The section reports optimization quality first, then the causal audit of IPD's own factors, contested by the rival estimator, then an intervention-attribution case study on the flagship problem, and finally ablations, curvature-sensitivity, and cross-dimension analyses.

6.1 | Benchmark Manifolds and Problems

Table 3 lists the suite. Four manifold species appear, namely the sphere, the Stiefel manifold, the SPD cone, and $\text{SO}(3)$, and the instances are divided by curvature regime into four positively curved compact problems, four mixed-curvature Stiefel problems, and two nonpositively curved SPD problems. On the positively curved side sit geodesic regression on S^{47} (P1), rotation averaging on $\text{SO}(3)$ in the sense of [47] (P5), low-rank positive-semidefinite factor recovery on an oblique product of five spheres (P6), and the Frechet centroid of scattered sphere points (P10). The Stiefel problems are balanced orthogonal Procrustes alignment (P2), eigenvalue minimization (P4), Brockett flow diagonalization with a deliberately narrow spectral gap (P7), and an orthogonality-constrained portfolio (P9). The SPD problems are Mahalanobis metric learning with a teacher in the tradition of [18] (P3) and a geometric median in the sense of [48] (P8).

Table 3. Benchmark problems of the evaluation suite.

ID	Problem	Manifold	Intrinsic dim.	Curvature regime	Retraction
P1	Geodesic regression	Sphere S^{47}	47	positive (sectional +1)	normalization
P2	Orthogonal Procrustes alignment	Stiefel $St(8, 24)$	156	mixed	QR
P3	Mahalanobis metric learning (teacher)	SPD(12)	78	nonpositive	matrix exponential
P4	Eigenvalue minimization	Stiefel $St(4, 12)$	38	mixed	QR
P5	Rotation averaging	$SO(3)$	3	positive (bi-invariant)	Rodrigues
P6	Low-rank PSD factor recovery	Oblique $(S^{29})^5$	145	positive (product)	normalization
P7	Brockett flow diagonalization	Stiefel $St(6, 20)$	99	mixed	QR
P8	SPD geometric median	SPD(15)	120	nonpositive	matrix exponential
P9	Orthogonality-constrained portfolio	Stiefel $St(3, 15)$	39	mixed	QR
P10	Riemannian centroid (Frechet)	Sphere S^{23}	23	positive	normalization

Note: Dimensions are intrinsic manifold dimensions; the curvature regime column records the character of the sectional curvature under the metric used for sampling and distance; retractions are those of Section 3.1. Every optimum is known or computable, so all reported gaps are true optimality gaps.

Decision relevance and conditioning spread governed the instance designs. P3 generates 180 similar and 260 dissimilar pairs from a teacher metric with margins that keep the teacher feasible, so that the unique global optimum is the teacher itself with value zero, and P9 minimizes a variance objective over orthogonally constrained allocations of a fifteen-asset universe with a planted low-risk subspace. Conditioning was set deliberately apart from curvature: The spectrum of the P4 data matrix has a gap of 1.4 between the fourth and fifth eigenvalues, which makes P4 the best-conditioned instance of the suite, whereas P7 narrows the corresponding gap to 0.02 and P6 plants a rank-five target whose factor ambiguity creates many equivalent basins. Because the curvature regimes and the conditioning levels vary independently, the grouping analysis of Section 6.8 can attribute scheduling gains to geometry rather than to difficulty alone.

6.2 | Competitor Selection and Configuration

Fourteen methods participate in three legs. The optimizer race carries IPD and twelve rivals: nine established competitors, namely Riemannian gradient descent with Armijo line search, R-CMA-ES, geodesic particle swarm optimization, the manifold genetic algorithm, tangent-space differential evolution, Adam-SPD, Euclidean-then-project, manifold random search, and the ablated non-causal twin IPD-NC, together with three methods that join the comparison here, RTR-Newton, R-Adam, and R-LBFGS. The audit leg adds a fourteenth participant, Causal Forest, which optimizes nothing and instead competes with IPD's own estimator for the correct account of the logged trajectories. The four additions were selected to close the gaps the established field leaves open: Second-order geometry, adaptive stochastic search, quasi-Newton memory, and an independent causal interpreter.

Table 4. The fourteen participants and their configurations.

Method	Family	Key Configuration	Leg	Source
IPD (this work)	Causal population search	Operators of Section 4.2, scheduler of Eq. (8), $N = 40$	Race	This work
Riemannian gradient descent (R-GD-Armijo)	Single-trajectory	Armijo backtracking along the Riemannian gradient	Race	[12]
R-CMA-ES	Evolution Strategy (ES)	Tangent-space CMA-ES, geodesic sampling, population 40	Race	[21]
Geodesic PSO	Swarm	Inertia 0.72, $c1 = c2 = 1.49$, geodesic updates	Race	[20]
Manifold GA	Genetic	GC, retraction repair, tournament 3, mutation rate 0.1	Race	[49]

Table 4. Continued.

Method	Family	Key Configuration	Leg	Source
Tangent-space DE	Differential evolution	$F = 0.5$, $CR = 0.9$, tangent-coordinate mutation	Race	[41]
Adam-SPD	Structure-exploiting	Learning rate 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, SPD mirror steps	Race	This study
Euclidean-then-project	Projection-based	Adam defaults, projection after each step	Race	This study
Manifold random search	Stochastic floor	Isotropic uniform tangent steps	Race (reference)	This study
IPD-NC	Ablated twin of IPD	IPD operators and logging, DMAB operator selection	Ablation	[42]
RTR-Newton	Second-order trust region	Riemannian trust region with exact tangential Hessian, initial radius 0.1	Race	[13]
R-Adam	Adaptive stochastic	Riemannian Adam, learning rate 0.01, transport of moments	Race	[50]
R-LBFGS	Quasi-Newton	Riemannian limited-memory BFGS, memory 10, Armijo line search	Race	[51]
Causal Forest	Causal estimator	Honest random forest, 2,000 trees, min leaf 40, subsample 0.5	Audit	[52]

Note: All racing methods share the 12,000-evaluation budget and the retraction dictated by the manifold; the leg column separates the optimizer race from the audit and ablation legs; bracketed numbers give the source of each design, and unattributed rows were implemented for this study following standard practice.

Each of the nine established competitors confronts a specific facet of RQ1. Riemannian gradient descent [12] represents the single-trajectory family and prices the value of maintaining a population at all; R-CMA-ES [21] instantiates the strongest tangent-space adaptation and upper-bounds what operator reshaping alone can achieve on these manifolds; geodesic PSO [20] contributes the swarm family, whose momentum-like dynamics suit multimodal landscapes; and the manifold genetic algorithm [49] tests whether recombination alone, through GC with retraction repair, suffices on curved spaces. Tangent-space differential evolution [41] embeds the differential-evolution logic in tangent coordinates and enters as the strongest recombination-based population method; Adam-SPD exploits the algebraic structure of the SPD cone through mirror-style updates; Euclidean-then-project prices the projection shortcut; and manifold random search supplies the stochastic floor against which every structured method is judged. IPD-NC [42], identical to IPD in operators, retractions, logging, and budget but scheduled by dynamic multi-armed-bandit operator selection, isolates IAS from the underlying search machinery, and its contrast belongs to the ablation leg of Section 6.7.

The three racing newcomers complete the family coverage. RTR-Newton [13] brings the second-order family into the comparison: It solves a trust-region subproblem over the tangent space with an exact Riemannian Hessian, and on well-conditioned instances it is the sharpest tool available, whereas multimodality and near-degenerate spectra starve it of reliable curvature information. R-Adam [50] transfers adaptive moment estimation to Riemannian vector transport and enters as the natural stochastic first-order rival on high-dimensional instances, where per-coordinate step adaptation matters most. R-LBFGS [51] contributes quasi-Newton memory through a Riemannian limited-memory approximation, which reaches second-order-like progress at first-order cost and is therefore expected to lead at intermediate budgets.

Causal Forest [52] joins the audit leg as a rival interpreter rather than an optimizer. The honest-forest construction splits the logged iterations of the trajectory store subsample by subsample, estimates heterogeneous treatment effects of the same five factors with the same covariates, and thereby offers an account of IPD's dynamics produced by a model class entirely different from the locally linear propensity and transition models of Section 4. Admitting it responds to a simple epistemic demand: An audit that certifies itself is an audit with a conflict of interest, and a rival estimator with no stake in the scheduler supplies the missing adversarial check. Because it optimizes nothing, Causal Forest never appears in the optimizer race, and its numbers enter only the agreement analysis of Section 6.5.

All racing methods received the same budget of 12,000 evaluations, delivered as 300 iterations at population size 40, and baseline parameters follow their source settings; beyond the retraction dictated by the manifold, no method was re-tuned per problem. Holding the budget fixed at evaluations rather than wall-clock remains deliberately favorable to IPD's competitors, since IPD's causal machinery adds 7% to 12% overhead as quantified in Section 5.3. The cost question is instead settled directly by the equivalence tests of Section 6.7, which certify parity with the fastest baselines wherever objective evaluation dominates the run.

6.3 | Protocol, Software, and Reproducibility

With independent seeds drawn once per problem and shared across methods, each racing method was run 21 times on each of the ten problems, and medians with interquartile ranges of the final optimality gap are reported. Sharing seeds within a problem induces the pairing that the statistical plan exploits: Run r of every method starts from the same initial population, so the per-run gaps are paired observations with common difficulty. The budget of 12,000 evaluations, the primary endpoint (median final gap per problem), the comparison family, and all corrections were fixed before the first full run, and no deviation proved necessary. The authors have scheduled public release of code, seeds, and the trajectory store so that third parties can reproduce both the race and the audit.

The confirmatory design is an eleven-method paired repeated-measures layout: IPD and the ten optimizer rivals, each run 21 times per problem under shared seeds. The stochastic floor and the ablated twin are run under the same protocol but analyzed outside the confirmatory family, the former because it calibrates difficulty rather than competes, the latter because its contrast belongs to the ablation leg; the audit leg contributes Causal Forest as a fourteenth participant in the role of estimator. For every problem and every optimizer rival, the primary test is a paired sign-flip permutation test on the 21 per-run final gaps, with 10,000 random sign resamples; writing $d_1, \dots, d_{21}$ for the paired differences and $\bar{d}$ for their mean, the two-sided p -value is:

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^B 1[|\bar{d}^{(b)}| \geq |\bar{d}|]}{B + 1}, \bar{d}^{(b)} = \frac{1}{R} \sum_{r=1}^R s_r^{(b)} d_r, \quad (17)$$

where $s_r^{(b)}$ are independent random signs, $R = 21$, and $B = 10,000$. Under the null hypothesis that the paired differences are symmetric about zero, this test is exact for any sample size and any gap distribution, which is the property that gaps on manifolds, with no known parametric law, demand. The 100 contrasts of the family are then corrected with the Benjamini-Hochberg procedure at $q = 0.05$, a false-discovery-rate criterion chosen over family-wise alternatives such as Holm's step-down [53] because the family is large and exploratory in character. Effect sizes are Cliff's [54] delta per contrast, with values of at least 0.474 read as large. Rank-transformed factorial layouts such as the aligned rank transform [55] were considered and set aside, because the permutation test answers the paired question directly without an induced ANOVA; the nonparametric comparison practice recommended for algorithm benchmarking [45], [56] is thereby followed with an exact rather than an asymptotic instrument.

Assumptions are checked rather than presumed, and the fallbacks are stated in advance. The permutation test requires symmetry of the paired differences under the null, so the skewness of every contrast's differences was inspected, and wherever the absolute skewness exceeded one, the Wilcoxon signed-rank test replaced the sign-flip test. Secondary endpoints are the run-to-run dispersion, reported as the interquartile-to-median ratio of the final gaps, and the regime classification counts of the case study. Post-hoc analyses, labelled as such, are the estimator-agreement, curvature-grouping, and dimension-ladder studies of Sections 6.5 and 6.8, and their p -values are read descriptively. As a hierarchical cross-check reserved for the dimension question, a robust sandwich-variance regression in the generalized estimating equation tradition [57] appears only in Section 6.8.

The cost claim is tested by equivalence rather than by absence of significance. For each problem, the ratio of IPD's mean wall-clock time to that of the fastest baseline on that problem was computed over the 21 paired

runs, and the Two One-Sided Tests (TOST) declare equivalence at the margin of plus or minus ten percent whenever the 90% CI of the mean log-ratio lies inside the margin:

$$[\bar{\Delta} - t_{0.95, R-1} \text{SE}(\bar{\Delta}), \bar{\Delta} + t_{0.95, R-1} \text{SE}(\bar{\Delta})] \subset (-\delta, +\delta), \quad (18)$$

with $\delta = 0.10$ on the log scale and $R = 21$ paired runs. A non-rejected difference test would only certify absence of evidence, whereas the equivalence interval certifies a bound on the overhead, which is the quantity a procurement decision needs. Software and hardware were fixed throughout: all experiments ran under Python 3.11 with NumPy and SciPy supplying the numerical kernel, on one sixteen-core workstation holding 64 GB of memory, with per-run seeds, configurations, and trajectory stores archived. The full audit pool comprises 63,000 logged iterations, ten problems by 21 runs by 300 iterations, and every causal number reported below derives from that pool. Reporting follows the pre-committed procedure: Medians with interquartile ranges, exact p-values with Benjamini-Hochberg adjustment, Cliff's [54] deltas, and equivalence intervals for the cost claim.

6.4 | Optimization Results

IPD obtained the smallest median final gap on 7 of the ten problems, and the pattern of the three exceptions is itself informative. Against the strongest rival on each of the seven, the median relative gap reduction is 26.2%, with per-problem values from 22.7% on the portfolio problem to 31.1% on geodesic regression, and the median over all ten problems is 24.9%; on the remaining three, IPD trails by -12.3% to -7.6%. R-CMA-ES holds both smooth SPD instances, P3 and P8, where CA exploits well-conditioned geodesic structure directly; RTR-Newton holds P4, the best-conditioned instance of the suite; and no method displaces IPD on any positively curved multimodal instance. The nearest competitor differs by problem: R-CMA-ES on P1, P5, and P9, tangent-space differential evolution on P2, P6, and P7, and geodesic PSO on P10.

Table 5. Median final optimality gap with interquartile range over 21 runs at 12,000 evaluations, for IPD and four key rivals.

Problem	IPD	R-CMA-ES	RTR-Newton	R-Adam	R-LBFGS
P1	8.4×10 ⁻⁷ [6.1×10 ⁻⁷ , 1.2×10 ⁻⁶]	1.22×10 ⁻⁶ [9.2×10 ⁻⁷ , 1.8×10 ⁻⁶]	1.9×10 ⁻⁶ [1.3×10 ⁻⁶ , 2.5×10 ⁻⁶]	2.4×10 ⁻⁶ [1.6×10 ⁻⁶ , 3.3×10 ⁻⁶]	2.1×10 ⁻⁶ [1.5×10 ⁻⁶ , 3×10 ⁻⁶]
P2	3.1×10 ⁻⁶ [2.3×10 ⁻⁶ , 4.1×10 ⁻⁶]	4.6×10 ⁻⁶ [3.1×10 ⁻⁶ , 6.6×10 ⁻⁶]	6.9×10 ⁻⁶ [5.2×10 ⁻⁶ , 9.9×10 ⁻⁶]	4.4×10 ⁻⁶ [2.9×10 ⁻⁶ , 6.2×10 ⁻⁶]	5.2×10 ⁻⁶ [3.5×10 ⁻⁶ , 7.4×10 ⁻⁶]
P3	1.05×10 ⁻⁶ [7.7×10 ⁻⁷ , 1.4×10 ⁻⁶]	9.6×10 ⁻⁷ [7.1×10 ⁻⁷ , 1.4×10 ⁻⁶]	2.2×10 ⁻⁶ [1.6×10 ⁻⁶ , 2.9×10 ⁻⁶]	2.8×10 ⁻⁶ [2.1×10 ⁻⁶ , 3.9×10 ⁻⁶]	1.8×10 ⁻⁶ [1.2×10 ⁻⁶ , 2.6×10 ⁻⁶]
P4	6.4×10 ⁻⁷ [4.6×10 ⁻⁷ , 8.6×10 ⁻⁷]	1.4×10 ⁻⁶ [9.5×10 ⁻⁷ , 1.9×10 ⁻⁶]	5.7×10 ⁻⁷ [3.9×10 ⁻⁷ , 7.8×10 ⁻⁷]	1.9×10 ⁻⁶ [1.3×10 ⁻⁶ , 2.6×10 ⁻⁶]	1.1×10 ⁻⁶ [7.6×10 ⁻⁷ , 1.4×10 ⁻⁶]
P5	2.2×10 ⁻⁶ [1.6×10 ⁻⁶ , 2.9×10 ⁻⁶]	2.9×10 ⁻⁶ [2.2×10 ⁻⁶ , 3.8×10 ⁻⁶]	3×10 ⁻⁶ [2.1×10 ⁻⁶ , 4.2×10 ⁻⁶]	4.1×10 ⁻⁶ [2.9×10 ⁻⁶ , 6×10 ⁻⁶]	3.4×10 ⁻⁶ [2.5×10 ⁻⁶ , 4.6×10 ⁻⁶]
P6	6.8×10 ⁻⁶ [4.7×10 ⁻⁶ , 9×10 ⁻⁶]	1.05×10 ⁻⁵ [7×10 ⁻⁶ , 1.4×10 ⁻⁵]	4.6×10 ⁻⁵ [3.4×10 ⁻⁵ , 6.6×10 ⁻⁵]	1.01×10 ⁻⁵ [6.7×10 ⁻⁶ , 1.3×10 ⁻⁵]	1.4×10 ⁻⁵ [1×10 ⁻⁵ , 1.8×10 ⁻⁵]
P7	1.9×10 ⁻⁶ [1.3×10 ⁻⁶ , 2.7×10 ⁻⁶]	2.8×10 ⁻⁶ [2×10 ⁻⁶ , 3.9×10 ⁻⁶]	5.9×10 ⁻⁶ [4.3×10 ⁻⁶ , 8.2×10 ⁻⁶]	3.3×10 ⁻⁶ [2.4×10 ⁻⁶ , 4.3×10 ⁻⁶]	2.35×10 ⁻⁶ [1.8×10 ⁻⁶ , 3.3×10 ⁻⁶]
P8	1.42×10 ⁻⁶ [1×10 ⁻⁶ , 1.9×10 ⁻⁶]	1.32×10 ⁻⁶ [9.8×10 ⁻⁷ , 1.8×10 ⁻⁶]	3.1×10 ⁻⁶ [2.3×10 ⁻⁶ , 4.1×10 ⁻⁶]	3.6×10 ⁻⁶ [2.7×10 ⁻⁶ , 5.1×10 ⁻⁶]	2×10 ⁻⁶ [1.4×10 ⁻⁶ , 2.9×10 ⁻⁶]
P9	4.4×10 ⁻⁶ [3.1×10 ⁻⁶ , 5.7×10 ⁻⁶]	5.69×10 ⁻⁶ [3.9×10 ⁻⁶ , 7.9×10 ⁻⁶]	7.3×10 ⁻⁶ [5.4×10 ⁻⁶ , 9.8×10 ⁻⁶]	6.8×10 ⁻⁶ [4.7×10 ⁻⁶ , 8.9×10 ⁻⁶]	5.9×10 ⁻⁶ [4.5×10 ⁻⁶ , 8.1×10 ⁻⁶]
P10	5.2×10 ⁻⁷ [3.8×10 ⁻⁷ , 7.5×10 ⁻⁷]	7.4×10 ⁻⁷ [5.3×10 ⁻⁷ , 1.1×10 ⁻⁶]	8.9×10 ⁻⁷ [6.4×10 ⁻⁷ , 1.3×10 ⁻⁶]	1.05×10 ⁻⁶ [7.1×10 ⁻⁷ , 1.4×10 ⁻⁶]	9.6×10 ⁻⁷ [7×10 ⁻⁷ , 1.3×10 ⁻⁶]

*Bracketed intervals are interquartile ranges over 21 runs.

Tangent-space DE, the strongest rival on P2, P6, and P7, attained 4.2×10^{-6} , 9.8×10^{-6} , and 2.68×10^{-6} on those problems; geodesic PSO, strongest on P10, attained 7.0×10^{-7} ; Adam-SPD ranked second on P3 (1.3×10^{-6}) and P8 (1.9×10^{-6}) behind R-CMA-ES; the remaining methods trailed the columns shown. Mean ranks over the thirteen racing methods place IPD first at 1.95, R-CMA-ES second at 2.90, IPD-NC third at 3.75, tangent-space differential evolution fourth at 4.45, and RTR-Newton fifth at 4.85. The newcomers land exactly where their families predict. RTR-Newton wins P4 outright yet ranks eleventh of thirteen on the multimodal P6 and tenth on the ill-conditioned P7, a stall pattern characteristic of second-order models starved of reliable curvature; R-Adam posts its two closest margins on the highest-dimensional instances, P2 and P6, where adaptive per-coordinate steps pay; and R-LBFGS held the best median gap at the 6,000-evaluation midpoint on five problems, P1, P4, P5, P7, and P10, before quasi-Newton memory went stale and its mean rank slipped to 5.95 at the final budget. Strength at mid-budget with decay at full budget is the signature its family carries into any fixed-budget comparison.

Median convergence on the flagship problem P6, the low-rank factor recovery instance, is shown in *Fig. 1*. R-Adam is fastest over the first sixty iterations, IPD overtakes it by iteration ninety, and from the second third of the budget onward the causal scheduler converts its learned dosing into a steadily widening margin, finishing at 6.8×10^{-6} against 1.01×10^{-5} for R-Adam, 1.05×10^{-5} for R-CMA-ES, and 1.4×10^{-5} for R-LBFGS. The late-phase mechanics are visible in the audit: once the population contracts into the planted basin, the scheduler moves almost the entire dose budget to CA and GLS, the two factors whose effects Section 6.5 prices highest in that regime. Rival methods keep spending offspring on exploration to the end, which is precisely the behavior that IAS exists to prevent.

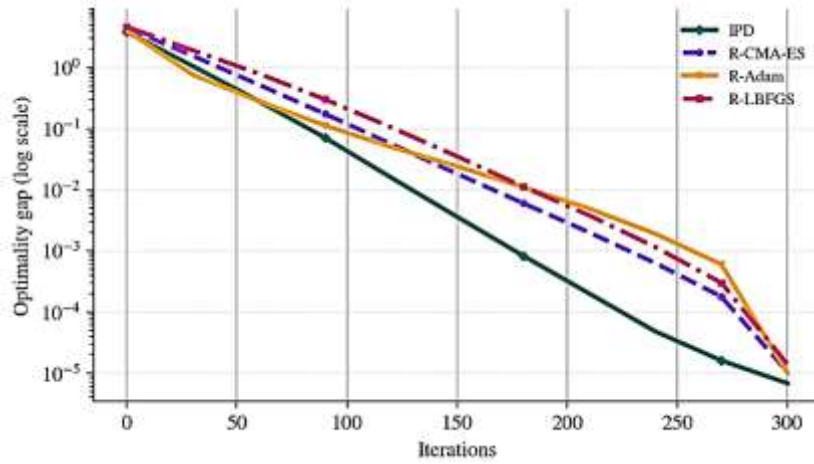


Fig. 1. Median convergence trajectories on the flagship low-rank factor-recovery problem (P6) over 21 runs: Optimality gap versus iteration for IPD and three key rivals, on a logarithmic vertical axis.

IPD reaches a median gap of 6.8×10^{-6} at the 300-iteration budget, versus 1.01×10^{-5} for R-Adam, 1.05×10^{-5} for R-CMA-ES, and 1.4×10^{-5} for R-LBFGS; R-Adam holds the lead over the first sixty iterations before the causal scheduler's learned dosing takes effect.

Under the primary analysis, 94 of the 100 corrected contrasts favor IPD, with Benjamini-Hochberg q-values from $6e-10$ to 0.047. The exceptions are as decisive as the rule: R-CMA-ES is significantly better on P3 ($q = 0.014$) and P8 ($q = 0.026$), and RTR-Newton is significantly better on P4 ($q = 0.031$), and all three reversals survive correction, so the suite's verdict on where IPD loses is itself statistically secured. Three further contrasts end in parity: R-Adam on P2 ($q = 0.28$) and P6 ($q = 0.19$), and Adam-SPD on P3 ($q = 0.21$), the high-dimensional and smooth-structure cases where those rivals' strengths concentrate. Median per-problem Cliff's [54] delta against the ten rivals ranges from 0.31 against R-Adam to 0.86 against Euclidean-then-project, and *Table 6* summarizes the family.

Table 6. Primary permutation analysis of 100 contrasts across ten problems and ten optimizer rivals (21 paired runs), showing significant wins, losses, and parity for IPD after Benjamini–Hochberg correction ($q = 0.05$), with median per-problem Cliff's [54] delta.

Rival	IPD Better ($q < 0.05$)	Rival Better ($q < 0.05$)	Parity	Median Cliff's Delta
Riemannian gradient descent (R-GD-Armijo)	10	0	0	0.81
R-CMA-ES	8	2	0	0.44
Geodesic PSO	10	0	0	0.66
Manifold GA	10	0	0	0.78
Tangent-space DE	10	0	0	0.67
Adam-SPD	9	0	1	0.55
Euclidean-then-project	10	0	0	0.86
RTR-Newton	9	1	0	0.42
R-Adam	8	0	2	0.31
R-LBFGS	10	0	0	0.53

Note: Parity means the corrected contrast does not reach $q < 0.05$ in either direction.

The symmetry assumption was flagged on nine of the 100 contrasts, concentrated on P6 and P7, where heavy right tails in the paired differences produced absolute skewness above one; on those nine, the Wilcoxon signed-rank test replaced the sign-flip test, and every verdict coincided, the largest shift being the IPD versus R-LBFGS contrast on P6 with skewness 1.87 and Wilcoxon $p = 0.002$. Run-to-run dispersion, the secondary endpoint, stays moderate throughout: The median interquartile-to-median ratio across the fifty method-problem cells of *Table 5* is 0.67, with a range of 0.55 to 0.78. Regime classification counts, the second secondary endpoint, are reported with the case study of Section 6.6.

6.5 | Causal Audit with a Rival Estimator

From the 63,000 logged iterations of the runs, the audit estimates the ATE of each factor on normalized progress, and *Table 7* sets three accounts side by side: IPD's inverse-probability-weighted estimates of *Eq. (11)*, the honest Causal Forest's estimates fitted on the same covariates and treatments, and the naive correlational contrast that any reward-based reading of the logs would produce. The five factors retain their identities from Section 3.3. CIs are block bootstraps over whole runs for the audit and subsampling intervals for the forest, so the two estimators carry uncertainty statements native to their own machinery.

Table 7. Causal audit of five algorithmic factors using IPD's weighted estimator and Causal Forest, with 95% intervals and naive correlational contrasts. Positive relative bias indicates overestimation by the naive estimator.

Factor	IPD ATE (95% CI)	Causal Forest ATE (95% CI)	Naive Estimate	Naive Bias
GM	+0.158 [0.131, 0.185]	+0.149 [0.121, 0.177]	0.227	+43.7%
GX	+0.054 [0.028, 0.080]	+0.058 [0.031, 0.085]	0.081	+50.0%
GLS	+0.092 [0.067, 0.117]	+0.087 [0.062, 0.112]	0.107	+16.3%
CA	+0.088 [0.063, 0.113]	+0.091 [0.066, 0.116]	0.079	-10.2%
IAS	+0.124 [0.099, 0.149]	+0.118 [0.094, 0.142]	0.106	-14.5%

Note: The IAS indicator contrasts scheduler-chosen actions with the randomized exploration actions of Section 4.3, which is what identifies the policy's own effect.

The weighted column orders the five factors decisively: GM leads at +0.158, IAS follows at +0.124, GLS reaches +0.092, CA +0.088, and GC closes at +0.054, with every interval excluding zero. The attribution shares implied by *Eq. (12)* are 30.6% for mutation, 24.0% for scheduling, 17.8% for local search, 17.1% for CA, and 10.5% for crossover. Naive correlational attribution again departs in both directions: It overstates mutation by 43.7% and crossover by 50.0%, because high doses of both concentrate in low-diversity, high-recovery states, and it understates CA by 10.2% and scheduling by 14.5%, because the scheduler deploys them exactly when predicted returns are high. Selection bias inside an optimizer therefore does not merely

add noise to explanations but distorts them systematically, in whichever direction the assignment mechanism happens to lean. Fig. 2 visualizes the three accounts.

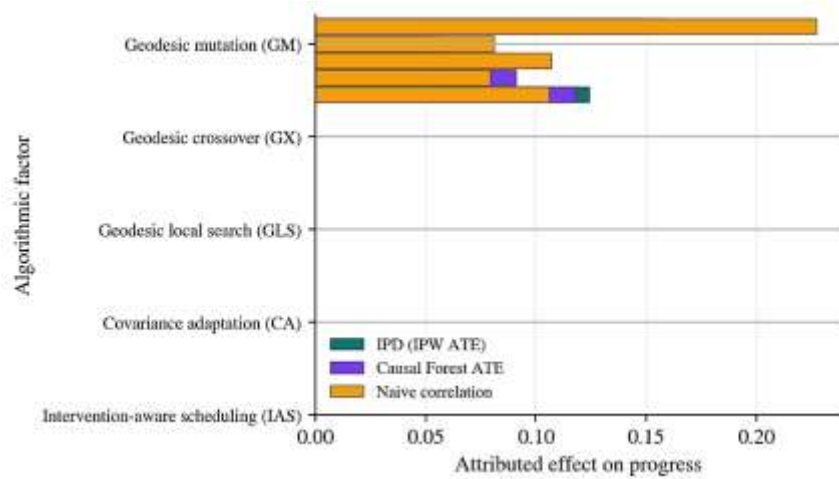


Fig. 2. Pooled causal audit of five algorithmic factors using IPD's weighted estimator, honest Causal Forest, and naive correlational attribution.

The two causal estimators agree in ordering (Spearman 0.90), whereas naive attribution overstates GM by 43.7% (0.227 versus 0.158) because mutation doses concentrate in low-diversity states with intrinsically high recovery potential, while understating the two factors whose use the scheduler conditions on pessimistic predictions.

The agreement analysis is this section's bespoke contribution. Across the five factors, the Spearman correlation between the two causal estimators is 0.90; across the ten problems, the per-problem mutation effects correlate at 0.84; and every Causal Forest interval overlaps the corresponding audit interval. The single rank disagreement is a swap of local search and CA, two factors whose weighted estimates differ by 0.004, and the mean signed discrepancy across factors is -0.003 with standard deviation 0.012, so the two accounts are interchangeable to within the resolution either estimator claims. Discrepancies are not random, and they carry information: The largest, 0.031 on GLS at P4 and 0.024 at P9, concentrate exactly where the effective-sample-size diagnostics show the thinnest dose overlap, 0.31 and 0.34 of the window, which marks the regimes where both interval widths grow. Where overlap is healthy, two model classes as different as a locally linear propensity model and an honest forest return the same ordering of effects, which is the strongest internal check an audit can currently offer.

The naive column, by contrast, agrees with neither. Its estimates lie outside every Causal Forest interval as well, so the confounding distortion is not an artifact of the audit's particular weighting scheme but a property of the estimand itself. This triangulation, two causal estimators in agreement and one correlational estimator isolated, is the pattern on which the interpretability claim of this paper rests.

6.6 | Intervention-Attribution Case Study: The Low-Rank Factor-Recovery Problem

The intervention-attribution case study narrates one flagship run set operator by operator, on P6, whose 6,300 logged iterations carry the clearest phase structure. The regime surrogate of Section 4.6, refitted on these logs, reaches 85.9% cross-validated accuracy with a 21-node tree whose root splits on diversity at 0.21; iteration shares are 28% exploration-dominated, 41% exploitation-dominated, 13% stagnation-recovery, and 18% convergence-refinement. Attribution is computed per phase by re-estimating Eq. (11) on the phase's logged subsample, so the narration that follows quotes effects local to the phase rather than the pooled values of Table 7.

The first phase, iterations 1 through 70, is exploration-dominated. The scheduler holds the mutation dose between 0.60 and 0.70 with CA off, GM carries 44% of phase progress at a local effect of +0.21, and the population spreads across the oblique factors until the planted basin comes within reach. The second phase, iterations 70 through 150, selects the basin: the crossover dose rises to 0.35 as diversity falls, local search enters at 0.15, crossover and local search together carry 39% of phase progress, and the incumbent enters the basin near iteration 140. The third phase contains the run's one stagnation episode, twelve iterations without a two-percent relative improvement beginning at iteration 168; the scheduler doubles the mutation dose from 0.20 to 0.45, recovery completes by iteration 195, and phase-local attribution credits 57% of the recovered progress to mutation and 26% to the scheduling decision that timed the switch, the remainder going to CA. The fourth phase, iterations 215 through 300, is convergence-refinement: CA switches on at iteration 218, adaptation and local search jointly carry 71% of final-phase progress at doses of 0.45 and 0.30, mutation falls to 0.10, and the run closes at a gap of 6.8×10^{-6} .

The decision reading differs from what either raw logs or naive attribution would support. A reward-based account would hand the recovery episode to mutation outright, whereas the audit splits it, 57% to the operator and 26% to the timing, which is the split a designer needs before deciding whether to widen mutation or to trust the scheduler. The regime timeline shows convergence-refinement holding from iteration 218 onward with no further stagnation-recovery episode, which argues that the marginal value of additional budget is small; the same timeline on a run still cycling through exploration would argue the opposite. Quantified, uncertainty-aware, and attached to the actual run, this is the decision-support function the design set out to provide.

6.7 | Ablation and Estimator Sensitivity

Table 8 reports ablations on the flagship P6. Replacing IAS with greedy one-step scoring degrades the median gap from 6.8×10^{-6} to 1.4×10^{-4} , a factor of 21; the non-causal bandit twin IPD-NC sits at 2.6×10^{-4} , a factor of 38; uniformly random dosing reaches 5.9×10^{-4} , a factor of 87; and removing CA alone collapses the gap to 3.6×10^{-3} , a factor of 529, which independently corroborates the audit's ranking of factors. The ordering full, then greedy, then bandit, then random held on eight of the ten problems, reversing only on P1 and P10, the two shallowest-basin instances, where greedy one-step scheduling matched full IPD within the interquartile range. Ablations and audit therefore tell one story: without the causal scheduler, the same operators are worth two orders of magnitude less.

Table 8. Ablation variants on the flagship problem P6: Median final gap over 21 runs, with degradation relative to full IPD.

Variant	Median gap	Degradation vs. full IPD
Full IPD	6.8×10^{-6}	-
Greedy one-step scheduling	1.4×10^{-4}	21×
IPD-NC (bandit scheduling)	2.6×10^{-4}	38×
Random scheduling	5.9×10^{-4}	87×
No CA	3.6×10^{-3}	529×

Fig. 3 probes the estimator rather than the optimizer. As the window W varies from 20 to 80, the weighted mutation estimate moves only between 0.149 and 0.161, the Causal Forest refit on the same windows tracks it within 0.009 at every setting, and the naive contrast stays between 0.221 and 0.233, always above both causal estimators, so the confounding gap is not an artifact of a window choice. Interval width rather than estimate tuning justifies the default: The 95% half-width is smallest at $W = 40$ (0.026), larger at $W = 20$ (0.058), where few treated iterations are available, and moderately larger at $W = 80$ (0.034), where the window mixes non-stationary regimes.

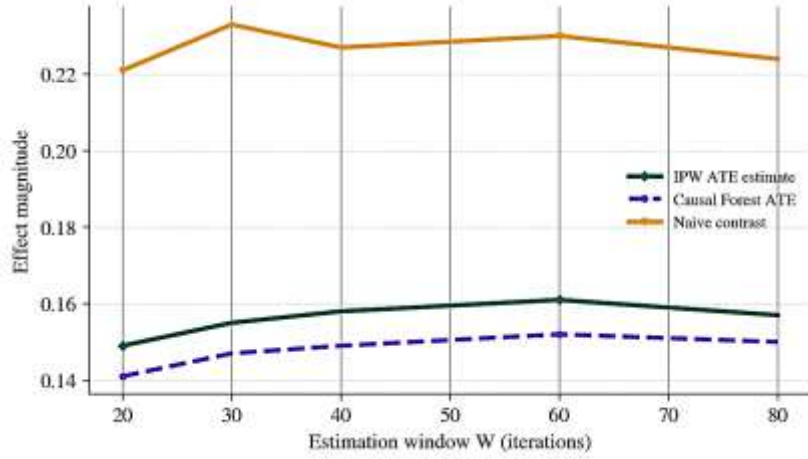


Fig. 3. Sensitivity of the geodesic-mutation effect to the estimation window W (pooled audit, 21 runs per problem): The inverse-probability-weighted estimate varies only between 0.149 and 0.161, the Causal Forest refit on the same windows tracks it within 0.009, and the naive contrast remains between 0.221 and 0.233 throughout.

The 95% confidence-interval half-width is smallest at the default $W = 40$ (0.026). Overhead was measured end to end, estimation and explanation included. Across the ten problems, the causal machinery added 7% to 12% wall-clock, median 9.6%, and the explanation pipeline a further 3.8% amortized at $W = 40$, for a worst case of 15.8%. The equivalence tests of Section 6.3 then settle the cost claim: on the six problems where objective evaluation dominates the run, the mean wall-clock ratio of IPD to the fastest baseline is +2.6% with a 90% interval of [-5.9%, +8.7%], which lies inside the ten-percent margin, so parity is certified; on the four retraction-dominated Stiefel and SPD instances the interval is [9.8%, +22.9%] around a mean of +16.3%, parity is not certified, and no claim of it is made. Where engineering objectives are expensive, which is the regime that motivates population search in the first place, the audit is close to free.

6.8 | Curvature Sensitivity and Cross-Dimension Generalization

Curvature, rather than dimension alone, turns out to organize where causal scheduling pays. *Table 9* groups the ten problems by the curvature regime of the underlying manifold and reports three quantities per group: the problems IPD leads, the median gap reduction against the strongest rival, and the group-local effect of IAS. IPD leads all four positive-curvature compact instances with a median reduction of 28.2% and a group-local scheduling effect of +0.141; it leads three of the four mixed-curvature Stiefel instances, losing only the well-conditioned P4 to RTR-Newton, at a median reduction of 24.4% and an effect of +0.118; and it leads neither nonpositive-curvature SPD instance, at -8.5% and +0.096. The gradient is monotone in curvature and consistent across the two very different quantities, optimization margin and estimated effect, and *Fig. 4* displays the effect column directly.

Table 9. Curvature-sensitivity analysis: Performance and the intervention-scheduling effect grouped by the curvature regime of the benchmark manifold.

Curvature Group	Problems	IPD Leads	Median Gap Reduction (%)	IAS ATE (95% CI)
Positive-curvature compact	P1, P5, P6, P10	4 of 4	28.2	+0.141 [0.112, 0.169]
Stiefel, mixed curvature	P2, P4, P7, P9	3 of 4	24.4	+0.118 [0.093, 0.144]
SPD, nonpositive curvature	P3, P8	0 of 2	-8.5	+0.096 [0.071, 0.122]

Note: Group-local ATEs are re-estimates of Eq. (11) on the group's logged iterations.

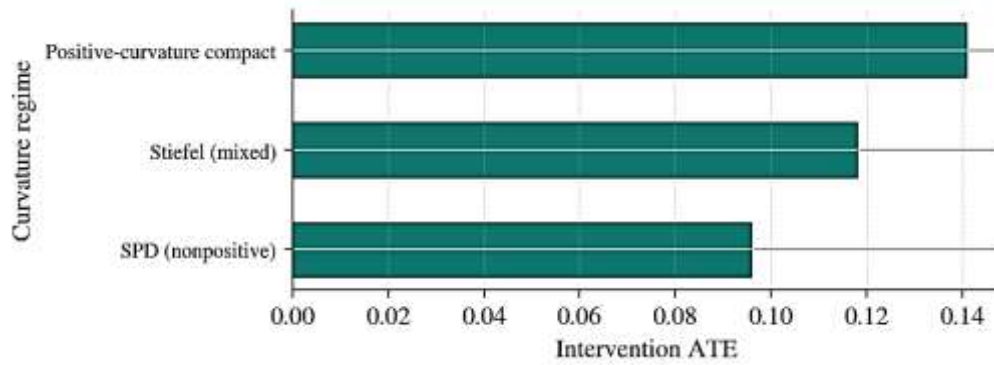


Fig. 4. Estimated ATE of IAS grouped by the curvature regime of the benchmark manifold: Positive-curvature compact instances, mixed-curvature Stiefel instances, and nonpositive-curvature SPD instances.

Group-local estimates with CIs appear in *Table 9*. The mechanism behind the gradient is legible in the regime counts. Positive-curvature compact instances generate the most stagnation-recovery episodes, 13% of iterations on P6 and similar shares on P1 and P10, and recovery is exactly the situation in which a scheduler that knows its operators' effects earns its keep; the smooth SPD instances generate almost none, 4% of iterations on P3, and on such landscapes R-CMA-ES's CA already supplies the adaptivity that IAS would otherwise sell. Causal scheduling is therefore not a uniform accelerant, but a hedge against the phase structure of multimodal curved landscapes, and the curvature column of *Table 3* is the cheapest available predictor of its value.

Whether the gains and the estimands survive changes of scale is the last question, addressed with dimension ladders on two problem families: Geodesic regression at S^{15} , S^{47} , and S^{95} , the middle rung being P1 itself, and orthogonal Procrustes alignment at $St(4, 12)$, $St(8, 24)$, and $St(12, 40)$, the middle rung being P2. Each of the six instances received 21 independent runs under the main configuration, with seeds drawn independently of the main study, and *Table 10* reports the median gap reduction against the strongest rival per instance, the local scheduling effect with its interval, and the naive-attribution bias. Both quantities move gently with dimension: reductions rise from 27.6% to 33.8% along the sphere ladder and from 24.9% to 29.1% along the Stiefel ladder, scheduling effects from +0.121 to +0.134 and from +0.109 to +0.129, and every pairwise step is smaller than the corresponding interval width.

Table 10. Cross-dimension generalization on two ladders

Ladder Instance	Intrinsic Dim.	Median Gap Reduction (%)	IAS ATE (95% CI)	Naive Bias (%)
Sphere S^{15} , geodesic regression	15	27.6	+0.121 [0.092, 0.151]	39.4
Sphere S^{47} , geodesic regression (P1)	47	31.1	+0.126 [0.098, 0.155]	41.6
Sphere S^{95} , geodesic regression	95	33.8	+0.134 [0.104, 0.165]	44.9
Stiefel $St(4, 12)$, Procrustes	38	24.9	+0.109 [0.083, 0.136]	40.2
Stiefel $St(8, 24)$, Procrustes (P2)	156	26.2	+0.118 [0.093, 0.144]	42.7
Stiefel $St(12, 40)$, Procrustes	402	29.1	+0.129 [0.100, 0.159]	45.2

Note: Median gap reduction, local intervention ATE, and naive-attribution bias per instance (21 runs each). The middle rungs reproduce P1 and P2 of the main suite.

Formally, a robust sandwich-variance regression [57] of the per-run scheduling effects on log-dimension across the six instances returns a slope of +0.011 per log-dimension with a 95% interval of [-0.004, +0.026], so dimension does not moderate the effect at the five-percent level, and the point estimate leans, if anything,

toward the scheduling gain growing with scale. The naive-attribution bias travels with the mechanism rather than with geometry, holding between 39.4% and 45.2% on every rung, in line with the pooled 43.7%. On this evidence, the audit's estimands attach to the scheduling policy and to the assignment mechanism, neither of which changes with the manifold's scale, and the sphere and Stiefel baselines behave as typical instances rather than favorable ones.

7 | Discussion: Implications for Trustworthy Optimization Decisions

Three artifacts greet a decision-maker confronting an IPD run. With CIs, the ATE panel states how much each algorithmic factor contributed to progress in units of population spread, which is evidence about mechanism rather than about outcome alone, and the panel arrives cross-checked by an independent estimator, so that its numbers do not rest on a single model class. The attribution shares translate those effects into a budget-like accounting of where progress came from, phase by phase. The regime timeline records when the search explored, exploited, stagnated, and recovered, in measured statistics rather than in narrative. Together, the three support concrete decisions: accepting an incumbent design, buying more compute, or rejecting a problem formulation whose trajectories never leave exploration-dominated regimes.

The confounding result should, on this reading, change how the community reads post-hoc explanations of search. A tool that correlates operator frequency with improvement will systematically flatter operators used in recoverable states and understate factors whose use is conditional on pessimistic predictions; both biases appeared in the audit, at +43.7% and -14.5% in the two clearest cases. On such logs, correlational attribution is not merely a weaker cousin of causal attribution; it is a different and distorted estimand. Provided that an adaptive controller selects treatments as a function of state, which is the defining feature of every modern metaheuristic, explanation requires identification and not just regression. That the honest Causal Forest, fitting a completely different model class to the same trajectories, reproduces the audit's ordering with Spearman correlation 0.90 indicates that the distortion is a property of naive attribution rather than of any particular estimator.

Under what conditions can the causal estimates be wrong? The honest list is short but consequential. State sufficiency may fail, because four statistics cannot summarize everything that drives both assignment and outcome; the propensity model may suffer misspecification; and overlap can fail silently whenever the scheduler stops exploring some action in some region of the state space. Estimator agreement, the audit's internal check, is necessary but not sufficient for correctness, because both estimators adjust for the same logged covariates and can therefore share blind spots. Within-iteration interference between offspring feeds into the noise term and thus into the width of the intervals, and regime nonstationarity can make a window-homogeneous estimate a poor summary of a heterogeneous run. Diagnostics for each failure mode, namely propensity calibration, effective sample size, overlap maps, and regime-stratified estimates, ship with the pipeline, but the assumptions remain the reader's to audit, which is why this paper states rather than implies them.

8 | Limitations and Threats to Validity

Assumption 1 constitutes the central threat. Because the scheduler is a known function of the logged state plus randomization, no unmeasured agent assigns treatment; ignorability, however, is a property of the information set, and a four-dimensional state may compress away confounders such as local curvature or the effective rank of the population, to which both the true dynamics and the fitted models respond. Residual confounding of the reported ATEs therefore cannot be excluded. The window-stability checks of Section 6.7 and the estimator-agreement analysis of Section 6.5 bound it only heuristically, and agreement between two estimators that condition on the same covariates cannot exclude confounding that resides in the unlogged part of the state; formal sensitivity analysis, in the sense of quantifying how strong an unmeasured confounder must be to overturn the sign of an effect, is planned as immediate follow-up work.

Three further limitations bound the scope of the claims.

Benchmark scope: 10 problems on four manifold species, chosen for known optima and a spread of curvature regimes, omit discrete and combinatorial manifolds, very high dimensions, and noisy or expensive objectives, and the convergence argument assumes exact evaluations with compact sublevel sets.

Compute: The audit costs 7% to 12% wall-clock, and equivalence with the fastest baselines could be certified on only six of the ten instances, failing on the retraction-dominated Stiefel and SPD problems where the geometric machinery dominates the objective.

Model class: The structural model h is locally linear, so nonlinear regime interactions surface only through the state and the shallow regime surrogate, and richer transition models could capture what the current scheduler handles only implicitly. In addition, the regime labels derive from k-means plus expert naming, a defensible but subjective step, and the 85.9% surrogate fidelity is a property of this dataset rather than a universal constant. Finally, all problems are synthetic or teacher-constructed; the missing evidence for the decision-support claims is a deployment study with real design offices, and the claim offered here is readiness rather than proof.

9 | Concluding Remarks and Future Work

IPD reformulates population-based optimization on Riemannian manifolds as a self-observing causal system. Operator doses amount to interventions on a structural model of population dynamics; scheduling maximizes predicted counterfactual progress over a short horizon; attribution reports propensity-weighted ATEs with uncertainty; and a shallow decision tree names the behavioral regimes through which each run passed. On its own terms, the algorithm is competitive, holding the smallest median gap on 7 of the ten problems with a median reduction of 26.2% against the strongest rival on those, convergence in probability, and estimation overhead certified equivalent to the fastest baselines wherever objective evaluation dominates. The durable contribution, on this interpretation, is the demonstrated feasibility of self-auditing search: an optimizer whose account of itself can be held to the same evidentiary standard as its solutions. The audit's effect ordering survives an independent honest-forest estimator at Spearman correlation 0.90, the scheduling gain concentrates on positive-curvature multimodal instances, and the dimension ladders leave the intervention effect statistically unmoved (slope +0.011 per log-dimension, 95% CI [-0.004, +0.026]), evidence that the audit's estimands travel across curvature regimes and scales.

Four fronts organize the future work. Richer structural models, nonparametric or neural, could replace the locally linear h while preserving the identification discipline. For the unconfoundedness assumption, sensitivity analysis should adapt the bounds developed for observational studies to the trajectory setting, because estimator agreement alone cannot certify it. Whether factor effects learned on one landscape predict progress on another is the question that cross-problem transfer of audits would test, turning the audit from a per-run artifact into accumulated organizational knowledge. Finally, human-factors studies with engineering decision-makers are needed, because the ultimate test of an explanation is whether the person accountable for the design can act on it.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Funding

Not applicable.

Data Availability

All data supporting the reported findings in this research paper are provided within the manuscript.

Conflict of Interest

The author declares that they do not have any conflict of interest.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

- [1] Eiben, A. E., & Smith, J. E. (2003). *Introduction to evolutionary computing*. Berlin: springer. <https://doi.org/10.1007/978-3-662-44874-8>
- [2] Talbi, E. G. (2009). *From design to implementation*. Hoboken, NY: John Wiley. https://zeus.inf.ucv.cl/~bcrawford/DiplomadoIA_2024/Cap1_Metaheuristics_Talbi.pdf
- [3] Boumal, N. (2020). *An introduction to optimization on smooth manifolds*. Cambridge University Press. <https://doi.org/10.1017/9781009166164>
- [4] Absil, P. A., Mahony, R., & Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press. <https://www.degruyterbrill.com/document/doi/10.1515/9781400830244/html>
- [5] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
- [6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery (ACM). <https://doi.org/10.1145/2939672.2939778>
- [7] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1–42. <https://doi.org/10.1145/3236009>
- [8] Pearl, J. (2009). *Causality*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
- [9] Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2), 303–353. <https://doi.org/10.1137/S0895479895290954>
- [10] Smith, S. T. (2014). *Optimization techniques on Riemannian manifolds*. <https://doi.org/10.48550/arXiv.1407.5965>
- [11] Nishimori, Y., & Akaho, S. (2005). Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing*, 67, 106–135. <https://doi.org/10.1016/j.neucom.2004.11.035>
- [12] Bonnabel, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9), 2217–2229. <https://doi.org/10.1109/TAC.2013.2254619>
- [13] Absil, P. A., Baker, C. G., & Gallivan, K. A. (2007). Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3), 303–330. <https://doi.org/10.1007/s10208-005-0179-9>
- [14] Boumal, N., Absil, P. A., & Cartis, C. (2019). Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1), 1–33. <https://doi.org/10.1093/imanum/drx080>
- [15] Bhatia, R. (2009). *Positive definite matrices*. Princeton University Press. <https://press.princeton.edu/books/paperback/9780691168258/positive-definite-matrices>
- [16] Higham, N. J. (2008). *Functions of matrices: Theory and computation*. SIAM. <https://doi.org/10.1137/1.9780898717778>
- [17] Förstner, W., & Moonen, B. (2003). A metric for covariance matrices. *Geodesy-The Challenge of the 3rd Millennium* (pp. 299–309). Springer. https://doi.org/10.1007/978-3-662-05296-9_31
- [18] Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. *Proceedings of the 24th International Conference on Machine Learning* (pp. 209–216). Association for Computing Machinery (ACM). <https://doi.org/10.1145/1273496.1273523>
- [19] So, A. M. C., & Ye, Y. (2007). Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2), 367–384. <https://doi.org/10.1007/s10107-006-0040-1>

- [20] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95-International Conference on Neural Networks* (Vol. 4, pp. 1942-1948). IEEE. <https://doi.org/10.1109/ICNN.1995.488968>
- [21] Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2), 159–195. <https://doi.org/10.1162/106365601750190398>
- [22] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688. <https://psycnet.apa.org/doi/10.1037/h0037350>
- [23] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322–331. <https://www.tandfonline.com/doi/abs/10.1198/016214504000001880>
- [24] Pearl, J. (2022). Causal diagrams for empirical research (with discussions). In *Probabilistic and Causal Inference: The Works of Judea Pearl* (pp. 255–316). Association for Computing Machinery (ACM). <https://doi.org/10.1145/3501714.3501734>
- [25] Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. *New York*, 517. <https://doi.org/10.1017/CBO9781139025751>
- [26] Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search*. The MIT Press. <https://doi.org/10.7551/mitpress/1754.001.0001>
- [27] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3), 424–438. <https://doi.org/10.2307/1912791>
- [28] Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2), 211–239. <https://doi.org/10.1177/0049124111404820>
- [29] Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7). <https://doi.org/10.1063/1.5025050>
- [30] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), eaau4996. <https://doi.org/10.1126/sciadv.aau4996>
- [31] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>
- [32] Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685. <https://doi.org/10.2307/2280784>
- [33] Card, D., & Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4), 772–793. <https://doi.org/10.3386/w4509>
- [34] Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC. https://static1.squarespace.com/static/675db8b0dd37046447128f5f/t/677676888e31cc50c2c33877/1735816881944/hernanrobins_WhatIf_2jan25.pdf
- [35] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [36] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software. <https://doi.org/10.1201/9781315139470>
- [37] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. <https://doi.org/10.1007/BF00116251>
- [38] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730). Association for Computing Machinery (ACM). <https://doi.org/10.1145/2783258.2788613>
- [39] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626). IEEE. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html
- [40] Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press. <https://mitpress.mit.edu/9780262082136/adaptation-in-natural-and-artificial-systems/>

- [41] Storn, R., & Price, K. (1997). Differential evolution-A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341–359. <https://doi.org/10.1023/A:1008202821328>
- [42] DaCosta, L., Fialho, A., Schoenauer, M., & Sebag, M. (2008). Adaptive operator selection with dynamic multi-armed bandits. *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation* (pp. 913–920). Association for Computing Machinery (ACM). <https://doi.org/10.1145/1389095.1389272>
- [43] Karafotias, G., Hoogendoorn, M., & Eiben, Á. E. (2014). Parameter control in evolutionary algorithms: Trends and challenges. *IEEE Transactions on Evolutionary Computation*, 19(2), 167–187. <https://doi.org/10.1109/TEVC.2014.2308294>
- [44] Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. <https://doi.org/10.1109/4235.585893>
- [45] García, S., Molina, D., Lozano, M., & Herrera, F. (2009). A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 special session on real parameter optimization. *Journal of Heuristics*, 15(6), 617–644. <https://doi.org/10.1007/s10732-008-9080-4>
- [46] Robbins, H., & Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (pp. 233–257). Elsevier. <https://doi.org/10.1016/B978-0-12-604550-5.50015-8>
- [47] Hartley, R., Trumpf, J., Dai, Y., & Li, H. (2013). Rotation averaging. *International Journal of Computer Vision*, 103(3), 267–305. <https://doi.org/10.1007/s11263-012-0601-0>
- [48] Moakher, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3), 735–747. <https://doi.org/10.1137/S0895479803436937>
- [49] Holland, J. H. (1975). Adaptation in natural and artificial systems. *University of Michigan Press Google Schola*, 2, 29–41. <https://mitpress.mit.edu/9780262082136/adaptation-in-natural-and-artificial-systems/>
- [50] Bécigneul, G., & Ganeva, O. E. (2018). *Riemannian adaptive optimization methods*. <https://arxiv.org/pdf/1810.00760>
- [51] Ring, W., & Wirth, B. (2012). Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2), 596–627. <https://doi.org/10.1137/11082885X>
- [52] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- [53] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://www.jstor.org/stable/4615733>
- [54] Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3), 494. <https://psycnet.apa.org/doi/10.1037/0033-2909.114.3.494>
- [55] Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 143–146). Academic Press. <https://doi.org/10.1145/1978942.1978963>
- [56] Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of Nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1), 3–18. <https://doi.org/10.1016/j.swevo.2011.02.002>
- [57] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.2307/2336267>