DASH: FAST, VALID COUNTERFACTUALS FOR DEEP NETWORKS VIA BATCHED DIRECTIONAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual explanations are most useful when they can be generated with low latency, remain close to the factual input, and satisfy input-domain, categorical, and actionability constraints. Achieving these objectives simultaneously is challenging for deep neural networks. Heuristic methods are often fast but may return invalid counterfactuals, whereas exact methods can certify global proximity but may not finish within practical time limits. We introduce DASH, a batched heuristic search method for finding close, valid, and actionable counterfactuals for deep neural networks under ℓ_1, ℓ_2, and ℓ_∞ objectives. DASH uses directional Lipschitz bounds and local affine models to generate anchors, then ranks and expands promising regions with batched network evaluations. We compare DASH against nine prior heuristic methods and time-limited exact mixed-integer baselines on four tabular datasets, with network depths from 2 to 32, and evaluate scalability on PBMC3k. Across 9,000 tabular query-norm cases, DASH returns a valid counterfactual within 5% of the best heuristic-observed valid distance in 94.6% of cases, with a median CPU search runtime of 0.061 s. PGD-bisect, the baseline with the highest pooled within-5% coverage, meets this criterion in 41.1% of cases, with a median runtime of 0.298 s. These results show that the proposed search maintains high valid proximity across norms while keeping its search runtime practical.

1 INTRODUCTION

Neural networks are increasingly used as predictive models that inform decisions, in settings such as mortgage default assessment from financial transactions (Kvamme et al., 2018), clinical-risk predictions from electronic health records (Rajkomar et al., 2018), and process fault diagnosis from operating measurements (Wu & Zhao, 2018). In such applications, a prediction can prompt a further question: *what sufficiently small change to the input would make the model predict a different, desired outcome?* For example, a borrower may seek changes to financial attributes that would enable them to cross a credit approval threshold, while an engineer may seek nearby operating conditions associated with a desired process state. Counterfactual explanations (CFEs) formalize this query by finding an alternative input that produces a desired target prediction while minimizing its deviation from the original or *factual* input (Wachter et al., 2018; Karimi et al., 2020). Let f denote a fixed predictor, x^0 a factual input, $\mathcal{Y}_{\text{tar}}$ the target-output set, and $\mathcal{A}(x^0)$ the set of *actionable* modifications. CFEs can be mathematically formulated as an optimization problem

$$\min_x \quad d_\nu(x, x^0) \quad \text{subject to} \quad f(x) \in \mathcal{Y}_{\text{tar}}, \quad x \in \mathcal{A}(x^0), \tag{1}$$

where $d_\nu(x, x^0)$ is a normalized ℓ_ν distance between x and x^0. We call a point *target-valid* when it satisfies the target constraint $f(x) \in \mathcal{Y}_{\text{tar}}$, *actionable* when $x \in \mathcal{A}(x^0)$, and a *valid* counterfactual when both conditions hold. The actionable set may impose feature bounds (e.g., employment duration), immutable coordinates (e.g., location), permitted directions of change (e.g., age), and categorical assignments (e.g., employment type) (Ustun et al., 2019). CFE methods can be measured on four evaluation criteria: likelihood of producing valid counterfactuals (*validity*), time it takes to produce a counterfactual (*runtime*), proximity to the factual (*distance*), and whether it provides a certificate of proximity or global optimality (*optimality certificate*).

Existing CFE methods make different compromises among these criteria. Penalty-based heuristic methods like Wachter (Wachter et al., 2018), S-CFE (Sadiku et al., 2025), and DiCE (Mothilal et al.,

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

2020) replace the target constraint with a differentiable prediction-loss term and optimize it jointly with an input distance penalty. Consequently, their termination condition does not guarantee validity or distance optimality of the proposed counterfactual. Other heuristics use discrete or sample-based search (Brughmans et al., 2024). However, they neither enforce actionability constraints nor give an optimality guarantee. Exact approaches instead encode the predictor neural network, target constraint, and the actionable set as a satisfiability or mixed-integer optimization problem (Karimi et al., 2020; Mohammadi et al., 2021). ReLU networks permit exact mixed-integer linear encoding due to their piecewise affine nature, whereas other non-linear activations require either mixed-integer non-linear encoding or mixed-integer linear approximation. Finding a minimum-distance counterfactual even for ReLU networks is NP-hard (Sälzer & Lange, 2021; Weng et al., 2018). Consequently, solvers can spend substantial computation proving optimality even after a valid counterfactual has been found, limiting their applicability to large networks under practical runtime budgets.

These approaches leave a gap between heuristic methods, whose returned outputs may violate target or actionability conditions, and exact methods, which may not complete within a practical time limit. To address target-validity, actionability and runtime, we propose DASH, a heuristic search method that balances all three axes for ℓ_1, ℓ_2, and ℓ_∞ norms. DASH uses local sensitivity information and linearization to estimate the decision boundary to achieve high validity, a strict enforcement of actionability constraints at each iteration, and batching the network evaluations to reduce computational overhead. We evaluate this approach against prior heuristic methods, and exact, time-limited, mixed-integer baselines on four tabular datasets and the PBMC3k single-cell RNA sequencing dataset. Pooled across all 9,000 tabular query-norm cases, DASH finds a valid counterfactual within 5% of the best heuristic-observed valid distance ~ 2.3 times more often than PGD-bisect, the baseline with the highest pooled within-5% coverage, with a ~ 5 times faster median CPU search runtime. Whereas prior scalability experiments (Mohammadi et al., 2021) considered architectures containing at most approximately $18{,}000$ parameters, our PBMC3k evaluation includes a network with more than $100{,}000$ parameters. As a heuristic, DASH does not certify non-existence or global optimality. Nevertheless, on the $3{,}448$ tabular query-norm cases for which a time-limited exact solver certified the optimum, DASH returned a counterfactual within 1% of that optimum in 82.7% of cases. We measure distance and validity of the found counterfactual, and runtime of the method, under ℓ_1, ℓ_2, and ℓ_∞ objectives, and additionally evaluate a GPU implementation on PBMC3k.

2 Related Work

Adversarial Methods Related to CFEs are minimum-norm adversarial boundary-crossing problems (Pawelczyk et al., 2022). Adversarial attack methods solve this through margin penalties or barriers (Carlini & Wagner, 2017; Finlay et al., 2019), adaptive norm constraints (Pintor et al., 2021) or local affine boundary models (Moosavi-Dezfooli et al., 2016; Croce & Hein, 2020). Our PGD-bisect baseline combines projected gradient descent (Madry et al., 2018) with bisection. Applying these search ideas to mixed-feature counterfactuals requires explicit treatment of actionable constraints like immutable features, categorical assignments, and permitted directions of change (Ustun et al., 2019).

Heuristic CFE methods. For differentiable predictors such as neural networks, several CFE methods have used first-order optimization. Wachter et al. (2018) convert the target constraint into a differentiable quadratic loss, and jointly optimize it with a normalized distance function, using gradient descent algorithms. DiCE (Mothilal et al., 2020) and S-CFE (Sadiku et al., 2025) add a diversity term and a non-smooth ℓ_p ($0 \leq p \leq 1$) sparsity penalty term respectively. These objectives guide the search, but minimizing a penalty does not by itself certify that the returned point satisfies every target and actionability constraint. Other heuristic CFE methods rely on sampling. NICE (Brughmans et al., 2024) identifies the nearest training instance with the desired class and successively replaces coordinates of the factual input with the corresponding values from the chosen training instance. Growing Spheres (Laugel et al., 2018) samples increasingly distant hyperspherical shells to locate a point with the desired prediction, whereas the Greedy method of Schut et al. (2021) uses Jacobian-guided feature updates. The limitation of these heuristic methods is that they often fail to find valid counterfactuals.

Exact counterfactual methods. Certifying a nearest counterfactual requires finding a valid counterfactual and ruling out all closer actionable points. One family of exact methods uses satisfiability modulo theories (SMT). ReLUplex (Katz et al., 2017) develops an SMT-based verification method for ReLU networks, while MACE (Karimi et al., 2020) applies SMT to counterfactual explanations using repeated queries to bound the optimal distance. An alternative is mixed-integer programming (MIP), where ReLU networks can be encoded exactly with mixed-integer linear constraints (Fischetti & Jo, 2018; Tjeng et al., 2019). Mohammadi et al. (2021) use this approach to compute nearest counterfactuals with guarantees. Bound tightening, presolve, and stronger ReLU formulations can improve the tractability of the MIP formulations (Fischetti & Jo, 2018; Grimstad & Andersson, 2019; Tjeng et al., 2019; Tsay et al., 2021; Huchette et al., 2026). Despite these advances, exact certification can remain computationally expensive for larger networks.

3 METHODOLOGY

Notations and problem definition The counterfactual problem and distance d_ν are defined in Eq. 1. The target constraint is defined as $f(x) \leq \tau$, where $f : \mathbb{R}^d \to \mathbb{R}$. We assume that f is represented by a ReLU neural network and is therefore piecewise affine. Appendix A.1 details the class of counterfactuals covered by this definition. DASH searches using the scalar score $\phi(x) = f(x) - \tau$, with the desired output satisfying $\phi(x) \leq -\mu$, where $\mu \geq 0$ is a verification margin. We define the target gap and target-improving gradient as

$$\Delta(x) = [\phi(x) + \mu]_+, \qquad g(x) = -\nabla\phi(x). \tag{2}$$

For a continuous feature j, the notation $x^{j\leftarrow u}$ replaces only $x_j \in \mathbb{R}$ with $u \in \mathbb{R}$. The sets Λ_{ax}, Λ_{gr}, and Λ_{fac} denote the step-size multipliers used by the anchor-generation mechanisms detailed in the next subsection. Their values are provided in Appendix A.5.

Overview of DASH We call every point generated during the search an *anchor*. At iteration t, $\mathcal{B}_t$ is the beam of target-invalid anchors. DASH first perturbs single features to produce one-directional anchors $\mathcal{Y}_t^{\mathrm{dir}}$. One-feature movements test whether a small intervention can reach the target. DASH then ranks them to select at most K_{dir} target-invalid anchors from $\mathcal{Y}_t^{\mathrm{dir}} \cup \mathcal{B}_t$ for gradient evaluation; these selected anchors form $\mathcal{G}_t$. Local gradient-based anchors are generated from $\mathcal{G}_t$. As the different anchors could lie in different activation regions of the network (Berzins, 2023), DASH also generates factual-origin anchors from x^0 using local affine models at $\mathcal{B}_t \cup \mathcal{G}_t$. Target-valid anchors update the incumbent. A final ranking retains at most K_{beam} target-invalid anchors for $\mathcal{B}_{t+1}$. The search runs for at most T iterations, with the map Π_{act} enforcing actionability on generated anchors. The following subsections define these anchor sets and their ranking rule; Algorithm 1 assembles them.

3.1 ANCHOR GENERATION

One-directional anchors. One-directional anchors change one feature at a time. For each continuous feature j, DASH precomputes a global coordinate-wise Lipschitz bound L_j from the network weights (Appendix A.2; see also Weng et al., 2018) satisfying

$$|\phi(x + he_j) - \phi(x)| \leq L_j|h| \tag{3}$$

over the bounded input domain.

Proposition 1 (One-feature exclusion bound) *Let $\phi(x) > -\mu$ and $L_j > 0$. For any $y = x + he_j$ in the input domain satisfying $\phi(y) \leq -\mu$,*

$$|h| \geq \frac{\Delta(x)}{L_j}.$$

The bound motivates the one-directional anchor scale $\rho_j(x) = \Delta(x)/(L_j + \epsilon)$. Let J_{cont} contain mutable continuous-feature indices with permitted intervals $[a_j, b_j]$. Let J_{cat} contain the one-hot coordinate blocks representing mutable categorical features. For $j \in J_{\mathrm{cat}}$, let $V_j(x^0)$ be the set of

permitted vertices and let $x^{j\leftarrow v}$ replace the entire block by v. The resulting one-feature anchor set is

$$\mathcal{Y}_t^{\mathrm{dir}} = \Pi_{\mathrm{act}}\Bigg(\bigcup_{x\in\mathcal{B}_t}\Big[\bigcup_{j\in J_{\mathrm{cont}}}\big(\{x \pm m\rho_j(x)e_j : m\in\Lambda_{\mathrm{ax}}\}\cup\{x^{j\leftarrow a_j}, x^{j\leftarrow b_j}\}\big) \cup \bigcup_{j\in J_{\mathrm{cat}}}\{x^{j\leftarrow v} : v\in V_j(x^0),\ v\neq x_j\}\Big]\Bigg), \tag{4}$$

The first union searches the continuous variables while the second union searches the categorical variables. If ϕ is monotone along a permitted feature interval, evaluating its endpoint guarantees that a target-valid anchor is found if one exists on that interval.

Gradient-based anchors While the one-directional anchors perturb a single feature, gradient-based anchors change multiple features simultaneously by using the gradient at a target-invalid anchor x^k in two complementary ways. The gradient at x^k defines a local affine model

$$\phi(x^k + h) \approx \phi(x^k) - g(x^k)^\top h. \tag{5}$$

Under this model, the displacement h must satisfy $g(x^k)^\top h \geq \Delta(x^k)$ to reach the decision boundary. Let ν^* denote the dual-norm of norm ν. By norm duality, the estimated minimum distance and the associated steepest directions are

$$\widehat{r}_\nu(x^k) = \frac{\Delta(x^k)}{\|g(x^k)\|_{\nu^*} + \epsilon}, \tag{6}$$

$$v_\infty(g) = \mathrm{sign}(g), \quad v_2(g) = \frac{g}{\|g\|_2}, \quad v_1(g) = \mathrm{sign}(g_{j^\star})e_{j^\star}, \quad j^\star \in \arg\max_j |g_j|. \tag{7}$$

Local gradient-based anchors continue from x^k along $v_\nu(g(x^k))$. From the selected target-invalid anchors $\mathcal{G}_t$, the local gradient-based anchor set is

$$\mathcal{Y}_t^{\mathrm{loc}} = \Pi_{\mathrm{act}}\Bigg(\bigcup_{x\in\mathcal{G}_t}\{x + m\widehat{r}_\nu(x)v_\nu(g(x)) : m\in\Lambda_{\mathrm{gr}}\}\Bigg). \tag{8}$$

A full-gradient step can change many features which decreases sparsity. As most counterfactual domains prefer sparser actions, DASH also generates support-restricted versions using the largest-magnitude gradient features. The configured support sizes are given in Appendix A.5.

Proposition 2 (Exactness of the affine boundary estimate) *Let $\nu \in \{1, 2, \infty\}$, and let x^k satisfy $\phi(x^k) > -\mu$ and $g(x^k) \neq 0$. Define the unregularized affine-boundary anchor*

$$x^\star_\nu = x^k + \frac{\Delta(x^k)}{\|g(x^k)\|_{\nu^*}}\, v_\nu(g(x^k)).$$

If the segment from x^k to $x^\star_\nu$ remains in one affine region of ϕ, then $\phi(x^\star_\nu) = -\mu$. Moreover, $x^\star_\nu$ minimizes $\|x - x^k\|_\nu$ over the target half-space of that local affine model.

Factual-origin anchors use the same direction and local affine model but originate from the factual input. This construction is motivated by seeking a boundary point near x^0 under that local model, rather than taking another step from x^k. DASH estimates the decision boundary radius $r_{\mathrm{fac}}(x^k)$ from the factual, based on the local affine model at x^k,

$$\phi(x^k) - g(x^k)^\top\big(x^0 + r_{\mathrm{fac}}(x^k)v_\nu(g(x^k)) - x^k\big) = -\mu. \tag{9}$$

The corresponding anchor set generated from the factual uses gradients evaluated at both the current beam $\mathcal{B}_t$ and the anchors selected for gradient evaluation $\mathcal{G}_t$:

$$\mathcal{Y}_t^{\mathrm{fac}} = \Pi_{\mathrm{act}}\Bigg(\bigcup_{x\in\mathcal{B}_t\cup\mathcal{G}_t}\{x^0 + mr_{\mathrm{fac}}(x)v_\nu(g(x)) : m\in\Lambda_{\mathrm{fac}}\}\Bigg). \tag{10}$$

The two sets may overlap when a beam anchor is selected again in $\mathcal{G}_t$. Geometrically, Eqn. 9 intersects the ray from x^0 with the decision hyperplane predicted by the affine model at x^k. Figure 1 illustrates how the two anchor-generation mechanisms expose different affine regions. Appendix D.1 reports a matched ablation testing the contribution of factual-origin anchors to proximity.

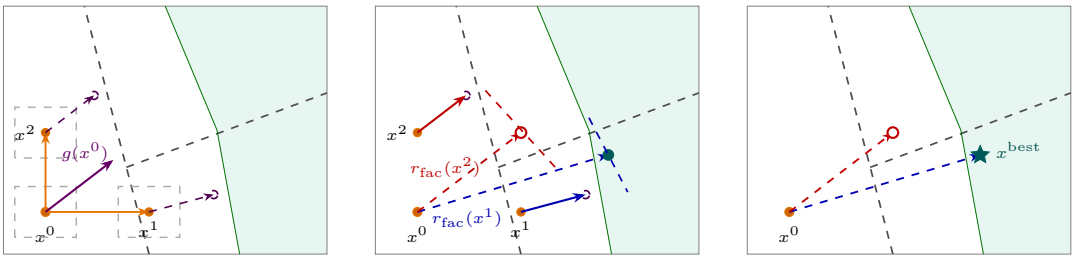

Figure 1: **Anchor generation across ReLU regions.** Black dashed lines separate activation regions, the green line is the target boundary, and the shaded region is target-valid. **left:** One-directional anchors x^1 and x^2 are in different regions. The gradient at each anchor is illustrated with colored dashed lines. **center:** Solid arrows show local gradient-based anchors from x^1 and x^2; dashed arrows show factual-origin anchors from the factual input x^0. The latter use radii $r_{\text{fac}}(x^1)$ and $r_{\text{fac}}(x^2)$, as estimates to the target boundary (Eq. 9). **right:** The red anchor will be used for further anchor generation. The blue anchor is the new incumbent x^{best}.

Actionability enforcement. For DASH, search-time actionability comprises feature bounds, immutable features, permitted directions of change, and valid categorical assignments. The enforcement map Π_{act} applies these constraints to each generated anchor and acts elementwise on anchor sets. Its implementation is given in Appendix A.3.

Proposition 3 (Actionability preservation) *Every anchor at every iteration evaluated by* DASH *belongs to* $\mathcal{A}(x^0)$. *In particular,* $x \in \mathcal{A}(x^0), \quad \forall x \in \mathcal{B}_t, \quad \forall t.$

Proofs of Propositions 2 and 3 are provided in Appendix A.4.

3.2 Anchor Ranking

After evaluating a target-invalid anchor x^ℓ, DASH estimates the total distance of a continuation through it using the beam ranking

$$q_{\text{beam}}(x^\ell) = d_\nu(x^\ell, x^0) + \frac{\Delta(x^\ell)}{\kappa(x^\ell) + \epsilon}, \qquad \kappa(x^\ell) = \frac{[\phi(x^0) - \phi(x^\ell)]_+}{d_\nu(x^\ell, x^0) + \epsilon}. \tag{11}$$

Here, $\kappa(x^\ell)$ is the observed target improvement per unit distance from x^0. The first term is the distance already incurred, while the second estimates the additional distance required if the observed rate persists.

At iteration t, the evaluated target-invalid anchors are considered in increasing order of q_{beam} to choose at most K_{beam} anchors. A diversity term is also applied to choose a diverse set of regions.

$$\min_{z \in \mathcal{S}} d_\nu(x^\ell, z) \geq \delta_{\text{div}}, \tag{12}$$

where $\mathcal{S}$ is the set already selected. Applied with capacity K_{dir}, this rule selects the anchors expanded by the gradient-based search in the current iteration. Applied with capacity K_{beam}, it constructs the beam carried to the next iteration. The score concentrates computation on anchors with favorable estimated distance. In Algorithm 1, R_K applies this rule with capacity K.

3.3 Efficiency Enhancements

DASH has CPU and GPU implementations. Both evaluate anchors and compute gradients in batches. The CPU implementation screens one-directional anchors before network evaluation using the directional score q_{dir} (Appendix A.2). The GPU implementation omits this screening and performs anchor generation, ranking, network evaluation, gradient computation, and beam selection on the GPU device. Finite step-size multiplier sets, support-size sets, and beam capacity control time and memory. The exact numerical settings and tie-breaking rules are reported in Appendix A.5 and Appendix A.2. Additionally, if a target-valid anchor has been found when the search terminates,

Algorithm 1 DASH: batched anchor search

Require: Predictor f, factual x^0, threshold τ, actionable set $\mathcal{A}(x^0)$, Lipschitz constants $\{L_j\}_{j=1}^{d}$, iteration limit T, directional-selection capacity K_{dir}, and beam capacity K_{beam}
Ensure: Verified target-valid, search-actionable counterfactual x^{best} or NONE
Set $d_\nu(\mathrm{NONE}, x^0) = +\infty$ for incumbent comparisons.
Initialize: $\mathcal{B}_0 \leftarrow \{x^0\}$, $x^{\mathrm{best}} \leftarrow$ NONE
for $t = 0, \ldots, T-1$ **do**
One-directional anchors and selection for gradient evaluation
    One-directional anchors: $\mathcal{Y}_t \leftarrow \mathcal{Y}_t^{\mathrm{dir}}$ (Eq. 4); $x^{\mathrm{best}} \leftarrow \mathsf{Best}(x^{\mathrm{best}}, \mathcal{Y}_t)$
    Rank anchors for gradient evaluation: $\mathcal{G}_t \leftarrow \mathsf{R}_{K_{\mathrm{dir}}}(\{y \in \mathcal{Y}_t \cup \mathcal{B}_t : \phi(y) > -\mu,\ d_\nu(y, x^0) < d_\nu(x^{\mathrm{best}}, x^0)\})$ (Eqs. 11, 12)
Local gradient-based and factual-origin anchors
    if $\mathcal{G}_t \neq \varnothing$ **then**
        Local gradient-based anchors: $\mathcal{Y}_t \leftarrow \mathcal{Y}_t \cup \mathcal{Y}_t^{\mathrm{loc}}$ (Eq. 8)
        Factual-origin anchors: $\mathcal{Y}_t \leftarrow \mathcal{Y}_t \cup \mathcal{Y}_t^{\mathrm{fac}}$ (Eq. 10)
Beam update with target-invalid anchors that are closer than incumbent
    $x^{\mathrm{best}} \leftarrow \mathsf{Best}(x^{\mathrm{best}}, \mathcal{Y}_t)$
    if $\mathcal{G}_t = \varnothing$ **then**
        break
    Rank next beam: $\mathcal{B}_{t+1} \leftarrow \mathsf{R}_{K_{\mathrm{beam}}}(\{y \in \mathcal{Y}_t : \phi(y) > -\mu,\ d_\nu(y, x^0) < d_\nu(x^{\mathrm{best}}, x^0)\})$ (Eqs. 11, 12)
    if $\mathcal{B}_{t+1} = \varnothing$ **then**
        break
Final polishing and bisection
if $x^{\mathrm{best}} \neq$ NONE **then**
    $\widetilde{x} \leftarrow \mathsf{Polish}(x^0, x^{\mathrm{best}})$ (Subsection 3.3)
    if $\widetilde{x} \in \mathcal{A}(x^0)$, $\phi(\widetilde{x}) \leq -\mu$, and $d_\nu(\widetilde{x}, x^0) < d_\nu(x^{\mathrm{best}}, x^0)$ **then**
        Accept verified improvement: $x^{\mathrm{best}} \leftarrow \widetilde{x}$
return x^{best}

DASH can reduce its distance by bisection over the continuous features while holding its categorical assignments fixed. We denote this operation by $\mathsf{Polish}(x^0, x^{\mathrm{best}})$. The polished anchor is accepted only after validity verification.

4 Experiments and Results

4.1 Experimental Setup

Datasets and predictors. Our primary testbed is actionable tabular counterfactual search, where continuous and categorical features coexist. We evaluate HELOC, Give Me Some Credit (GMC), Adult Income, and COMPAS (Pawelczyk et al., 2021) using fully connected ReLU networks of width 64 and depths $2, 4, 8, 16, 32$. We also evaluate PBMC3k (10x Genomics, 2017) to test the same search on a wider network. MNIST is used to test a no-actionable dataset, whose results are reported in Appendix C.1. For each of three training seeds, we select 50 correctly classified held-out factual inputs and seek the opposite class with verification margin $\mu = 10^{-4}$. Each method receives the same factual inputs under each objective norm. This yields 9,000 tabular and 450 PBMC3k queries per method. Data preparation and query selection are described in Appendices B.1 and B.4.

Baselines and budgets. We compare DASH with Wachter (Wachter et al., 2018), DiCE (Mothilal et al., 2020), Greedy (Schut et al., 2021), NICE (Brughmans et al., 2024), FMN (Pintor et al., 2021), FAB (Croce & Hein, 2020), DeepFool (Moosavi-Dezfooli et al., 2016), norm-extended versions of Growing Spheres (GS) (Laugel et al., 2018) and PGD-bisect (Madry et al., 2018), and with the MIP-OBJ and MIP-EXP exact formulations (Mohammadi et al., 2021) on tabular networks of depths $2, 4, 8$ due to the heavy computational overhead. Per-query limits are 60 seconds for tabular and PBMC3k. The MNIST budget is reported in Appendix C.1. CPU experiments used AMD EPYC

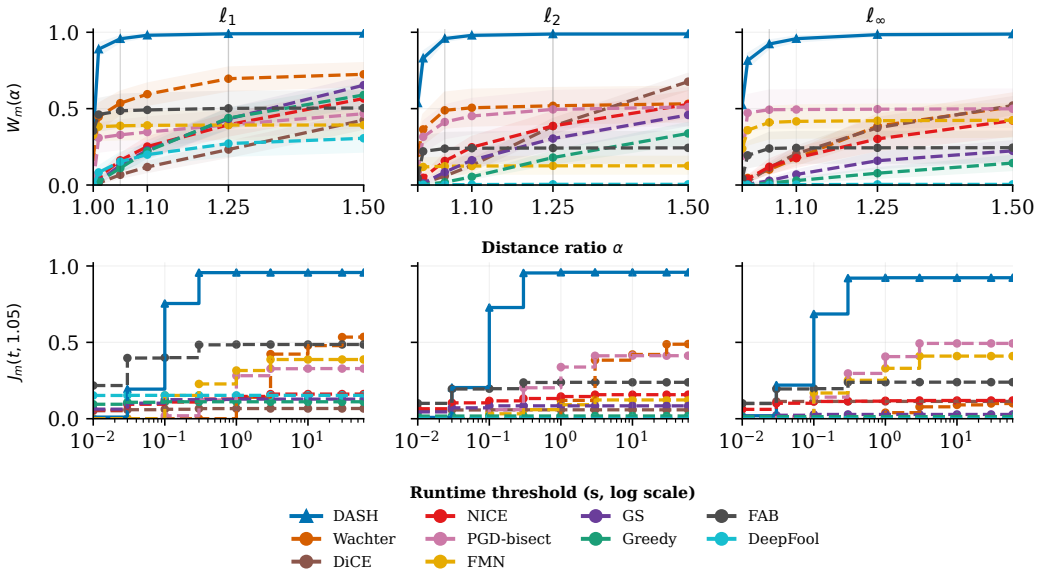

Figure 2: **Tabular performance profiles.** Columns show the ℓ_1, ℓ_2, and ℓ_∞ objectives; rows show $W_m(\alpha)$ and $J_m(t, 1.05)$. Results pool four datasets, five network depths, and three training seeds. Top-row shading gives 95% bootstrap intervals clustered by trained model (dataset $\times$ depth $\times$ seed).

7643; GPU experiments used one NVIDIA A100-40gb. Further hardware details, baseline settings, and common verification are given in Appendices B.4 and B.5.

Constraints and output verification. Tabular queries use dataset-specific feature bounds, immutability, permitted change directions, and categorical constraints. PBMC3k additionally requires increases of at least 0.05 units in two markers from a prespecified eight-gene pool. Outputs are then verified. The principal DASH search uses $T = 10$, beam capacity $K_{\text{beam}} = 16$, and directional-selection capacity $K_{\text{dir}} = 16$; other settings and detailed constraints are given in Appendices A.5, B.2, and B.3.

Metrics Let $m \in \mathcal{M}$ index DASH and the heuristic baselines available for the corresponding queries, and let $q \in \mathcal{Q}$ index the factual-predictor-norm queries. Let $x_{q,m}$ and $t_{q,m}$ denote method m's verified output and recorded search time on query q, respectively. Define

$$d_q^{\text{best}} = \min_{\substack{m \in \mathcal{M} \\ x_{q,m} \text{ valid}}} d_\nu(x_{q,m}, x_q^0), \qquad R_{q,m} = \begin{cases} d_\nu(x_{q,m}, x_q^0)/d_q^{\text{best}}, & x_{q,m} \text{ valid}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{13}$$

The reference d_q^{best} is the best heuristic-observed valid distance, not a certified optimum; the MIP formulations are compared with certified optima in Section 4.2. Thus, $R_{q,m}$ measures the distance achieved by method m relative to this reference for query q, with $R_{q,m} = 1$ indicating the best heuristic-observed performance. Our distance and joint quality-runtime profiles are

$$W_m(\alpha) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}[R_{q,m} \le \alpha], \qquad J_m(t, \alpha) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}[R_{q,m} \le \alpha \wedge t_{q,m} \le t]. \tag{14}$$

Queries for which a method returns an invalid output or no output remain in the denominators and therefore count as failures. Thus, $W_m(1.05)$ is the fraction of queries for which method m returns a valid counterfactual within 5% of the best heuristic-observed valid distance, while $J_m(t, 1.05)$ additionally requires the search to finish within time t. These profiles follow the relative-performance comparison of Dolan & Moré (2002).

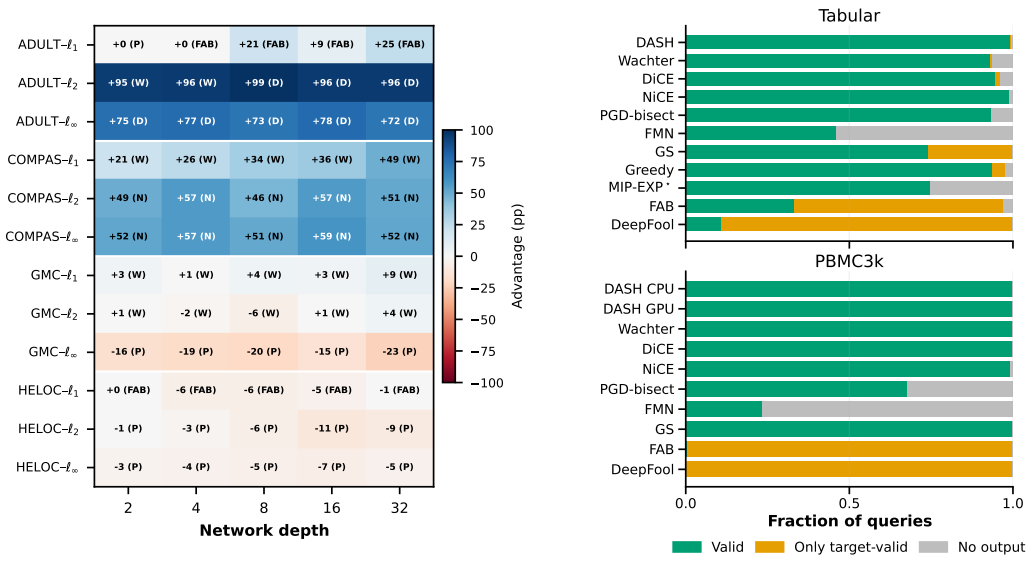

Figure 3: **Tabular robustness and output outcomes.** Left: Percentage-point difference in $W_m(1.05)$ between DASH and the cell-wise best prior heuristic. Parentheses identify that heuristic (W: Wachter, D: DiCE, N: NICE, P: PGD-bisect; FAB is written in full). Right: Final verification outcomes for the methods in the outcome analysis on tabular and PBMC3k queries, pooled across norms. "Only target-valid" denotes a target-valid but constraint-invalid output. MIP-EXP* covers tabular depths 2, 4, and 8 only.

4.2 Results

Tabular counterfactuals Figure 2 separates distance and joint quality-runtime performance by objective norm. DASH's distance profile lies above the comparator profiles for every $\alpha \geq 1.01$ under each norm. At $\alpha = 1.05$, it covers 95.7%, 95.9%, and 92.3% of tabular queries under ℓ_1, ℓ_2, and ℓ_∞, respectively. Wachter has the second-highest coverage under ℓ_1 and ℓ_2, whereas PGD-bisect does under ℓ_∞. The runtime row shows that the separation persists when unsuccessful queries and search time are counted jointly: DASH leads every comparator from 0.1 s on and reaches a higher plateau under each norm. The other baselines ordering changes with the norm, especially as PGD-bisect and FMN improve under ℓ_∞.

Depth and certified distance The left side of Figure 3 resolves the pooled tabular comparison by dataset, norm, and depth. DASH has a positive valid-within-5% advantage over the strongest comparator in 36 of 60 configurations, including every COMPAS cell and every Adult cell under ℓ_2 and ℓ_∞, and ties in three. The reversals are GMC under ℓ_∞ and HELOC, where the norm-adapted PGD-bisect (ℓ_2, ℓ_∞) and FAB (ℓ_1, depths 4-32) have higher coverage, and two GMC ℓ_2 cells, where Wachter does. The heatmap shows no shared monotonic decline in the DASH advantage as depth increases.

The bottom row of Figure 4 compares outputs with certified optima on the subset of tabular queries for which an exact method finishes within the time limit with a certified optimum. DASH returns a valid output on all 1,071, 1,022, and 1,355 certified queries under ℓ_1, ℓ_2, and ℓ_∞, respectively; 99.0%, 95.9%, and 92.3% are within 5% of the corresponding certified optimum. These fractions include all certified queries in their denominators rather than conditioning on a method's successful returns. Appendix C.2 defines the plotted gap.

Wider network The upper row of Figure 4 tests whether the tabular search behavior carries to the wider PBMC3k network. The corresponding MNIST comparison appears in Appendix C.1. On PBMC3k, all methods receive the same returned-output marker completion before validity and distance are scored; Greedy did not run at this scale. DASH CPU and GPU end above the comparators under ℓ_1, with final valid-within-5% coverages of 70.7% and 72.0% (Growing Spheres, the next

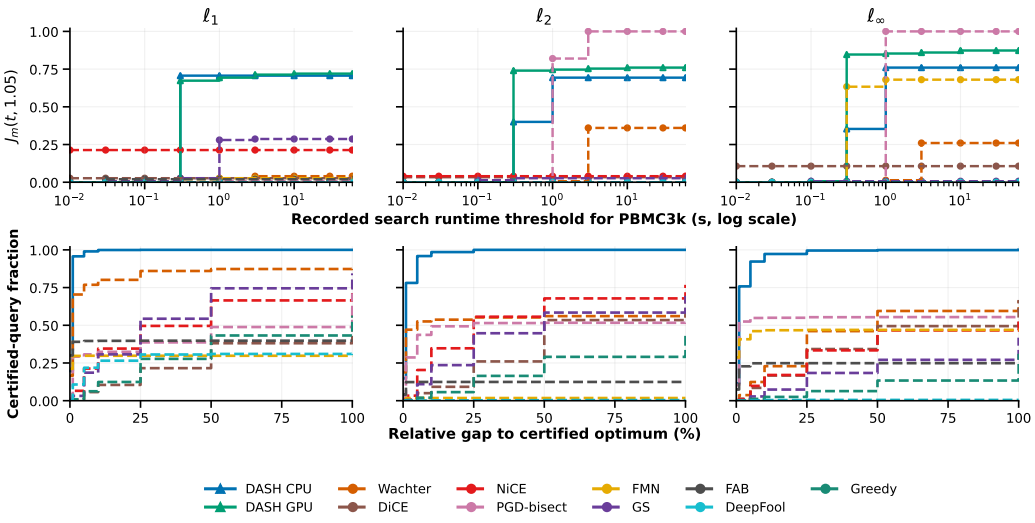

Figure 4: **PBMC3k runtime and tabular certified distance.** Columns show the ℓ_1, ℓ_2, and ℓ_∞ objectives. The upper row reports PBMC3k joint quality-search-runtime profiles after the same output-only marker completion for every method; recorded times exclude completion. The lower row reports the fraction of tabular queries with a valid output within each relative gap to a certified optimum. Certification covers depths 2, 4, and 8; unsuccessful returns remain in the denominator.

best, reaches 28.7%). Under ℓ_2 and ℓ_∞, PGD-bisect reaches 100%, above DASH GPU (76.0% and 87.3%) and CPU (69.3% and 76.0%). The three norms therefore do not exhibit the same method ordering in this high-dimensional setting.

Failure modes. The right side of Figure 3 separates invalid returns from queries with no output. On tabular queries, PGD-bisect returns valid counterfactuals in 93.6% of cases despite meeting the within-5% criterion in 41.1%; FMN more often returns no output, while FAB and DeepFool predominantly return target-valid but constraint-invalid points, including on PBMC3k after shared completion. This figure illustrates a major difference between CFE methods, built mostly for tabular recourse, and adversarial attack methods, built mostly for high dimensional datasets with simple bound constraints.

4.3 Conclusion

Discussion. The results distinguish finding a target-valid point from finding a close counterfactual that satisfies the specified actionability constraints. Across the tabular tasks that contain both continuous and categorical variables, DASH combines high valid-within-5% coverage with short search times under all three norms. Its proximity is also supported on the subset where an exact solver certifies the optimum. These findings support boundary-guided, actionable search as a practical approach to tabular counterfactual generation, especially when categorical variables coexist with continuous variables. For continuous only datasets, like PBMC3k, DASH outperforms all methods on ℓ_1 and remains competitive otherwise.

Limitations and future work. DASH does not certify global optimality or non-existence: a finite beam and runtime budget can cause it to miss an existing counterfactual. Its search also relies on local gradient and boundary estimates, and the current implementation is limited to ReLU networks and the actionability constraints considered here. Future work could combine DASH's incumbent with certified lower bounds and extend the search to causal, user-specific, or more general linear constraints. On some continuous-only settings and norms, adversarial attack methods outperform the current configuration; adapting the balance between one-directional and gradient-based anchors may improve performance in those settings.

AI USE STATEMENT

In this work, generative AI tools assisted with code implementation, figure preparation, and manuscript polishing. The code was reviewed and run through multiple tests to ensure correctness. The figure and plots were manually verified with final results. It was not used for other tasks with required disclosures. The authors take responsibility for the final content, including all AI-assisted claims and artifacts.

REPRODUCIBILITY STATEMENT

Algorithm 1 specifies the search. The appendix provides proofs, implementation settings, dataset preprocessing and actionability rules, predictor training and query selection, baseline configurations, and evaluation procedures (Appendices A.4-B.5). Code, the trained predictors, every result row, and scripts that regenerate all tables and figures are available in an anonymized repository: `https://anonymous.4open.science/r/dash-counterfactuals-D37E`.

REFERENCES

10x Genomics. 3k PBMCs from a healthy donor. Single Cell Gene Expression Dataset, 2017.

Franck Barthe, Olivier Guédon, Shahar Mendelson, and Assaf Naor. A probabilistic approach to the geometry of the ℓ_p^n-ball. *The Annals of Probability*, 33(2):480–513, 2005. doi: 10.1214/009117904000000874.

Arturs Berzins. Polyhedral complex extraction from relu networks using edge subdivision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Dieter Brughmans, Pieter Leyman, and David Martens. NICE: An algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery*, 38(5):2665–2703, 2024. doi: 10.1007/s10618-023-00930-y.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2196–2205. PMLR, 2020.

Francesco Croce and Matthias Hein. Mind the box: l_1-APGD for sparse adversarial attacks on image classifiers. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2201–2211, 2021.

Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002. doi: 10.1007/s101070100263.

Chris Finlay, Aram-Alexandre Pooladian, and Adam M. Oberman. The LogBarrier adversarial attack: Making effective use of decision boundary information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4862–4870, 2019. doi: 10.1109/ICCV.2019.00496.

Matteo Fischetti and Jason Jo. Deep neural networks and mixed integer linear optimization. *Constraints*, 23(3):296–309, 2018. doi: 10.1007/s10601-018-9285-6.

Bjarne Grimstad and Henrik Andersson. Relu networks as surrogate models in mixed-integer linear programs. *Computers & Chemical Engineering*, 131:106580, 2019. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2019.106580.

Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*, 2024.

Joey Huchette, Gonzalo Muñoz, Thiago Serra, and Calvin Tsay. When deep learning meets polyhedral theory: A survey. *INFORMS Journal on Computing*, 2026.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 895–905. PMLR, 26–28 Aug 2020.

Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification (CAV)*, volume 10426 of *Lecture Notes in Computer Science*, pp. 97–117. Springer, 2017. doi: 10.1007/978-3-319-63387-9_5.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

Håvard Kvamme, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.02.029.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations (IPMU)*, Communications in Computer and Information Science, pp. 100–111. Springer, 2018. doi: 10.1007/978-3-319-91473-2_9.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. Scaling guarantees for nearest counterfactual explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 177–187. ACM, 2021. doi: 10.1145/3461702.3462514.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016. doi: 10.1109/CVPR.2016.282.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 607–617. ACM, 2020. doi: 10.1145/3351095.3372850.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387, 2016. doi: 10.1109/EuroSP.2016.36.

Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4574–4594. PMLR, 2022.

Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20052–20062, 2021.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, May 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1.

Shpresim Sadiku, Moritz Wagner, Sai Ganesh Nagarajan, and Sebastian Pokutta. S-cfe: Simple counterfactual explanations. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 172–180. PMLR, 03–05 May 2025.

Marco Sälzer and Martin Lange. Reachability is NP-complete even for the simplest neural networks. In *Reachability Problems (RP 2021)*, volume 13035 of *Lecture Notes in Computer Science*, pp. 149–164. Springer, 2021. doi: 10.1007/978-3-030-89716-1_10.

Lisa Schut, Oscar Key, Rory McGrath, Luca Costabello, Bogdan Sacaleanu, Medb Corcoran, and Yarin Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1756–1764. PMLR, 2021.

Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

Calvin Tsay, Jan Kronqvist, Alexander Thebelt, and Ruth Misener. Partition-based formulations for mixed-integer optimization of trained ReLU neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3068–3080, 2021.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 10–19. ACM, 2019. doi: 10.1145/3287560.3287566.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018.

Zhendong Wang, Ioanna Miliou, Isak Samsten, and Panagiotis Papapetrou. Counterfactual explanations for time series forecasting. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1391–1396, 2023. doi: 10.1109/ICDM58522.2023.00180.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5276–5285. PMLR, 2018.

Hao Wu and Jinsong Zhao. Deep convolutional neural network model based chemical process fault diagnosis. *Computers & Chemical Engineering*, 115:185–197, 2018. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2018.04.009.

A DASH IMPLEMENTATION DETAILS

A.1 TARGET ORIENTATION AND OUTPUT SCORES

The binary target may lie on either side of a scalar decision threshold. Let f_{raw} and τ_{raw} denote the original predictor and threshold. Set $s = 1$ when the target is $f_{\mathrm{raw}}(x) \le \tau_{\mathrm{raw}} - \mu$ and $s = -1$ when it is $f_{\mathrm{raw}}(x) \ge \tau_{\mathrm{raw}} + \mu$. The main text uses the oriented quantities $f = sf_{\mathrm{raw}}$ and $\tau = s\tau_{\mathrm{raw}}$. Hence $\phi(x) = f(x) - \tau \le -\mu$ in either case, and $g(x) = -\nabla\phi(x)$ points toward decreasing oriented output. This sign change is without loss of generality for a scalar threshold target; it does not by itself turn an arbitrary vector target into a scalar one.

For a K-class predictor with pre-softmax logits $z(x) \in \mathbb{R}^K$, a specified target class c has the scalar margin

$$\phi_c(x) = \max_{j \ne c}\{z_j(x) - z_c(x)\}. \tag{15}$$

The condition $\phi_c(x) \le -\mu$ says that class c exceeds every other logit by at least μ. If the target is any class other than the factual class y, use $\phi_{\ne y}(x) = z_y(x) - \max_{j \ne y} z_j(x)$; its value is at most $-\mu$ exactly when some competing class exceeds y by that margin. At $\mu = 0$, ties must also be resolved according to the classifier's label rule during final verification. Logit-margin objectives of this form are used for targeted multiclass attacks (Carlini & Wagner, 2017).

For a vector regressor $F(x) \in \mathbb{R}^p$ and a desired output box $\underline{y}_i \le F_i(x) \le \overline{y}_i$ for all i, define

$$\phi_{\mathrm{box}}(x) = \max_i\{\underline{y}_i - F_i(x),\ F_i(x) - \overline{y}_i\}. \tag{16}$$

Then $\phi_{\mathrm{box}}(x) \le -\mu$ is equivalent to $\underline{y}_i + \mu \le F_i(x) \le \overline{y}_i - \mu$ for every output. With $\mu = 0$, this recovers the original box. For a nonempty tightened box, each interval must have width at least 2μ. Output coordinates should be scaled before applying a common margin. Desired output intervals across a forecast horizon have been studied in counterfactual forecasting (Wang et al., 2023). More generally, a polyhedral target $\{y : Ay \le b\}$ has score $\phi(x) = \max_r\{(AF(x))_r - b_r\}$, after scaling its inequalities when a shared verification margin is used.

These reductions preserve the quantities used by the search. Because $|\max_r h_r(x + he_j) - \max_r h_r(x)| \le \max_r |h_r(x + he_j) - h_r(x)|$, if every constituent score h_r has coordinate bound L_{rj}, then a finite maximum or minimum has bound $L_j = \max_r L_{rj}$. For a vector-output ReLU network, the (i, j) entry of $|W_D| \cdots |W_1|$ is a valid bound L^F_{ij} for output i along input feature j, by the same chain-rule argument used for the scalar network in Appendix A.2. A logit difference $z_k - z_c$ has a coordinate bound no larger than the sum of the bounds for z_k and z_c. For the polyhedral score above, a valid bound for constituent r is $L_{rj} \le \sum_i |A_{ri}| L^F_{ij}$, where L^F_{ij} bounds the change in output F_i along input feature j. ReLU logits and ReLU regression outputs make these scores piecewise affine (Berzins, 2023). At a tie, an active branch can supply a search direction, but Proposition 2 assumes that ϕ is differentiable at the anchor and applies only while the network activation pattern and active score term remain fixed along the segment. The current implementation and experiments use scalar-output binary predictors. Applying these reductions to vector-output networks requires supplying the scalar score, its gradient, and its directional bounds; arbitrary nonlinear output targets do not inherit the piecewise-affine exactness claim.

A.2 DIRECTIONAL SCALES AND ANCHOR SCREENING

For a scalar-output Dense/ReLU network with weight matrices $W_1, \ldots, W_D$, the row vector $L = |W_D| \cdots |W_1|$ gives directional bounds L_j. Each ReLU slope lies in $[0, 1]$, so the chain rule shows that this activation-agnostic product bounds the output change under a displacement of feature j as in Eq. 3. Related coordinate-wise bound propagation is studied by Weng et al. (2018). For $L_j > 0$, DASH uses $\rho_j(x) = \Delta(x)/(L_j + \epsilon)$ and the multiplier and endpoint anchors of Eq. 4. A direction with $L_j = 0$ is skipped. The bounds supply anchor scales; they do not certify that a proposed anchor reaches the target.

Proof of Proposition 1. Let $x + he_j$ be target-valid while x is target-invalid. Then $|\phi(x + he_j) - \phi(x)| \ge \phi(x) + \mu = \Delta(x)$, so Eq. 3 gives $L_j|h| \ge \Delta(x)$.

For a one-directional anchor x^ℓ generated from x^k, the CPU screening score is

$$q_{\mathrm{dir}}(x^k, x^\ell) = \frac{d_\nu(x^\ell, x^k)}{\max\{g(x^k)^\top (x^\ell - x^k), \epsilon\}}. \tag{17}$$

The numerator is the movement from x^k, and the denominator is the target improvement predicted by its local affine model. A smaller score indicates more predicted progress per unit movement. The CPU implementation associates each new anchor with the anchor that generated it when computing q_{dir}. If multiple parents generate the same point, score ordering and duplicate detection determine which anchor is admitted first. Screening preserves the lowest-scoring one-directional anchor for each feature-and-sign tag. An explicitly added feature-bound endpoint has a separate tag. Among the remaining continuous anchors, it retains the lowest scores until the total reaches

$$\max\{n_{\mathrm{protected}}, 1, \lceil \gamma N_{\mathrm{cont}} \rceil\},$$

where N_{cont} counts continuous one-directional anchors and $n_{\mathrm{protected}}$ counts protected tags. Permitted categorical transitions are retained separately. This coverage floor applies to one-directional anchors; gradient-based anchors are ranked by the same score without feature-and-sign protection. Factual-origin anchors are evaluated without this CPU screening step. The final tabular CPU suite sets $\gamma = 1$. The batched GPU implementation does not use the CPU directional-screening filter.

For beam selection, anchors are ordered by q_{beam} from Eqn. 11. An anchor is accepted when it is at least δ_{div} from every previously accepted anchor. If fewer than the requested number of anchors meet this separation criterion, the remaining positions are filled by the lowest-scoring rejected anchors. Previously evaluated anchors are identified using a coordinate hash with tolerance 10^{-10}.

A.3 Actionability enforcement

For each continuous feature j, let $I_j(x^0) = [a_j(x^0), b_j(x^0)]$ combine its input bounds, immutability, and permitted direction of change. An immutable feature has $a_j(x^0) = b_j(x^0) = x^0_j$. For a generated anchor z, the enforcement map is

$$[\Pi_{\mathrm{act}}(z)]_j = \min\{b_j(x^0), \max\{a_j(x^0), z_j\}\}.$$

Categorical anchors are generated as permitted one-hot vertices; assignments outside that set are discarded. The map therefore clips feature-wise constraints and preserves valid categorical assignments. It is not a metric projection onto a general mixed-integer actionable set. The PBMC3k marker-group criterion is applied to returned outputs as described in Appendix B.3.

A.4 Proofs of method properties

Proof of Proposition 2. Write $\delta = \Delta(x^k) = \phi(x^k) + \mu > 0$ and $g = g(x^k)$. The directions in Eq. 7 satisfy $\|v_\nu(g)\|_\nu = 1$ and $g^\top v_\nu(g) = \|g\|_{\nu^*}$. On the assumed segment the affine model is exact, so $\phi(x^\star_\nu) = \phi(x^k) - g^\top(x^\star_\nu - x^k) = \phi(x^k) - \delta = -\mu$. Any z in the affine target half-space satisfies $g^\top(z - x^k) \geq \delta$. Hölder's inequality gives $\|z - x^k\|_\nu \geq \delta / \|g\|_{\nu^*}$, and $x^\star_\nu$ attains this bound. The minimizer need not be unique.

Proof of Proposition 3. The factual anchor x^0 satisfies the search-time feature constraints. Continuous clipping maps every generated feature value into its permitted interval, while categorical transitions replace a valid one-hot vertex with another permitted vertex. Thus every admitted anchor belongs to $\mathcal{A}(x^0)$. Each later beam is drawn from evaluated anchors, so induction over beam iterations proves the claim.

A.5 Search configuration and final verification

Table 1 reports the CPU search settings for the final suite. The plotted high-dimensional batched GPU runs likewise use $K_{\mathrm{beam}} = K_{\mathrm{dir}} = 16$, $T = 10$, and the wide gradient-step multipliers, but do not apply the CPU directional-screening filter. Step multipliers probe shorter and longer movements than the local boundary estimate. For each configured support size, the gradient construction retains the actionable features with largest absolute gradient components. A zero usable gradient produces no local gradient-based or factual-origin anchor; a factual-origin ray is also skipped when its estimated

Table 1: DASH search settings.

Setting	Value
Beam capacity K_{beam}	16
Directional-selection capacity K_{dir}	16
Iteration limit T	10
Verification margin μ	10^{-4}
CPU tabular directional-screening fraction γ	1
Axis multipliers Λ_{ax}	$\{1, 2, 4, 8, 16, 32\}$
Gradient-step multipliers Λ_{gr}	$\{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32\}$
Gradient support sizes	$\{1, 2, 4, 8, 16\}$
Factual-origin multipliers Λ_{fac}	$\{0.75, 1, 1.25, 1.5, 2\}$
Numerical gradient and Lipschitz tolerance	10^{-12}
Beam-diversity tolerance δ_{div}	10^{-5}
Anchor-hash tolerance	10^{-10}
Continuous polishing steps	20

Table 2: Processed datasets and query directions.

Dataset	Dim.	Groups	Factual → target	Train / test
Adult	108	8	$\leq$ \$50K →> \$50K	30,162 / 15,060
COMPAS	13	3	recidivated → not recidivated	4,938 / 1,234
GMC	10	0	delinquent → not delinquent	120,000 / 30,000
HELOC	23	0	Bad → Good	7,897 / 1,974
MNIST 3, 8	784	0	3 → 8	11,982 / 1,984
PBMC3k	50	0	other → CD14+ monocyte	2,110 / 528

intersection radius is nonpositive or undefined. All returned anchors undergo the target-validity and search-time actionability checks of Eq. 1. If a valid incumbent exists, DASH performs at most 20 bisection steps on its continuous features toward x^0, holding the incumbent's categorical assignments fixed. The polished result is kept only if it remains target-valid and actionable. A returned anchor is valid under these checks, while NONE does not establish infeasibility.

B EXPERIMENTAL PROTOCOL

B.1 DATASETS AND PREPROCESSING

Table 2 reports the processed input dimensions, one-hot groups, class directions, and train-test counts. Continuous tabular features are scaled to $[0, 1]$ using training-set extrema. Categorical features are one-hot encoded using levels observed in the training data. GMC and HELOC missing values are imputed using training-set medians. COMPAS length of stay is derived from jail admission and release dates. MNIST pixels are scaled to $[0, 1]$. PBMC3k uses 50 selected genes and training-set min-max scaling. The reported d_ν values are ℓ_ν distances in these processed coordinates.

B.2 DATASET-SPECIFIC ACTIONABILITY SPECIFICATIONS

All returned outputs are first required to satisfy the feature bounds induced by preprocessing. Table 3 specifies the additional dataset-specific restrictions. Immutable features remain equal to their factual values, nondecreasing and nonincreasing features may change only in the indicated direction, and categorical features must remain valid one-hot assignments.

The PBMC3k constraints define a marker-based computational intervention model. They restrict the search to expression changes consistent with an increased monocyte-marker signal and a reduced T-cell-marker signal. They do not establish that the corresponding expression changes are biologically feasible interventions.

Table 3: Actionability specifications used in the experiments.

Dataset	Immutable features	Directional constraints	Additional constraints
Adult	Age, fnlwgt, race, sex, and native-country features	Education-num is nondecreasing	Categorical groups retain valid one-hot assignments
COMPAS	None beyond feature bounds	Age and priors-count are nondecreasing	Categorical groups retain valid one-hot assignments
GMC	Age, NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse, NumberOfTimes90DaysLate, and NumberOfDependents	RevolvingUtilizationOfUnsecuredLines and DebtRatio are nonincreasing. MonthlyIncome is nondecreasing	None
HELOC	None beyond feature bounds	ExternalRiskEstimate is nondecreasing. NetFractionRevolvingBurden and NetFractionInstallBurden are nonincreasing	None
MNIST	None beyond pixel bounds	None	Every pixel remains in $[0, 1]$
PBMC3k	All genes outside the marker panels are immutable	CD14, LYZ, FCN1, LST1, CTSS, TYROBP, S100A8, and S100A9 are nondecreasing. CD3D, CD3E, LCK, CD2, CD7, IL7R, LTB, and PTPRCAP are nonincreasing	At least two monocyte-marker genes must increase by at least 0.05

B.3 Common PBMC3k output completion

The marker-group requirement in Table 3 is an additional PBMC3k output constraint applied equally to every method. DASH's search-time actionability set consists of feature bounds, immutable features, permitted directions, and valid categorical assignments. After a method returns an output x, the common final check verifies target validity and all these base constraints. An output that passes is retained. If x reaches the target and satisfies the base constraints but fails the marker-group condition, we enumerate the $\binom{8}{2} = 28$ unordered pairs (i, j) of monocyte markers. For each pair, we set $x_i \leftarrow \max\{x_i, x_i^0 + 0.05\}$ and $x_j \leftarrow \max\{x_j, x_j^0 + 0.05\}$, clip the result to the processed feature bounds, and reevaluate the predictor and every actionability condition. Among passing completed outputs, we keep the one with smallest $d_\nu(\cdot, x^0)$. If none passes, that method has no valid output for the query. This completion is applied once to the returned output of each method, with the same selection rule and verification. The reported search time excludes completion time. The saved benchmark records do not isolate completion wall-clock time, so we do not report a numerical completion-time summary.

B.4 Predictors, queries, and compute budgets

Tabular predictors are fully connected ReLU networks of width 64 and depths $2, 4, 8, 16$, and 32. MNIST uses a width-256, eight-layer ReLU network, and PBMC3k uses a width-128, eight-layer ReLU network. Models are trained with Adam (Kingma & Ba, 2015), learning rate 10^{-3}, weight decay 10^{-4}, and at most 1,000 epochs. Early stopping uses a 20% validation partition of the training data and patience 100. Tabular models use full-batch training, while MNIST and PBMC3k use minibatches of 512 and 256, respectively.

We train each architecture using seeds 17, 29, and 43. For each trained predictor, we select 50 correctly classified held-out factual inputs and seek the opposite prediction under each of the three norms. All methods receive the same factual inputs for a given predictor and norm. The per-query limit is 60 seconds for tabular and PBMC3k experiments and 360 seconds for MNIST experiments. Every CPU heuristic runs on one thread per query; the exact MIP baselines use four Gurobi threads.

Table 4: Baseline implementation settings.

Method	Configuration
Wachter	10 restarts, learning rate 10^{-3}, 100 inner steps, λ-growth 5
DiCE	proximity weight 0.5, Adam (Kingma & Ba, 2015) learning rate 0.05, at most 5,000 iterations
NICE	target-side training pool, norm-specific nearest neighbor, grouped one-hot copying
Growing Spheres	1,000 samples per shell, shell width 0.2, 1,000 shells
PGD-bisect	10 restarts, 100 gradient steps, bisection tolerance 10^{-4}
FMN	1,000 steps, $\alpha : 1 \to 0.01$, $\gamma : 0.05 \to 0.001$
Greedy	coordinate step 0.1, at most 10,000 steps
FAB	Norm-specific box-constrained projections; 100 iterations, $\alpha_{\max} = 0.1$, $\eta = 1.05$, $\beta = 0.9$, 30 polish steps
DeepFool	norm-specific linearized steps; at most 50 iterations, overshoot 0.02, no box clipping
MIP-OBJ and MIP-EXP	tolerance 10^{-4}, 4 Gurobi threads

The CPU host has one AMD EPYC 7643 processor, two hardware threads per core, and 994 GiB of memory. CPU-run metadata records Julia 1.12.1. The GPU host provides an NVIDIA A100 with 40 GiB of memory, driver 590.48.01, and Julia 1.11.4.

B.5 Baseline configurations and common evaluation

Table 4 reports the implemented baseline settings. Methods receive the same coordinate bounds, except FAB, which is clipped only to the feature range, and DeepFool, which is unconstrained; neither receives the actionability constraints. Methods differ in whether they enforce one-hot assignments during search. Consequently, every returned output is evaluated using the common target-validity and actionability checks before inclusion in validity or proximity metrics.

Wachter and DiCE project categorical groups before evaluating output validity. NICE copies one-hot groups jointly from its target-side training instance. PGD-bisect and FMN apply categorical projection before final evaluation. Under ℓ_2, FMN returns no counterfactual on any of the 750 Adult queries; these queries count as failures. Growing Spheres and Greedy use coordinate bounds but do not consistently repair categorical assignments during their search. The mixed-integer formulations encode the applicable bounds and categorical constraints directly. MIP methods are evaluated only on tabular predictors of depths 2, 4, and 8.

FAB reimplementation. We use a repository reimplementation of FAB (Croce & Hein, 2020), not the authors' released code. At each iteration, it forms a local affine decision hyperplane and computes box-constrained projections under the requested ℓ_1, ℓ_2, or ℓ_∞ norm. The ℓ_1 projection uses a continuous-knapsack solution, while the scalar parameters for the ℓ_2 and ℓ_∞ projections are found by bisection. It then combines projections from the current point and factual, applies an overshoot and backward correction, and performs a final boundary search. We use at most 100 iterations, $\alpha_{\max} = 0.1$, overshoot $\eta = 1.05$, backward-step factor $\beta = 0.9$, and 30 final boundary-search steps. The implementation enforces box bounds but does not repair categorical or other tabular actionability constraints during its search; all outputs undergo the common final validity checks.

Norm-adapted PGD-bisect. Our PGD-bisect baseline combines a projected-gradient search at a fixed radius with bisection over that radius, following the projected-gradient attack of Madry et al. (2018). For each requested $p \in \{1, 2, \infty\}$, an inner restart searches the intersection $S_p(\epsilon) = \{x : \|x - x^0\|_p \le \epsilon,\ \ell \le x \le u\}$, where ℓ and u are feature-bound vectors. Its ascent direction is the signed gradient for $p = \infty$, the Euclidean-normalized gradient for $p = 2$, and a signed coordinate of largest gradient magnitude for $p = 1$. Each iteration proposes a step of length $2.5\epsilon/100$ in the corresponding norm, then takes the Euclidean projection onto $S_p(\epsilon)$. For $p = \infty$, this projection is coordinate-wise clipping. For $p = 1$, we use the clipped soft-threshold form of the exact ball-box projection given by Croce & Hein (2021); a scalar multiplier is found by bisection so that the projected perturbation has ℓ_1 norm at most ϵ. The $p = 2$ projection similarly clips a radially shrunk perturbation

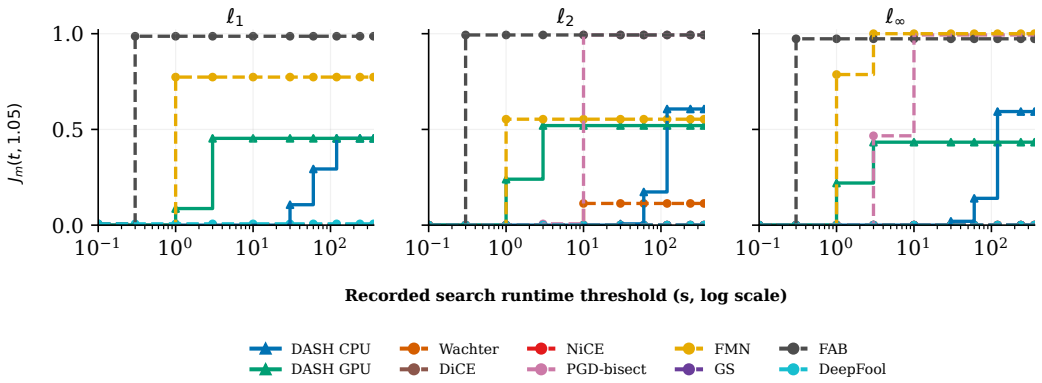

Figure 5: **MNIST joint quality-search-runtime profiles.** Columns show the ℓ_1, ℓ_2, and ℓ_∞ objectives. Each curve gives the fraction of all matched queries returning a valid counterfactual within 5% of the best heuristic-observed valid distance by the indicated search-time threshold.

and solves for its scalar multiplier. This is a norm adaptation of our PGD-bisect baseline, not a reproduction of the adaptive ℓ_1-APGD optimizer of Croce & Hein (2021). We use ten restarts and an outer-radius tolerance of 10^{-4}. Random starts are sampled from the requested norm ball and clipped to the feature bounds; categorical projection is applied only to the returned point before common verification.

Norm-adapted Growing Spheres. The original method searches Euclidean shells and then reduces unnecessary feature changes (Laugel et al., 2018). We retain those two phases while replacing the shell by $\{x : a \le \|x - x^0\|_p \le b\}$ for the requested $p \in \{1, 2, \infty\}$. Gaussian vectors normalized in ℓ_2 generate Euclidean directions, independently signed exponential variables normalized in ℓ_1 generate cross-polytope directions, and uniform-cube vectors normalized in ℓ_∞ generate cube directions. These constructions follow the probabilistic geometry of ℓ_p balls (Barthe et al., 2005). Conditional on a shell, the radius is sampled with R^d uniform on $[a^d, b^d]$, where d is the input dimension. We clip sampled points to the feature bounds, choose the target-valid point with smallest distance in the requested norm, and apply the same feature-reversion pass as in the Euclidean implementation. Clipping can move points inward and therefore changes the ideal uniform shell distribution near the domain boundary. The ℓ_1 and ℓ_∞ versions are our norm adaptations, not variants supplied by the original authors. All three use 1,000 samples per shell, shell width 0.2, and at most 1,000 shells; outputs receive the same target and actionability checks as every other method.

C Supplementary Analyses

C.1 MNIST scalability evaluation

We additionally evaluate the same search on MNIST (LeCun et al., 1998), using a width-256, eight-layer ReLU network. Pixels are scaled to $[0, 1]$, and bounded pixel changes are permitted. For each of three model seeds, we select 50 correctly classified held-out images, yielding 150 factual inputs and 450 query-norm cases per method. The per-query limit is 360 seconds. Figure 5 reports joint valid-within-5% quality-search-runtime profiles for the CPU and GPU implementations and available baselines. At the final threshold, FAB reaches 98.7%, 99.3%, and 97.3% under ℓ_1, ℓ_2, and ℓ_∞, respectively, while DASH CPU reaches 45.3%, 60.7%, and 59.3%. Native ℓ_1 PGD-bisect and Growing Spheres runs do not cover the full matched query set and are omitted from that panel. The MNIST setting tests scaling on a larger input space without the tabular actionability constraints of the primary evaluation.

C.2 CERTIFIED-DISTANCE GAP

For queries on which a mixed-integer formulation certifies an optimum $d_\nu^\star$, we report the relative gap of method m as

$$g_m = 100 \left(\frac{d_\nu(x_m, x^0)}{d_\nu^\star} - 1 \right).$$

A method that returns no valid counterfactual is assigned $g_m = \infty$. The certified-gap curves therefore retain unsuccessful returns in their denominators and should be interpreted only over queries for which the exact formulation completed with an optimum certificate.

D FULL BENCHMARK RESULTS

Protocol. Four tabular datasets (Adult, COMPAS, GMC, HELOC), fully connected ReLU networks of width 64 and depth $\{2, 4, 8, 16, 32\}$, three model seeds (17, 29, 43), 50 held-out factuals per (dataset, depth, seed) cell, the same factuals for every arm, and a 60 s budget per factual on one CPU thread. Every arm except FAB and DeepFool receives the same basic actionability box (immutable, non-decreasing and non-increasing coordinates); validity is checked identically for every arm (see Metrics). DASH (beam 16, 10 iterations) and every baseline except Greedy* search the objective norm of the table; Greedy* is norm-agnostic and is reported under every norm. *DASH + data seed* is an ablation that also generates anchors by copying single features, or one-hot groups, from the nearest training point already on the target side, as NICE does. It leaves the rate of Table 5 unchanged (94.6% for both) at 11% more search time (median 0.068 vs 0.061 s), so DASH is reported without it, and the ablation enters no best heuristic-observed reference. MNIST (3 vs 8, $n = 784$, depth 8, width 256, no actionability constraints, 360 s) uses the same seeds and 50 factuals; Greedy* did not run there, and under ℓ_1 GS* and PGD-bisect* cover 70 of the 150 factuals. PBMC3k is reported in the main text only: its numbers there include the marker completion of Appendix B.3, which these tables do not apply. The exact references MIP-EXP (bounded-distance feasibility queries with radius refinement) and MIP-OBJ (direct distance minimization) use Gurobi (Gurobi Optimization, LLC, 2024) with 4 threads, a 60 s budget per factual and an absolute optimality gap of 10^{-4}; they were run at depths 2, 4 and 8 on the same factuals. In the per-cell tables their medians are over the instances they solve, which at depth 8 is a small and easy subset, so they are left out of the depth figures and compared through Tables 9 (certified optima) and 10 (all queries) instead. MIP-EXP proved four factual-model pairs infeasible (no point in the actionability box flips the label: one COMPAS and one GMC factual at depth 2, two GMC factuals at depth 8; the same four under each norm, 12 of the $3 \times 3{,}000$ queries). No arm returned a valid counterfactual on them; they are kept in every count below and count as a failure for every arm (0.13% of the queries), so that every denominator is the full query set.

Baselines marked with an asterisk. All baselines are our own Julia reimplementations. DiCE (Mothilal et al., 2020) and NICE (Brughmans et al., 2024) handle categorical features in their original formulation and are used as published; for Wachter (Wachter et al., 2018), whose original formulation enumerates a single binary attribute, we project categorical groups as described in Appendix B.5. GS*, PGD-bisect* and FMN* are continuous methods (Growing Spheres, Laugel et al., 2018; the PGD attack of Madry et al., 2018; the fast minimum-norm attack of Pintor et al., 2021), run in the variant that searches the objective norm of the table (Appendix B.5); PGD-bisect* wraps the attack in a bisection on the perturbation radius, since PGD alone has no notion of a minimal perturbation. PGD-bisect* and FMN* receive a one-hot repair that projects the returned point onto the nearest legal category assignment (or one alternative category per group) and re-checks the label; GS* and Greedy* receive none. Greedy* is the single-feature saliency-map search of Schut et al. (2021), an extension of JSMA (Papernot et al., 2016), which originally perturbs feature pairs, run on a single predictor rather than their ensemble. Scoring the un-repaired outputs instead changes the within-5% rates of Table 5 by at most 0.4 points (PGD-bisect* 40.7 vs 41.1, FMN* 30.5 vs 30.7); the raw outputs are stored alongside the repaired ones in every results file. FAB (Croce & Hein, 2020) and DeepFool (Moosavi-Dezfooli et al., 2016) run unmodified, in their ℓ_1, ℓ_2 and ℓ_∞ variants, and carry no asterisk: FAB is clipped to the range of each feature and DeepFool is unconstrained; neither receives the actionability box or a one-hot repair, and the common validity check counts the resulting violations as failures. The main text omits the asterisks.

Metrics. Success is the fraction of factuals for which the arm returns a point that crosses the threshold with margin 10^{-4} (CARLA's success rate); validity additionally requires the box and one-hot constraints to hold, and is the column every other metric conditions on. The gap between the two is the arm's constraint handling: GS* flips the label on every Adult and COMPAS query but is valid on only 45.6% and 52.0% of them (Greedy*: 91.3% and 87.1%), and FAB and DeepFool flip it on at least 90% of the queries of every tabular dataset but are valid on only 1.5-29.6% of the Adult, COMPAS and GMC queries. Every other column is a median over the valid rows of the cell (NA = no valid counterfactual). ℓ_0 counts changed features (a one-hot switch counts as one); redundancy is the number of changed features that can be reverted without losing validity (Pawelczyk et al., 2021); yNN is the fraction of the 5 nearest training points (Euclidean) that the model assigns to the counterfactual's class; yNN-dist is their mean distance; time is wall-clock per factual. Bold marks the best value per column among arms with validity $\geq 80\%$ (a threshold we adopt to exclude arms with low validity); distances within 10^{-3} of the best share the bold. Wins/ties/losses are paired per factual over instances valid for both arms with the same 10^{-3} tie band; the Wilcoxon signed-rank test excludes ties.

Summary across objectives. Table 5 is the statistic the abstract and the introduction quote (Section 4.2 quotes its per-norm counterparts, Tables 11, 19 and 27); Table 6 pools the same 9,000 tabular queries into one distribution per arm: the ratio of each returned distance to the best that DASH or any of the nine prior methods found on that query, with failures counted as infinite. Where FAB or FMN* is valid it is often as close as DASH or closer: on the instances valid for both, FAB is closer than DASH more often than the reverse under every norm, and FMN* under ℓ_1 and ℓ_2 (Tables 12, 20 and 28); their shortfall in Table 5 is validity (33.1% and 45.9%, against 99.4% for DASH).

Table 5: Headline statistic, all three objectives pooled, tabular datasets (Adult, COMPAS, GMC, HELOC, depths 2–32, 9,000 queries), no query excluded: share of queries on which the arm returns a valid counterfactual within 5% of the smallest valid distance found by DASH or any of the nine prior methods, and the arm's median search time over all its queries. These are the numbers the abstract and the introduction quote; Section 4.2 quotes the per-norm rates of Tables 11, 19 and 27. 95% CI = percentile cluster bootstrap over trained models (dataset × depth × seed), 2,000 resamples.

method	queries	valid %	within 5 %	95 % CI	median time (s)
DASH	9000	99.4	**94.6**	93.0–96.1	0.061
PGD-bisect*	9000	93.6	41.1	33.5–48.6	0.298
Wachter	9000	93.0	37.5	31.8–43.5	1.583
FAB	9000	33.1	32.1	23.2–41.6	0.014
FMN*	9000	45.9	30.7	24.9–36.9	0.156
NICE	9000	98.9	14.6	10.4–19.0	0.023
DiCE	9000	94.5	8.0	6.0–10.0	0.001
GS*	9000	74.1	8.0	5.6–10.8	0.014
DeepFool	9000	10.8	5.2	3.4–7.2	0.000
Greedy*	9000	93.6	4.8	3.8–5.7	0.001

Table 6: Distance relative to the best-known counterfactual over every query of the benchmark (9000 queries: adult, compas, gmc, heloc, all three objectives, all depths): per query, the arm's distance divided by the smallest valid distance among DASH and the nine prior methods; an invalid answer, or an arm that did not run on that query, counts as infinite (rows = the arm's actual rows). Quantiles are over all queries; within x = share of queries with ratio $\leq 1 + x$ (2x, 10x = ratio ≤ 2, ≤ 10).

method	rows	queries	valid %	median ratio	75th pct	90th pct	within 1 %	within 5 %	within 25 %	within 2x	within 10x
DASH	9000	9000	99.4	1.00	1.00	1.02	84.5	94.6	98.8	99.1	99.4
PGD-bisect*	9000	9000	93.6	1.55	5.91	43.45	36.5	41.1	46.2	56.1	80.7
Wachter	9000	9000	93.0	1.19	4.84	18.58	28.6	37.5	53.3	62.8	86.1
FAB	9000	9000	33.1	fail	fail	fail	29.2	32.1	32.9	33.1	33.1
FMN*	9000	9000	45.9	fail	fail	fail	28.6	30.7	31.3	32.1	40.9
NICE	9000	9000	98.9	1.49	3.06	6.33	4.6	14.6	36.1	63.6	95.1
GS*	9000	9000	74.1	1.63	fail	fail	1.3	8.0	29.9	57.6	72.4
DiCE	9000	9000	94.5	1.44	2.15	8.86	1.6	8.0	33.1	72.8	90.4
DeepFool	9000	9000	10.8	fail	fail	fail	2.8	5.2	9.3	10.7	10.8
Greedy*	9000	9000	93.6	1.94	5.26	22.20	0.9	4.8	23.2	51.6	84.8

D.1 FACTUAL-ORIGIN ANCHOR ABLATION

Table 7 removes the factual-origin anchors from DASH and keeps every other component (one-directional anchors, including categorical transitions; local gradient-based anchors; ranking; beam selection; and polishing) unchanged. Validity is unaffected (100% for both arms), but the within-5% rate drops from 95.6% to 90.1%: removing factual-origin anchors gives a worse distance on 180 queries and a better one on 55 (Wilcoxon signed-rank $p < 10^{-19}$). The effect is concentrated where the search must move several continuous features jointly: under ℓ_2 and ℓ_∞ and on HELOC and Adult. Under ℓ_1 a single-feature move is typically optimal, and the one-directional anchors already find it. Factual-origin anchors cost about 2% more median runtime. The ablation ran on an Apple M-series laptop (one thread per query), so runtimes are compared only within the table; on the same 900 queries its DASH arm matches the cluster run's validity on every query and its distance on 823, and floating-point differences between the machines change the search path on the other 77, moving the distance by at most 6% in either direction.

Table 7: Factual-origin anchor ablation: DASH against the same search without factual-origin anchors (all other components unchanged), tabular datasets, depths 2-32, seed 17, the first 15 factuals of each model (900 queries per arm), paper settings; both arms return a valid counterfactual on every query. Within 5% = valid and within 5% of the best-known distance of Table 5 (DASH and the nine prior methods, full paper run). B/T/L = queries where removing the factual-origin anchors gives a better/tied/worse distance (relative tie band 10^{-3}); 90th = 90th percentile of the distance ratio (no factual-origin anchors / DASH). Both arms ran on the same machine, each in its own process, so their runtimes are comparable with each other but not with the other tables.

	queries	DASH	no factual-origin	B/T/L	90th	median s: DASH	median s: no factual-origin anchors
ℓ_1	300	97.3	97.7	6/290/4	1.000	0.095	0.092
ℓ_2	300	96.3	88.7	25/181/94	1.034	0.110	0.105
ℓ_∞	300	93.0	84.0	24/194/82	1.051	0.117	0.113
Adult	225	100.0	90.2	15/157/53	1.035	0.159	0.152
COMPAS	225	99.6	100.0	5/209/11	1.000	0.052	0.048
GMC	225	90.2	92.0	17/194/14	1.000	0.057	0.054
HELOC	225	92.4	78.2	18/105/102	1.079	0.149	0.147
all	900	95.6	90.1	55/665/180	1.022	0.109	0.107

D.2 ℓ_1 OBJECTIVE (HEADLINE)

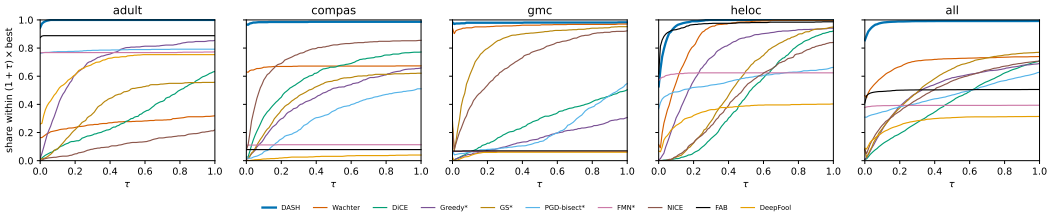

Figure 6: Performance profile, ℓ_1 objective: share of the tabular queries (750 per dataset panel, 3,000 in the pooled panel) where the arm is valid and within $(1+\tau)\times$ the best heuristic-observed valid distance of Table 11 ($\tau = \alpha - 1$ in the notation of Section 4.2; at $\tau = 0$ the 10^{-3} tie band applies), as τ ranges from 0 (the heuristic reference) to 1 (twice that reference). The right end of a curve cannot exceed the arm's validity. The reference excludes the exact solvers; Table 9 compares with the certified optimum where it exists.

Table 8: CARLA axes at depth 8 (ℓ_1 objective, basic actionability; seeds 17/29/43).

dataset	method	success %	valid %	ℓ_1	l0	redund.	yNN	yNN-dist	time (s)
adult	DASH	**100**	**100**	0.694	**1.0**	**0.0**	0.000	1.004	0.0873
	DASH + data seed	**100**	**100**	0.694	**1.0**	**0.0**	0.000	1.004	0.0940
	Wachter	83	78	2.432	5.0	2.0	0.000	1.438	2.0056
	DiCE	97	95	1.074	4.0	**0.0**	0.000	**0.872**	0.0021
	Greedy*	**100**	82	**0.600**	**1.0**	**0.0**	0.000	0.933	**0.0012**
	GS*	100	37	0.584	1.0	0.0	0.000	0.634	0.1199
	PGD-bisect*	70	70	0.478	1.0	0.0	0.000	0.773	0.4647
	FMN*	67	67	0.478	1.0	0.0	0.000	0.769	0.1431
	NICE	**100**	**100**	2.565	3.0	**0.0**	**0.200**	0.972	0.0248
	FAB	100	78	0.514	1.0	0.0	0.000	0.913	0.0151
	DeepFool	100	60	0.460	1.0	0.0	0.000	0.715	0.0003
	MIP-EXP	2	2	0.017	1.0	0.0	0.200	0.067	7.6279
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA
compas	DASH	**98**	**98**	**0.144**	**1.0**	**0.0**	0.400	0.027	0.0388
	DASH + data seed	**98**	**98**	**0.144**	**1.0**	**0.0**	0.400	0.027	0.0407
	Wachter	91	91	**0.144**	2.0	1.0	0.600	**0.027**	1.7761
	DiCE	94	94	0.149	**1.0**	**0.0**	0.800	**0.026**	0.0009
	Greedy*	97	85	0.200	**1.0**	**0.0**	**1.000**	0.029	**0.0004**
	GS*	100	59	0.101	1.0	0.0	0.600	0.025	0.0101
	PGD-bisect*	95	95	0.432	2.0	1.0	0.800	0.032	0.2912
	FMN*	7	7	0.389	1.0	0.0	0.600	0.070	0.1478
	NICE	**98**	**98**	0.154	**1.0**	**0.0**	0.600	**0.026**	0.0075
	FAB	99	7	0.006	1.0	0.0	0.000	0.005	0.0138
	DeepFool	100	10	0.322	1.0	0.0	0.800	0.040	0.0002
	MIP-EXP	45	45	0.049	1.0	0.0	0.600	0.015	5.8384
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA
gmc	DASH	**99**	**99**	0.039	**1.0**	**0.0**	0.800	0.049	0.0478
	DASH + data seed	**99**	**99**	0.039	**1.0**	**0.0**	0.800	0.049	0.0543
	Wachter	97	97	0.038	3.0	1.0	0.800	0.049	1.3670
	DiCE	91	91	0.050	**1.0**	**0.0**	0.800	0.056	0.0004
	Greedy*	91	91	0.100	**1.0**	**0.0**	**1.000**	0.095	**0.0003**
	GS*	**99**	**99**	0.040	**1.0**	**0.0**	0.800	0.050	0.0097
	PGD-bisect*	97	97	0.087	4.0	2.0	0.800	0.055	0.1687
	FMN*	1	1	0.072	1.0	0.0	1.000	0.077	0.1489
	NICE	98	98	0.048	**1.0**	**0.0**	0.800	0.050	0.7981
	FAB	100	1	0.005	1.0	0.0	0.900	0.052	0.0139
	DeepFool	100	0	NA	NA	NA	NA	NA	NA
	MIP-EXP	87	87	**0.031**	**1.0**	**0.0**	0.800	**0.044**	4.4221
	MIP-OBJ	2	2	0.389	1.0	0.0	0.000	0.391	1.6557
heloc	DASH	**100**	**100**	0.436	**2.0**	**0.0**	0.200	0.416	0.2130
	DASH + data seed	**100**	**100**	0.430	**2.0**	**0.0**	0.200	0.411	0.2203
	Wachter	**100**	**100**	0.485	20.0	9.5	0.200	0.318	1.8254
	DiCE	99	99	0.699	9.0	**0.0**	0.200	0.317	**0.0005**
	Greedy*	99	99	0.500	3.0	**0.0**	0.200	0.384	0.0010
	GS*	**100**	**100**	0.582	6.0	**0.0**	0.200	0.360	0.0159
	PGD-bisect*	**100**	**100**	0.538	6.0	1.5	0.200	0.401	0.3132
	FMN*	67	67	0.372	3.0	0.0	0.200	0.367	0.1556
	NICE	**100**	**100**	0.634	6.0	1.0	**0.400**	**0.287**	0.0115
	FAB	98	98	**0.418**	3.0	**0.0**	0.200	0.371	0.0137
	DeepFool	99	35	0.305	1.0	0.0	0.400	0.359	0.0003
	MIP-EXP	7	7	0.032	1.0	0.0	0.400	0.222	11.0365
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA

success = the returned point crosses the threshold (CARLA's success rate); valid = it also satisfies the actionability box and one-hot constraints. Bold = best per column within a dataset among arms with validity $\geq$ 80%. Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 12) and the best-known table (Table 11).

Table 9: ℓ_1 relative to the certified optimum (MIP-EXP or MIP-OBJ; Gurobi, 4 threads, 60 s budget per factual, absolute optimality gap 10^{-4}, depths 2/4/8, same 50 factuals per cell as the main run). certified n = instances on which either exact formulation proved an optimum in that dataset; arm valid = how many of those the arm returns a valid point on; at optimum = arm valid and distance $\leq$ ref $+10^{-3}$ (one-sided: below the solver's incumbent is inside its gap); within 5% = arm valid and distance $\leq 1.05\times$ref (purely relative: where the optimum is below 0.02 the absolute tie band is the wider of the two, so the columns are not nested); both percentages over certified n; ratio columns over the arm-valid subset. A valid point can lie up to 6×10^{-4} below the certified optimum (the solver's absolute gap and numerical tolerances), so ratios below 1 occur where the optimum itself is small: every ratio below 0.95 has an optimum under 2.5×10^{-3}.

dataset	arm	certified n	arm valid	at optimum %	within 5 %	median ratio	max ratio
adult	DASH	187	187	100.0	100.0	1.000	1.001
adult	DASH + data seed	187	187	100.0	100.0	1.000	1.001
adult	Wachter	187	182	41.7	42.8	2.567	12.574
adult	DiCE	187	187	0.5	1.1	2.028	62.500
adult	Greedy*	187	187	1.1	10.2	1.260	125.000
adult	GS*	187	141	0.5	6.4	1.261	2.142
adult	PGD-bisect*	187	187	100.0	100.0	1.000	1.007
adult	FMN*	187	173	92.5	92.5	1.000	1.000
adult	NICE	187	187	0.0	1.6	5.690	548.469
adult	FAB	187	187	100.0	100.0	1.000	1.000
adult	DeepFool	187	186	55.6	82.4	1.011	1.531
compas	DASH	313	313	97.4	99.7	1.000	1.065
compas	DASH + data seed	313	313	97.1	99.7	1.000	1.065
compas	Wachter	313	302	84.3	85.9	1.000	86.976
compas	DiCE	313	311	1.9	15.0	1.269	166.667
compas	Greedy*	313	305	0.6	5.1	1.514	333.333
compas	GS*	313	237	3.2	9.6	1.211	23.564
compas	PGD-bisect*	313	308	2.2	2.6	1.510	2000.933
compas	FMN*	313	1	0.3	0.3	0.979	0.979
compas	NICE	313	311	6.7	35.8	1.079	9.149
compas	FAB	313	36	11.5	11.5	1.000	1.000
compas	DeepFool	313	2	0.3	0.3	1.049	1.080
gmc	DASH	413	413	99.3	99.0	1.000	1.148
gmc	DASH + data seed	413	413	99.3	99.0	1.000	1.148
gmc	Wachter	413	413	97.6	97.1	1.000	1.973
gmc	DiCE	413	406	0.5	2.7	2.240	521.374
gmc	Greedy*	413	407	0.2	1.9	4.367	1042.749
gmc	GS*	413	413	32.7	37.5	1.078	27.974
gmc	PGD-bisect*	413	406	12.1	9.2	1.913	15.053
gmc	FMN*	413	45	10.9	10.9	1.000	1.002
gmc	NICE	413	406	29.8	27.4	1.129	24.263
gmc	FAB	413	45	10.9	10.9	1.000	1.000
gmc	DeepFool	413	46	4.4	2.9	1.092	2.278
heloc	DASH	158	158	70.9	96.2	1.001	1.086
heloc	DASH + data seed	158	158	71.5	96.2	1.001	1.086
heloc	Wachter	158	158	24.7	46.2	1.062	1.413
heloc	DiCE	158	158	0.0	0.6	1.518	33.333
heloc	Greedy*	158	158	0.0	15.2	1.182	66.667
heloc	GS*	158	158	0.0	1.3	1.449	9.312
heloc	PGD-bisect*	158	158	55.7	60.8	1.001	6.897
heloc	FMN*	158	100	62.0	63.3	1.000	1.034
heloc	NICE	158	158	0.0	0.0	1.604	14.781
heloc	FAB	158	158	79.7	98.7	1.000	1.106
heloc	DeepFool	158	100	19.0	42.4	1.028	1.341
all	DASH	1071	1071	94.7	99.0	1.000	1.148
all	DASH + data seed	1071	1071	94.7	99.0	1.000	1.148
all	Wachter	1071	1055	73.2	76.8	1.000	86.976
all	DiCE	1071	1062	0.8	5.7	1.699	521.374
all	Greedy*	1071	1057	0.5	6.3	1.705	1042.749
all	GS*	1071	949	13.6	18.6	1.178	27.974
all	PGD-bisect*	1071	1059	31.0	30.7	1.512	2000.933
all	FMN*	1071	319	29.6	29.8	1.000	1.034
all	NICE	1071	1062	13.4	21.3	1.249	548.469
all	FAB	1071	426	36.8	39.6	1.000	1.106
all	DeepFool	1071	334	14.3	21.8	1.022	2.278

Table 10: MIP-EXP against DASH by depth, ℓ_1 objective, same queries, pooled over datasets and seeds. Median times are over all of these queries, solved or not (MIP-EXP stops at its 60 s budget).

depth	queries	MIP-EXP valid %	MIP-EXP certified %	MIP-EXP median time (s)	DASH valid %	DASH median time (s)
2	600	99.7	94.5	1.977	99.7	0.031
4	600	81.0	46.8	60.000	99.5	0.051
8	600	35.5	31.5	60.000	99.2	0.073

Table 11: Best-known counterfactual, ℓ_1 objective: population is every tabular query of this norm (the queries column); best = smallest valid distance among DASH and the nine prior methods on that query (the exact solvers and the ablation are not in the reference); a query counts for an arm at τ iff the arm is valid and its distance $\leq (1+\tau)\times$best; column best uses the absolute tie band best $+10^{-3}$ instead, which is wider than 1% where best is below 0.1; invalid rows never count, no penalty constant; 95% CI = percentile cluster bootstrap over trained models (dataset $\times$ depth $\times$ seed), 2,000 resamples. Sorted by within 5%; bold = best per column.

method	queries	valid %	best	within 1 %	within 5 %	95 % CI (within 5 %)	within 10 %
DASH	3000	**99.4**	**85.6**	**88.9**	**95.7**	93.7–97.5	**98.0**
Wachter	3000	93.6	46.3	44.6	53.6	45.8–61.0	59.5
FAB	3000	50.6	40.9	46.1	48.6	37.5–59.9	49.2
FMN*	3000	40.1	37.8	38.2	38.8	29.2–48.0	39.1
PGD-bisect*	3000	92.3	30.9	30.9	32.8	24.1–42.4	34.7
NICE	3000	98.9	7.9	5.0	16.1	12.1–20.2	25.2
DeepFool	3000	31.6	8.9	8.2	15.3	9.9–21.1	19.9
GS*	3000	79.8	7.9	2.4	12.8	9.0–16.9	22.7
Greedy*	3000	93.6	0.9	2.2	11.1	8.6–13.7	22.1
DiCE	3000	94.9	0.9	1.2	6.7	4.6–9.0	11.8

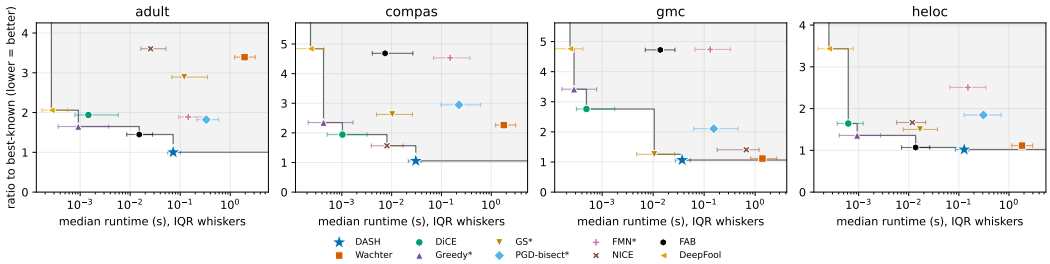

Figure 7: Runtime-quality trade-off, ℓ_1 objective, tabular datasets, all depths and seeds pooled. Horizontal axis: median runtime per query (25th-75th percentile whiskers). Vertical axis: mean failure-penalized ratio to the best heuristic-observed valid distance - a valid counterfactual scores its distance divided by the smallest valid distance DASH or any of the nine prior methods found on that query (the reference of Table 11); a failure, or a ratio above 5, scores 5. The staircase is the Pareto front, drawn to the axis edges; the shaded region is dominated.

Table 12: DASH against each arm on ℓ_1, paired per factual over the instances where both are valid: DASH wins = DASH is closer by more than 10^{-3}, losses = the other arm is closer; cells won = (depth, seed) cells ("all" row: dataset, depth, seed cells) with ≥ 1 pair, where the median paired difference favours DASH by more than 10^{-3}, printed as k/m; median ratio = median of DASH distance / other distance (below 1 favours DASH); two-sided Wilcoxon signed-rank on the non-tied pairs. Pairs within a cell share one trained model, so the pooled p-value overstates the evidence; cells won counts model-level agreement.

dataset	vs	n paired	DASH wins	ties	DASH losses	cells won	median ratio	Wilcoxon p
adult	DASH + data seed	750	9	732	9	0/15	1.000	0.52
adult	Wachter	654	531	123	0	15/15	0.296	1.1e-88
adult	DiCE	717	708	2	7	15/15	0.563	6.3e-118
adult	Greedy*	685	665	9	11	15/15	0.892	2.8e-109
adult	GS*	426	420	6	0	15/15	0.796	1.5e-70
adult	PGD-bisect*	604	31	572	1	0/15	1.000	1.4e-09
adult	FMN*	588	13	563	12	0/15	1.000	0.026
adult	NICE	750	748	2	0	15/15	0.249	4.2e-124
adult	FAB	666	7	629	30	0/15	1.000	1.4e-05
adult	DeepFool	565	369	196	0	11/15	0.966	3.2e-62
adult	MIP-EXP	303	109	194	0	3/8	1.000	1.3e-19
adult	MIP-OBJ	299	127	172	0	3/6	1.000	1.4e-22
compas	DASH + data seed	740	1	738	1	0/15	1.000	NA
compas	Wachter	669	194	467	8	0/15	1.000	4.3e-34
compas	DiCE	668	653	15	0	15/15	0.840	1.3e-108
compas	Greedy*	653	646	7	0	15/15	0.727	1.9e-107
compas	GS*	477	456	20	1	15/15	0.850	2.5e-75
compas	PGD-bisect*	685	664	20	1	15/15	0.570	1.7e-110
compas	FMN*	99	16	79	4	0/9	1.000	8.2e-05
compas	NICE	728	671	52	5	15/15	0.934	6.4e-109
compas	FAB	59	0	58	1	0/15	1.001	NA
compas	DeepFool	33	32	1	0	7/7	0.807	4.7e-10
compas	MIP-EXP	359	44	306	9	0/9	1.000	1.4e-07
compas	MIP-OBJ	292	62	223	7	1/6	1.000	1.8e-12
gmc	DASH + data seed	742	0	742	0	0/15	1.000	NA
gmc	Wachter	736	30	703	3	0/15	1.000	1.1e-05
gmc	DiCE	717	712	5	0	15/15	0.515	3.2e-118
gmc	Greedy*	720	714	3	3	15/15	0.294	6.5e-116
gmc	GS*	742	530	209	3	15/15	0.933	2.8e-85
gmc	PGD-bisect*	730	672	58	0	14/15	0.518	1.1e-111
gmc	FMN*	49	0	49	0	0/6	1.000	NA
gmc	NICE	735	558	177	0	15/15	0.863	4.3e-93
gmc	FAB	52	1	51	0	0/9	1.000	NA
gmc	DeepFool	46	28	18	0	2/3	0.916	7.5e-09
gmc	MIP-EXP	430	14	413	3	0/9	0.999	0.19
gmc	MIP-OBJ	297	1	293	3	0/8	1.000	NA
heloc	DASH + data seed	750	65	577	108	0/15	1.000	0.0016
heloc	Wachter	750	576	84	90	15/15	0.937	6.1e-79
heloc	DiCE	746	744	2	0	15/15	0.665	1.9e-123
heloc	Greedy*	749	706	4	39	15/15	0.872	3.4e-114
heloc	GS*	750	746	0	4	15/15	0.715	3.4e-124
heloc	PGD-bisect*	750	425	197	128	8/15	0.887	7.2e-67
heloc	FMN*	468	29	224	215	0/15	1.002	1.5e-21
heloc	NICE	750	748	0	2	15/15	0.689	2e-124
heloc	FAB	741	141	316	284	0/15	1.001	8.3e-08
heloc	DeepFool	303	224	56	23	14/15	0.956	3.6e-33
heloc	MIP-EXP	205	43	116	46	3/9	1.000	0.037
heloc	MIP-OBJ	299	108	118	73	3/6	1.000	1.6e-13
all	DASH + data seed	2982	75	2789	118	0/60	1.000	0.0043
all	Wachter	2809	1331	1377	101	30/60	0.992	6.9e-212
all	DiCE	2848	2817	24	7	60/60	0.641	$< 10^{-300}$
all	Greedy*	2807	2731	23	53	60/60	0.788	$< 10^{-300}$
all	GS*	2395	2152	235	8	60/60	0.827	$< 10^{-300}$
all	PGD-bisect*	2769	1792	847	130	37/60	0.674	6e-282
all	FMN*	1204	58	915	231	0/45	1.000	5.9e-09
all	NICE	2963	2725	231	7	60/60	0.732	$< 10^{-300}$
all	FAB	1518	149	1054	315	0/54	1.000	3.7e-10
all	DeepFool	947	653	271	23	34/40	0.953	1.1e-103
all	MIP-EXP	1297	210	1029	58	6/35	1.000	6.9e-27
all	MIP-OBJ	1187	298	806	83	7/26	1.000	3.2e-51

Table 13: Validity (%) per arm and cell, ℓ_1 objective.

dataset	depth	DASH	DASH + data seed	Wachter	DiCE	Greedy*	GS*	PGD-bisect*	FMN*	NICE	FAB	DeepFool	MIP-EXP	MIP-OBJ
adult	2	100	100	97	100	100	73	100	91	100	100	99	100	100
	4	100	100	90	100	100	79	100	92	100	100	95	100	99
	8	100	100	78	95	82	37	70	67	100	78	60	2	0
	16	100	100	87	95	91	66	84	89	100	91	84	–	–
	32	100	100	83	88	83	29	49	53	100	75	39	–	–
compas	2	99	99	93	98	97	77	96	0	98	9	1	99	99
	4	98	98	90	97	97	67	91	0	95	7	0	95	95
	8	98	98	91	94	85	59	95	7	98	7	10	45	0
	16	98	98	92	86	79	59	89	23	97	8	10	–	–
	32	100	100	80	70	77	56	87	35	99	7	1	–	–
gmc	2	99	99	99	99	99	99	98	28	97	27	30	99	98
	4	100	100	99	97	99	100	97	1	98	2	1	100	98
	8	99	99	97	91	91	99	97	1	98	1	0	87	2
	16	99	99	99	98	99	99	98	3	99	2	0	–	–
	32	98	98	96	92	93	98	96	0	98	3	0	–	–
heloc	2	100	100	100	100	100	100	100	64	100	100	59	100	99
	4	100	100	100	100	100	100	100	36	100	100	27	29	100
	8	100	100	100	100	99	99	100	100	67	100	98	35	7	0
	16	100	100	100	100	98	100	100	100	77	100	96	35	–	–
	32	100	100	100	100	100	100	100	100	68	100	100	47	–	–

Table 14: MNIST (3 vs 8, $n = 784$, depth 8, width 256, 150 factuals), ℓ_1 objective. Greedy* did not run at this scale; under ℓ_1 the norm-adapted GS* and PGD-bisect* runs cover 70 of the 150 factuals (validity over the arm's own rows). PBMC3k is reported in the main text, after the shared marker completion. DASH = the CPU implementation; NA = no valid counterfactual, or a metric the run did not record.

dataset	method	success %	valid %	ℓ_1	l0	redund.	yNN	yNN-dist	time (s)
mnist	DASH	**100**	**100**	8.890	**11.0**	**0.0**	0.000	5.583	61.8500
mnist	DASH + data seed	**100**	**100**	8.913	**11.0**	**0.0**	0.000	5.583	63.4336
mnist	Wachter	**100**	**100**	22.274	607.0	274.0	0.000	**5.136**	3.4466
mnist	DiCE	85	85	17.454	177.0	**0.0**	0.000	5.159	**0.0436**
mnist	GS*	57	57	15.816	34.5	17.5	0.000	5.695	27.8042
mnist	PGD-bisect*	23	23	173.9	462.0	459.0	0.000	11.548	3.0613
mnist	FMN*	77	77	8.135	12.0	0.0	0.000	5.742	0.8633
mnist	NICE	**100**	**100**	11.790	22.0	6.0	**0.200**	5.647	0.7475
mnist	FAB	**100**	**100**	**7.882**	16.0	**0.0**	0.000	5.688	0.1730
mnist	DeepFool	100	1	0.034	1.0	0.0	0.000	5.614	0.0011

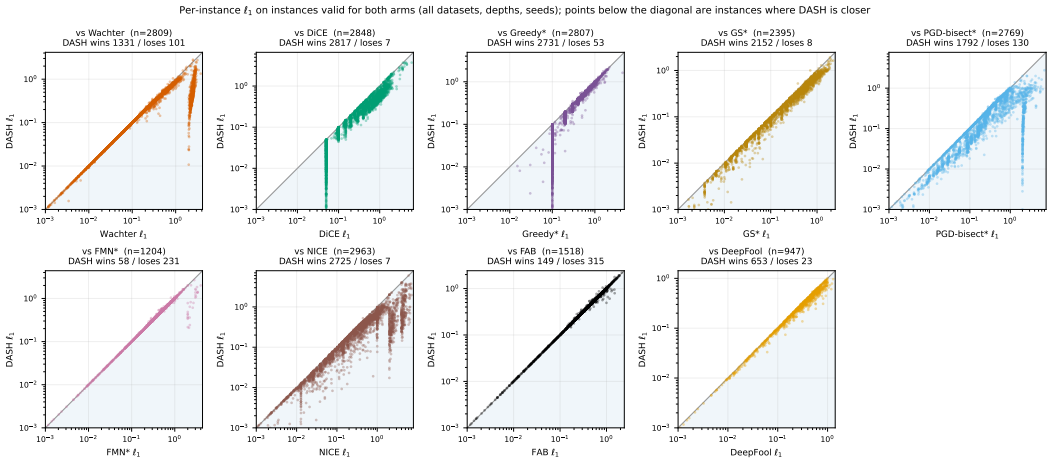

Figure 8: The paired comparison behind Table 12: for each baseline, one point per instance valid for both arms, DASH's ℓ_1 against the baseline's (log-log). Points below the diagonal are instances where DASH is closer; the vertical streaks are instances where the baseline returns the same distance for many factuals (a fixed step size or a one-hot flip) while DASH finds a closer point.

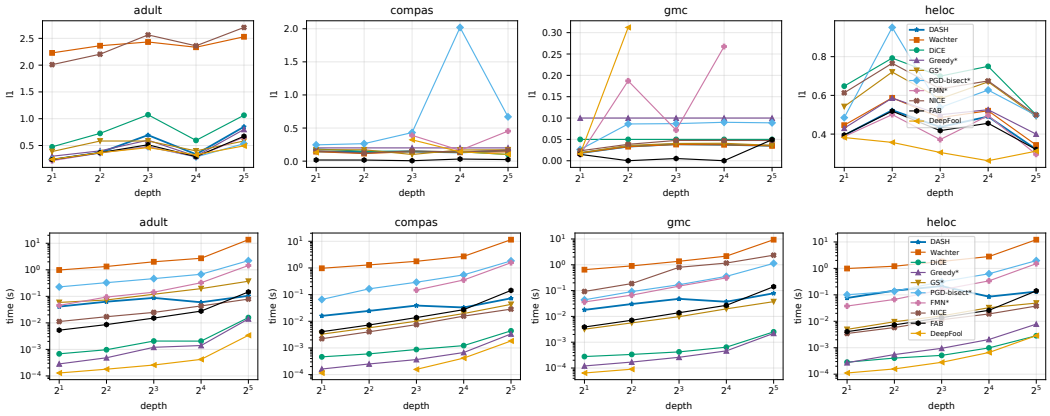

Figure 9: Median ℓ_1 (top) and median runtime (bottom, log scale) of the valid counterfactuals per arm as network depth grows, ℓ_1 objective. DASH + data seed is omitted (it overlaps DASH).

Table 15: All tabular cells, ℓ_1 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 12) and the best-known table (Table 11).

dataset	depth	method	success %	valid %	ℓ_1	l0	redund.	yNN	yNN-dist	time (s)	ℓ_2	ℓ_∞
adult	2	DASH	**100**	**100**	0.237	**1.0**	**0.0**	**0.000**	0.344	0.0411	0.237	0.237
adult	2	DASH + data seed	**100**	**100**	0.237	**1.0**	**0.0**	**0.000**	0.344	0.0462	0.237	0.237
adult	2	Wachter	**100**	97	2.229	4.0	1.0	**0.000**	1.426	0.9827	1.424	1.000
adult	2	DiCE	**100**	**100**	0.475	3.0	**0.0**	**0.000**	0.401	0.0007	0.293	**0.190**
adult	2	Greedy*	**100**	**100**	0.300	**1.0**	**0.0**	**0.000**	0.400	0.0003	0.300	0.300
adult	2	GS*	100	73	0.388	1.0	0.0	0.000	0.417	0.0589	0.388	0.388
adult	2	PGD-bisect*	**100**	**100**	0.237	**1.0**	**0.0**	**0.000**	0.344	0.2265	0.237	0.237
adult	2	FMN*	91	91	**0.216**	**1.0**	**0.0**	**0.000**	**0.341**	0.0454	**0.216**	0.216
adult	2	NICE	**100**	**100**	2.009	2.0	**0.0**	**0.000**	0.497	0.0112	1.414	1.000
adult	2	FAB	**100**	**100**	0.237	**1.0**	**0.0**	**0.000**	0.344	0.0053	0.237	0.237
adult	2	DeepFool	**100**	99	0.241	**1.0**	**0.0**	**0.000**	0.348	**0.0001**	0.241	0.241
adult	2	MIP-EXP	**100**	**100**	0.237	**1.0**	**0.0**	**0.000**	0.344	3.7562	0.237	0.237
adult	2	MIP-OBJ	**100**	**100**	0.237	**1.0**	**0.0**	**0.000**	0.344	1.4630	0.237	0.237
adult	4	DASH	**100**	**100**	0.367	**1.0**	**0.0**	0.000	0.619	0.0629	0.367	0.367
adult	4	DASH + data seed	**100**	**100**	0.367	**1.0**	**0.0**	0.000	0.619	0.0759	0.367	0.367
adult	4	Wachter	95	90	2.362	5.0	1.0	0.000	1.433	1.3335	1.431	1.000
adult	4	DiCE	**100**	**100**	0.727	3.0	**0.0**	0.000	0.630	0.0010	0.408	**0.224**
adult	4	Greedy*	**100**	**100**	0.400	**1.0**	**0.0**	0.000	0.648	0.0005	0.400	0.400
adult	4	GS*	100	79	0.584	1.0	0.0	0.000	0.719	0.0729	0.584	0.584
adult	4	PGD-bisect*	**100**	**100**	0.367	**1.0**	**0.0**	0.000	0.619	0.3274	0.367	0.367
adult	4	FMN*	92	92	**0.362**	**1.0**	**0.0**	0.000	**0.580**	0.0937	**0.362**	0.362
adult	4	NICE	**100**	**100**	2.204	2.0	**0.0**	**0.200**	0.996	0.0170	1.429	1.000
adult	4	FAB	**100**	**100**	0.367	**1.0**	**0.0**	0.000	0.619	0.0086	0.367	0.367
adult	4	DeepFool	**100**	95	0.366	**1.0**	**0.0**	0.000	0.629	**0.0002**	0.366	0.366
adult	4	MIP-EXP	**100**	**100**	0.384	**1.0**	**0.0**	0.000	0.766	60.0003	0.384	0.384
adult	4	MIP-OBJ	99	99	8.000	5.0	1.0	**0.200**	1.922	60.0061	2.828	1.000
adult	8	DASH	**100**	**100**	0.694	**1.0**	**0.0**	0.000	1.004	0.0873	0.694	0.694
adult	8	DASH + data seed	**100**	**100**	0.694	**1.0**	**0.0**	0.000	1.004	0.0940	0.694	0.694
adult	8	Wachter	83	78	2.432	5.0	2.0	0.000	1.438	2.0056	1.435	1.000
adult	8	DiCE	97	95	1.074	4.0	**0.0**	0.000	**0.872**	0.0021	**0.560**	**0.302**
adult	8	Greedy*	**100**	82	**0.600**	**1.0**	**0.0**	0.000	0.933	**0.0012**	0.600	0.600
adult	8	GS*	100	37	0.584	1.0	0.0	0.000	0.634	0.1199	0.584	0.584
adult	8	PGD-bisect*	70	70	0.478	1.0	0.0	0.000	0.773	0.4647	0.478	0.478
adult	8	FMN*	67	67	0.478	1.0	0.0	0.000	0.769	0.1431	0.478	0.478
adult	8	NICE	**100**	**100**	2.565	3.0	**0.0**	**0.200**	0.972	0.0248	1.491	1.000
adult	8	FAB	100	78	0.514	1.0	0.0	0.000	0.913	0.0151	0.514	0.514
adult	8	DeepFool	100	60	0.460	1.0	0.0	0.000	0.715	0.0003	0.460	0.460
adult	8	MIP-EXP	2	2	0.017	1.0	0.0	0.200	0.067	7.6279	0.017	0.017
adult	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	16	DASH	**100**	**100**	0.336	**1.0**	**0.0**	0.000	0.747	0.0597	0.336	0.336
adult	16	DASH + data seed	**100**	**100**	0.336	**1.0**	**0.0**	0.000	0.747	0.0643	0.336	0.336
adult	16	Wachter	92	87	2.334	5.0	2.0	0.000	1.433	2.7226	1.427	1.000
adult	16	DiCE	97	95	0.596	3.0	**0.0**	0.000	0.613	0.0020	0.358	**0.217**
adult	16	Greedy*	**100**	91	0.300	**1.0**	**0.0**	0.000	0.607	0.0014	0.300	0.300
adult	16	GS*	100	66	0.400	1.0	0.0	0.000	0.594	0.1953	0.400	0.400
adult	16	PGD-bisect*	84	84	**0.286**	**1.0**	**0.0**	0.000	**0.513**	0.6787	**0.286**	0.286
adult	16	FMN*	89	89	0.292	**1.0**	**0.0**	0.000	0.557	0.3253	0.292	0.292
adult	16	NICE	**100**	**100**	2.360	3.0	**0.0**	**0.200**	0.993	0.0435	1.444	1.000
adult	16	FAB	**100**	91	0.292	**1.0**	**0.0**	0.000	0.570	0.0277	0.292	0.292
adult	16	DeepFool	**100**	84	0.326	**1.0**	**0.0**	0.000	0.620	**0.0004**	0.326	0.326
adult	32	DASH	**100**	**100**	0.850	**1.0**	**0.0**	0.000	1.147	0.1049	0.844	0.839
adult	32	DASH + data seed	**100**	**100**	0.850	**1.0**	**0.0**	0.000	1.147	0.1100	0.844	0.839
adult	32	Wachter	85	83	2.529	5.0	2.0	0.000	1.444	13.6492	1.444	1.000
adult	32	DiCE	89	88	1.065	4.0	**0.0**	0.000	0.934	0.0158	**0.598**	**0.334**
adult	32	Greedy*	**100**	83	**0.800**	**1.0**	**0.0**	0.000	1.026	**0.0141**	0.800	0.800
adult	32	GS*	100	29	0.584	1.0	0.0	0.000	0.718	0.3705	0.584	0.584
adult	32	PGD-bisect*	49	49	0.534	1.0	0.0	0.000	0.785	2.2078	0.529	0.529
adult	32	FMN*	53	53	0.653	1.0	0.0	0.000	0.975	1.4514	0.653	0.653
adult	32	NICE	**100**	**100**	2.703	3.0	**0.0**	**0.200**	**0.859**	0.0786	1.519	1.000
adult	32	FAB	100	75	0.673	1.0	0.0	0.000	0.987	0.1463	0.661	0.661
adult	32	DeepFool	99	39	0.493	1.0	0.0	0.000	0.680	0.0034	0.493	0.493
compas	2	DASH	99	**99**	0.147	**1.0**	**0.0**	0.400	0.031	0.0159	0.147	0.147
compas	2	DASH + data seed	99	**99**	0.147	**1.0**	**0.0**	0.400	0.031	0.0163	0.147	0.147
compas	2	Wachter	93	93	**0.143**	2.0	1.0	0.600	**0.029**	0.9717	**0.136**	**0.136**
compas	2	DiCE	99	98	0.183	**1.0**	**0.0**	**0.800**	**0.029**	0.0005	0.150	0.149
compas	2	Greedy*	**100**	97	0.200	**1.0**	**0.0**	**0.800**	0.037	**0.0002**	0.200	0.200
compas	2	GS*	100	77	0.171	1.0	0.0	0.800	0.031	0.0033	0.171	0.171
compas	2	PGD-bisect*	96	96	0.246	2.0	1.0	**0.800**	0.033	0.0658	0.231	0.230
compas	2	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	NICE	98	98	0.154	**1.0**	**0.0**	0.600	**0.028**	0.0022	0.154	0.154

Continued on next page

Table 15: All tabular cells, ℓ_1 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 12) and the best-known table (Table 11).

dataset	depth	method	success %	valid %	ℓ_1	l0	redund.	yNN	yNN-dist	time (s)	ℓ_2	ℓ_∞
compas	2	FAB	100	9	0.017	1.0	0.0	0.000	0.005	0.0041	0.017	0.017
compas	2	DeepFool	100	1	0.132	1.0	0.0	0.600	0.040	0.0001	0.132	0.132
compas	2	MIP-EXP	99	**99**	0.147	**1.0**	**0.0**	0.400	0.031	1.0463	0.147	0.147
compas	2	MIP-OBJ	99	**99**	0.147	**1.0**	**0.0**	0.400	0.031	0.8786	0.147	0.147
compas	4	DASH	**98**	**98**	0.137	**1.0**	**0.0**	0.600	0.027	0.0245	0.137	0.137
compas	4	DASH + data seed	**98**	**98**	0.137	**1.0**	**0.0**	0.600	0.028	0.0253	0.137	0.137
compas	4	Wachter	90	90	**0.114**	2.0	1.0	0.600	0.026	1.2992	**0.114**	**0.114**
compas	4	DiCE	97	97	0.150	**1.0**	**0.0**	0.800	0.028	0.0006	0.150	0.149
compas	4	Greedy*	**98**	97	0.200	**1.0**	**0.0**	**1.000**	0.028	**0.0003**	0.200	0.200
compas	4	GS*	100	67	0.171	1.0	0.0	0.800	0.025	0.0057	0.171	0.171
compas	4	PGD-bisect*	91	91	0.264	2.0	1.0	**1.000**	0.028	0.1646	0.259	0.259
compas	4	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	NICE	95	95	0.122	**1.0**	**0.0**	0.800	**0.024**	0.0040	0.122	0.122
compas	4	FAB	100	7	0.016	1.0	0.0	0.600	0.004	0.0073	0.016	0.016
compas	4	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	MIP-EXP	95	95	0.120	**1.0**	**0.0**	0.600	0.027	11.0061	0.120	0.120
compas	4	MIP-OBJ	95	95	0.256	**1.0**	**0.0**	0.800	0.035	60.0039	0.256	0.256
compas	8	DASH	**98**	**98**	**0.144**	**1.0**	**0.0**	0.400	0.027	0.0388	**0.144**	0.144
compas	8	DASH + data seed	**98**	**98**	**0.144**	**1.0**	**0.0**	0.400	0.027	0.0407	**0.144**	0.144
compas	8	Wachter	91	91	**0.144**	2.0	1.0	0.600	**0.027**	1.7761	**0.144**	0.144
compas	8	DiCE	94	94	0.149	**1.0**	**0.0**	0.800	**0.026**	0.0009	0.149	0.149
compas	8	Greedy*	97	85	0.200	**1.0**	**0.0**	**1.000**	0.029	**0.0004**	0.200	0.200
compas	8	GS*	100	59	0.101	1.0	0.0	0.600	0.025	0.0101	0.101	0.101
compas	8	PGD-bisect*	95	95	0.432	2.0	1.0	0.800	0.032	0.2912	0.426	0.426
compas	8	FMN*	7	7	0.389	1.0	0.0	0.600	0.070	0.1478	0.389	0.389
compas	8	NICE	**98**	**98**	0.154	**1.0**	**0.0**	0.600	**0.026**	0.0075	0.154	**0.141**
compas	8	FAB	99	7	0.006	1.0	0.0	0.000	0.005	0.0138	0.006	0.006
compas	8	DeepFool	100	10	0.322	1.0	0.0	0.800	0.040	0.0002	0.322	0.322
compas	8	MIP-EXP	45	45	0.049	1.0	0.0	0.600	0.015	5.8384	0.048	0.048
compas	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	16	DASH	**98**	**98**	**0.132**	**1.0**	**0.0**	0.600	0.035	0.0329	0.132	0.132
compas	16	DASH + data seed	**98**	**98**	**0.132**	**1.0**	**0.0**	0.600	0.035	0.0349	0.132	0.132
compas	16	Wachter	92	92	**0.133**	2.0	1.0	0.600	0.031	2.7179	0.132	0.132
compas	16	DiCE	86	86	0.142	**1.0**	**0.0**	**0.800**	**0.027**	**0.0012**	0.140	0.140
compas	16	Greedy*	96	79	0.200	1.0	0.0	1.000	0.028	0.0007	0.200	0.200
compas	16	GS*	100	59	0.168	1.0	0.0	0.600	0.028	0.0195	0.152	0.152
compas	16	PGD-bisect*	89	89	2.020	3.0	1.0	**0.800**	0.059	0.5507	1.414	1.000
compas	16	FMN*	23	23	0.163	1.0	0.0	0.600	0.028	0.3538	0.163	0.163
compas	16	NICE	97	97	0.139	**1.0**	**0.0**	0.600	0.028	0.0156	**0.128**	**0.122**
compas	16	FAB	94	8	0.032	1.0	0.0	0.700	0.005	0.0273	0.032	0.032
compas	16	DeepFool	100	10	0.141	1.0	0.0	0.800	0.027	0.0004	0.141	0.141
compas	32	DASH	**100**	**100**	0.159	**1.0**	**0.0**	0.600	0.033	0.0726	0.155	0.155
compas	32	DASH + data seed	**100**	**100**	0.159	**1.0**	**0.0**	0.600	0.033	0.0748	0.155	0.155
compas	32	Wachter	80	80	**0.140**	2.0	1.0	**0.800**	**0.029**	11.4730	**0.131**	**0.131**
compas	32	DiCE	70	70	0.100	1.0	0.0	0.800	0.022	0.0044	0.100	0.100
compas	32	Greedy*	86	77	0.200	1.0	0.0	1.000	0.027	0.0034	0.171	0.150
compas	32	GS*	100	56	0.171	1.0	0.0	0.800	0.028	0.0431	0.171	0.171
compas	32	PGD-bisect*	87	87	0.669	2.0	1.0	**0.800**	0.072	1.8393	0.630	0.628
compas	32	FMN*	35	35	0.454	1.0	0.0	0.600	0.060	1.5718	0.454	0.454
compas	32	NICE	99	99	0.167	**1.0**	**0.0**	**0.800**	**0.030**	**0.0285**	0.167	0.167
compas	32	FAB	57	7	0.023	1.0	0.0	0.600	0.007	0.1435	0.023	0.023
compas	32	DeepFool	100	1	0.102	1.0	0.0	1.000	0.025	0.0018	0.102	0.102
gmc	2	DASH	**99**	**99**	**0.018**	**1.0**	**0.0**	0.600	**0.035**	0.0179	**0.018**	**0.018**
gmc	2	DASH + data seed	**99**	**99**	**0.018**	**1.0**	**0.0**	0.600	**0.035**	0.0239	**0.018**	**0.018**
gmc	2	Wachter	**99**	**99**	**0.018**	3.0	1.0	0.600	**0.035**	0.6458	**0.018**	**0.018**
gmc	2	DiCE	**99**	**99**	0.050	**1.0**	**0.0**	0.800	0.057	0.0003	0.050	0.050
gmc	2	Greedy*	**99**	**99**	0.100	**1.0**	**0.0**	**1.000**	0.098	**0.0001**	0.100	0.100
gmc	2	GS*	**99**	**99**	**0.019**	**1.0**	**0.0**	0.600	0.037	0.0032	**0.019**	**0.019**
gmc	2	PGD-bisect*	98	98	0.026	3.0	2.0	0.600	0.037	0.0438	0.022	0.021
gmc	2	FMN*	28	28	0.015	1.0	0.0	0.500	0.036	0.0363	0.015	0.015
gmc	2	NICE	97	97	0.023	**1.0**	**0.0**	0.600	0.036	0.0923	0.022	0.021
gmc	2	FAB	100	27	0.015	1.0	0.0	0.500	0.037	0.0039	0.015	0.015
gmc	2	DeepFool	100	30	0.016	1.0	0.0	0.600	0.035	0.0001	0.016	0.016
gmc	2	MIP-EXP	**99**	**99**	**0.018**	**1.0**	**0.0**	0.600	**0.035**	0.4350	**0.018**	**0.018**
gmc	2	MIP-OBJ	98	98	**0.018**	**1.0**	**0.0**	0.600	**0.034**	0.1949	**0.018**	**0.018**
gmc	4	DASH	**100**	**100**	0.033	**1.0**	**0.0**	0.600	**0.046**	0.0302	0.033	0.033
gmc	4	DASH + data seed	**100**	**100**	0.033	**1.0**	**0.0**	0.600	**0.046**	0.0353	0.033	0.033
gmc	4	Wachter	99	99	0.033	3.0	1.0	0.600	**0.046**	0.9046	0.033	0.033
gmc	4	DiCE	97	97	0.050	**1.0**	**0.0**	0.800	0.058	0.0003	0.050	0.050
gmc	4	Greedy*	99	99	0.100	**1.0**	**0.0**	**1.000**	0.098	**0.0002**	0.100	0.100

Continued on next page

Table 15: All tabular cells, ℓ_1 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 12) and the best-known table (Table 11).

dataset	depth	method	success %	valid %	ℓ_1	l0	redund.	yNN	yNN-dist	time (s)	ℓ_2	ℓ_∞
gmc	4	GS*	**100**	**100**	0.036	**1.0**	**0.0**	0.800	0.048	0.0056	0.036	0.036
gmc	4	PGD-bisect*	97	97	0.086	4.0	3.0	0.800	0.053	0.0914	0.061	0.048
gmc	4	FMN*	1	1	0.187	1.0	0.0	0.100	0.192	0.0666	0.187	0.187
gmc	4	NICE	98	98	0.038	**1.0**	**0.0**	0.800	0.048	0.1870	0.038	0.038
gmc	4	FAB	100	2	0.000	1.0	0.0	0.600	0.022	0.0071	0.000	0.000
gmc	4	DeepFool	100	1	0.312	1.0	0.0	1.000	0.205	0.0001	0.312	0.312
gmc	4	MIP-EXP	**100**	**100**	0.033	**1.0**	**0.0**	0.600	**0.047**	1.3364	0.033	0.033
gmc	4	MIP-OBJ	98	98	**0.032**	**1.0**	**0.0**	0.600	**0.046**	13.4699	**0.032**	**0.032**
gmc	8	DASH	**99**	**99**	0.039	**1.0**	**0.0**	0.800	0.049	0.0478	0.039	0.039
gmc	8	DASH + data seed	**99**	**99**	0.039	**1.0**	**0.0**	0.800	0.049	0.0543	0.039	0.039
gmc	8	Wachter	97	97	0.038	3.0	1.0	0.800	0.049	1.3670	0.038	0.038
gmc	8	DiCE	91	91	0.050	**1.0**	**0.0**	0.800	0.056	0.0004	0.050	0.050
gmc	8	Greedy*	91	91	0.100	**1.0**	**0.0**	**1.000**	0.095	**0.0003**	0.100	0.100
gmc	8	GS*	**99**	**99**	0.040	**1.0**	**0.0**	0.800	0.050	0.0097	0.039	0.039
gmc	8	PGD-bisect*	97	97	0.087	4.0	2.0	0.800	0.055	0.1687	0.062	0.049
gmc	8	FMN*	1	1	0.072	1.0	0.0	1.000	0.077	0.1489	0.072	0.072
gmc	8	NICE	98	98	0.048	**1.0**	**0.0**	0.800	0.050	0.7981	0.040	0.039
gmc	8	FAB	100	1	0.005	1.0	0.0	0.900	0.052	0.0139	0.005	0.005
gmc	8	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	8	MIP-EXP	87	87	**0.031**	**1.0**	**0.0**	0.800	**0.044**	4.4221	**0.031**	**0.031**
gmc	8	MIP-OBJ	2	2	0.389	1.0	0.0	0.000	0.391	1.6557	0.389	0.389
gmc	16	DASH	**99**	**99**	**0.037**	**1.0**	**0.0**	0.800	**0.050**	0.0369	**0.037**	**0.037**
gmc	16	DASH + data seed	**99**	**99**	**0.037**	**1.0**	**0.0**	0.800	**0.050**	0.0439	**0.037**	**0.037**
gmc	16	Wachter	**99**	**99**	**0.037**	3.0	1.0	0.800	**0.050**	2.1617	**0.037**	**0.037**
gmc	16	DiCE	98	98	0.050	**1.0**	**0.0**	0.800	0.059	0.0006	0.050	0.050
gmc	16	Greedy*	**99**	**99**	0.100	**1.0**	**0.0**	**1.000**	0.097	**0.0005**	0.100	0.100
gmc	16	GS*	**99**	**99**	0.041	**1.0**	**0.0**	0.800	**0.049**	0.0196	0.039	0.039
gmc	16	PGD-bisect*	98	98	0.090	4.0	3.0	0.800	0.055	0.3551	0.062	0.050
gmc	16	FMN*	3	3	0.268	1.0	0.0	0.100	0.257	0.3116	0.268	0.268
gmc	16	NICE	**99**	**99**	0.048	**1.0**	**0.0**	0.800	0.052	1.1659	0.047	0.047
gmc	16	FAB	100	2	0.000	1.0	0.0	1.000	0.026	0.0268	0.000	0.000
gmc	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	32	DASH	**98**	**98**	**0.035**	**1.0**	**0.0**	0.800	**0.046**	0.0784	**0.035**	**0.035**
gmc	32	DASH + data seed	**98**	**98**	**0.035**	**1.0**	**0.0**	0.800	**0.046**	0.0873	**0.035**	**0.035**
gmc	32	Wachter	96	96	**0.035**	3.0	1.0	0.800	**0.045**	9.3777	**0.035**	**0.035**
gmc	32	DiCE	92	92	0.050	**1.0**	**0.0**	0.800	0.055	0.0025	0.050	0.050
gmc	32	Greedy*	93	93	0.100	**1.0**	**0.0**	**1.000**	0.094	**0.0022**	0.100	0.100
gmc	32	GS*	**98**	**98**	0.036	**1.0**	**0.0**	0.800	0.046	0.0373	0.036	0.036
gmc	32	PGD-bisect*	96	96	0.089	4.0	3.0	0.800	0.055	1.1331	0.062	0.049
gmc	32	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	32	NICE	**98**	**98**	0.047	**1.0**	**0.0**	0.800	0.050	2.3543	0.043	0.040
gmc	32	FAB	98	3	0.049	1.5	0.0	0.600	0.072	0.1415	0.043	0.041
gmc	32	DeepFool	99	0	NA	NA	NA	NA	NA	NA	NA	NA
heloc	2	DASH	**100**	**100**	0.394	**2.0**	**0.0**	0.200	0.423	0.0760	0.363	0.350
heloc	2	DASH + data seed	**100**	**100**	0.394	**2.0**	**0.0**	0.200	0.423	0.0842	0.362	0.350
heloc	2	Wachter	**100**	**100**	0.446	20.0	10.0	0.200	0.325	0.9888	0.228	0.176
heloc	2	DiCE	**100**	**100**	0.648	8.0	**0.0**	0.200	0.325	0.0003	**0.219**	**0.100**
heloc	2	Greedy*	**100**	**100**	0.430	**2.0**	**0.0**	0.200	0.438	**0.0003**	0.400	0.350
heloc	2	GS*	**100**	**100**	0.542	6.0	**0.0**	0.200	0.372	0.0049	0.275	0.208
heloc	2	PGD-bisect*	**100**	**100**	0.485	3.0	**0.0**	0.200	0.414	0.1006	0.351	0.306
heloc	2	FMN*	64	64	0.387	2.0	0.0	0.200	0.396	0.0377	0.324	0.300
heloc	2	NICE	**100**	**100**	0.613	6.0	1.0	**0.400**	**0.270**	0.0035	0.318	0.230
heloc	2	FAB	**100**	**100**	0.395	**2.0**	**0.0**	0.200	0.420	0.0041	0.351	0.334
heloc	2	DeepFool	100	59	0.382	1.0	0.0	0.200	0.419	0.0001	0.364	0.364
heloc	2	MIP-EXP	**100**	**100**	0.390	**2.0**	**0.0**	0.200	0.415	19.1915	0.339	0.331
heloc	2	MIP-OBJ	99	99	**0.388**	**2.0**	**0.0**	0.200	0.414	6.6213	0.338	0.327
heloc	4	DASH	**100**	**100**	0.523	**2.0**	**0.0**	0.200	0.421	0.1440	0.442	0.422
heloc	4	DASH + data seed	**100**	**100**	**0.519**	**2.0**	**0.0**	0.200	0.421	0.1489	0.438	0.413
heloc	4	Wachter	**100**	**100**	0.587	20.0	10.0	0.000	0.330	1.2079	**0.256**	0.164
heloc	4	DiCE	**100**	**100**	0.793	9.0	**0.0**	0.200	0.328	**0.0004**	0.264	**0.100**
heloc	4	Greedy*	**100**	**100**	0.586	3.0	**0.0**	0.200	0.422	0.0006	0.419	0.400
heloc	4	GS*	**100**	**100**	0.721	6.0	**0.0**	0.200	0.382	0.0096	0.362	0.288
heloc	4	PGD-bisect*	**100**	**100**	0.951	18.5	11.0	0.200	0.456	0.1364	0.423	0.336
heloc	4	FMN*	36	36	0.501	2.0	0.0	0.200	0.406	0.0669	0.405	0.343
heloc	4	NICE	**100**	**100**	0.766	7.0	1.0	**0.400**	**0.282**	0.0058	0.378	0.246
heloc	4	FAB	**100**	**100**	**0.518**	3.0	**0.0**	0.200	0.407	0.0073	0.392	0.358
heloc	4	DeepFool	99	27	0.357	1.0	0.0	0.400	0.412	0.0002	0.345	0.345
heloc	4	MIP-EXP	29	29	0.205	2.0	0.0	0.400	0.305	60.0052	0.177	0.156
heloc	4	MIP-OBJ	**100**	**100**	0.748	6.0	**0.0**	0.200	0.429	60.0050	0.441	0.332
heloc	8	DASH	**100**	**100**	0.436	**2.0**	**0.0**	0.200	0.416	0.2130	0.372	0.363

Continued on next page

Table 15: All tabular cells, ℓ_1 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 12) and the best-known table (Table 11).

dataset	depth	method	success %	valid %	ℓ_1	l0	redund.	yNN	yNN-dist	time (s)	ℓ_2	ℓ_∞
heloc	8	DASH + data seed	**100**	**100**	0.430	**2.0**	**0.0**	0.200	0.411	0.2203	0.362	0.357
heloc	8	Wachter	**100**	**100**	0.485	20.0	9.5	0.200	0.318	1.8254	0.224	0.155
heloc	8	DiCE	99	99	0.699	9.0	**0.0**	0.200	0.317	**0.0005**	**0.218**	**0.050**
heloc	8	Greedy*	99	99	0.500	3.0	**0.0**	0.200	0.384	0.0010	0.333	0.300
heloc	8	GS*	**100**	**100**	0.582	6.0	**0.0**	0.200	0.360	0.0159	0.292	0.214
heloc	8	PGD-bisect*	**100**	**100**	0.538	6.0	1.5	0.200	0.401	0.3132	0.344	0.273
heloc	8	FMN*	67	67	0.372	3.0	0.0	0.200	0.367	0.1556	0.320	0.280
heloc	8	NICE	**100**	**100**	0.634	6.0	1.0	**0.400**	**0.287**	0.0115	0.344	0.248
heloc	8	FAB	98	98	**0.418**	3.0	**0.0**	0.200	0.371	0.0137	0.317	0.277
heloc	8	DeepFool	99	35	0.305	1.0	0.0	0.400	0.359	0.0003	0.278	0.273
heloc	8	MIP-EXP	7	7	0.032	1.0	0.0	0.400	0.222	11.0365	0.032	0.028
heloc	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
heloc	16	DASH	**100**	**100**	0.490	**3.0**	**0.0**	0.200	0.387	0.0852	0.349	0.321
heloc	16	DASH + data seed	**100**	**100**	0.496	**3.0**	**0.0**	0.200	0.381	0.0955	0.354	0.312
heloc	16	Wachter	**100**	**100**	0.521	20.0	9.0	0.200	0.322	2.8045	**0.239**	0.160
heloc	16	DiCE	98	98	0.750	9.0	**0.0**	0.200	0.330	**0.0010**	0.262	**0.100**
heloc	16	Greedy*	**100**	**100**	0.526	**3.0**	**0.0**	0.200	0.366	0.0020	0.334	0.300
heloc	16	GS*	**100**	**100**	0.670	6.0	**0.0**	0.200	0.363	0.0326	0.318	0.220
heloc	16	PGD-bisect*	**100**	**100**	0.627	5.0	**0.0**	0.200	0.375	0.6255	0.332	0.253
heloc	16	FMN*	77	77	0.495	3.0	0.0	0.200	0.358	0.3368	0.320	0.288
heloc	16	NICE	**100**	**100**	0.675	6.0	1.0	**0.400**	**0.289**	0.0189	0.354	0.258
heloc	16	FAB	96	96	**0.456**	4.0	**0.0**	0.200	0.358	0.0260	0.308	0.260
heloc	16	DeepFool	100	35	0.262	2.0	0.0	0.200	0.354	0.0007	0.225	0.216
heloc	32	DASH	**100**	**100**	0.322	**2.0**	**0.0**	0.200	0.343	0.1326	0.237	0.194
heloc	32	DASH + data seed	**100**	**100**	0.327	**2.0**	**0.0**	0.200	0.340	0.1462	0.245	0.203
heloc	32	Wachter	**100**	**100**	0.343	20.0	9.5	0.200	0.316	12.1622	0.184	0.131
heloc	32	DiCE	**100**	**100**	0.500	8.0	**0.0**	0.200	0.317	**0.0028**	**0.158**	**0.050**
heloc	32	Greedy*	**100**	**100**	0.400	3.0	**0.0**	0.200	0.354	0.0078	0.279	0.200
heloc	32	GS*	**100**	**100**	0.491	6.0	**0.0**	0.200	0.340	0.0487	0.237	0.160
heloc	32	PGD-bisect*	**100**	**100**	0.497	5.0	1.0	0.200	0.342	1.9645	0.242	0.189
heloc	32	FMN*	68	68	0.295	3.0	0.0	0.200	0.342	1.5060	0.214	0.181
heloc	32	NICE	**100**	**100**	0.499	6.0	1.0	**0.400**	**0.274**	0.0377	0.264	0.204
heloc	32	FAB	**100**	**100**	**0.320**	3.0	**0.0**	0.200	0.343	0.1411	0.235	0.196
heloc	32	DeepFool	100	47	0.312	2.0	0.0	0.200	0.366	0.0029	0.299	0.286

D.3 ℓ_2 OBJECTIVE

Table 16: CARLA axes at depth 8 (ℓ_2 objective, basic actionability; seeds 17/29/43).

dataset	method	success %	valid %	ℓ_2	l0	redund.	yNN	yNN-dist	time (s)
adult	DASH	**100**	**100**	**0.526**	4.0	**0.0**	0.000	0.904	0.1010
	DASH + data seed	**100**	**100**	**0.526**	4.0	**0.0**	0.000	0.903	0.1046
	Wachter	79	77	1.428	5.0	3.0	0.000	1.430	1.8042
	DiCE	99	92	0.550	4.0	**0.0**	0.000	**0.853**	0.0017
	Greedy*	**100**	82	0.600	**1.0**	**0.0**	0.000	0.933	**0.0012**
	GS*	100	17	0.544	2.0	0.0	0.000	0.941	0.0319
	PGD-bisect*	89	89	1.447	5.0	3.0	0.000	1.448	0.5200
	FMN*	0	0	NA	NA	NA	NA	NA	NA
	NICE	**100**	**100**	1.491	3.0	**0.0**	**0.200**	0.897	0.0247
	FAB	100	0	NA	NA	NA	NA	NA	NA
	DeepFool	100	0	NA	NA	NA	NA	NA	NA
	MIP-EXP	3	3	0.016	4.0	0.0	0.200	0.066	14.4452
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA
compas	DASH	**98**	**98**	**0.138**	2.0	**0.0**	0.400	**0.025**	0.0349
	DASH + data seed	**98**	**98**	**0.138**	2.0	**0.0**	0.400	**0.026**	0.0364
	Wachter	88	88	1.417	4.0	2.0	0.800	0.117	1.8195
	DiCE	93	93	0.149	**1.0**	**0.0**	0.800	**0.026**	0.0009
	Greedy*	97	85	0.200	**1.0**	**0.0**	**1.000**	0.029	**0.0004**
	GS*	100	37	0.089	1.0	0.0	0.800	0.021	0.0092
	PGD-bisect*	89	89	1.417	3.0	2.0	0.800	0.077	0.3575
	FMN*	11	11	1.437	3.0	2.0	0.800	0.330	0.1593
	NICE	**98**	**98**	0.144	**1.0**	**0.0**	0.600	**0.026**	0.0076
	FAB	99	1	0.000	5.0	0.0	0.600	0.003	0.0145
	DeepFool	100	0	NA	NA	NA	NA	NA	NA
	MIP-EXP	48	48	0.051	2.0	0.0	0.600	0.013	7.4927
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA
gmc	DASH	**99**	**99**	0.036	3.0	**0.0**	0.600	0.048	0.0465
	DASH + data seed	**99**	**99**	0.036	3.0	**0.0**	0.600	0.048	0.0534
	Wachter	97	97	0.035	3.0	**0.0**	0.600	0.047	1.5871
	DiCE	91	91	0.050	**1.0**	**0.0**	0.800	0.056	0.0004
	Greedy*	91	91	0.100	**1.0**	**0.0**	**1.000**	0.095	**0.0003**
	GS*	**99**	**99**	0.041	**1.0**	**0.0**	0.800	0.049	0.0117
	PGD-bisect*	**99**	**99**	0.040	3.0	1.0	0.800	0.047	0.3008
	FMN*	1	1	0.100	4.0	1.0	0.800	0.080	0.1472
	NICE	98	98	0.039	**1.0**	**0.0**	0.800	0.046	0.8054
	FAB	100	1	0.000	5.0	0.0	0.900	0.017	0.0144
	DeepFool	100	1	0.000	8.0	8.0	0.800	0.017	0.0002
	MIP-EXP	91	91	**0.030**	3.0	1.0	0.800	**0.044**	5.1556
	MIP-OBJ	2	2	0.333	2.0	0.0	0.000	0.337	1.7697
heloc	DASH	**100**	**100**	0.180	21.0	1.0	0.000	0.302	0.2332
	DASH + data seed	**100**	**100**	0.180	21.0	1.0	0.000	0.302	0.2461
	Wachter	**100**	**100**	**0.178**	22.0	1.0	0.000	0.303	1.6352
	DiCE	99	99	0.218	9.0	**0.0**	0.200	0.318	**0.0006**
	Greedy*	99	99	0.333	**3.0**	**0.0**	0.200	0.384	0.0010
	GS*	**100**	**100**	0.273	7.0	**0.0**	0.200	0.357	0.0075
	PGD-bisect*	**100**	**100**	**0.178**	21.0	**0.0**	0.000	0.301	0.3650
	FMN*	51	51	0.248	21.0	0.0	0.000	0.323	0.1556
	NICE	**100**	**100**	0.314	6.0	1.0	**0.400**	**0.294**	0.0114
	FAB	98	96	**0.178**	21.0	**0.0**	0.000	0.302	0.0147
	DeepFool	100	1	0.156	23.0	13.5	0.200	0.360	0.0002
	MIP-EXP	11	11	0.021	20.0	4.0	0.400	0.252	4.6867
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA

success = the returned point crosses the threshold (CARLA's success rate); valid = it also satisfies the actionability box and one-hot constraints. Bold = best per column within a dataset among arms with validity $\geq$ 80%. Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 20) and the best-known table (Table 19).

Table 17: ℓ_2 relative to the certified optimum (MIP-EXP or MIP-OBJ; Gurobi, 4 threads, 60 s budget per factual, absolute optimality gap 10^{-4}, depths 2/4/8, same 50 factuals per cell as the main run). certified n = instances on which either exact formulation proved an optimum in that dataset; arm valid = how many of those the arm returns a valid point on; at optimum = arm valid and distance $\leq$ ref $+10^{-3}$ (one-sided: below the solver's incumbent is inside its gap); within 5% = arm valid and distance $\leq 1.05\times$ref (purely relative: where the optimum is below 0.02 the absolute tie band is the wider of the two, so the columns are not nested); both percentages over certified n; ratio columns over the arm-valid subset. A valid point can lie up to 6×10^{-4} below the certified optimum (the solver's absolute gap and numerical tolerances), so ratios below 1 occur where the optimum itself is small: every ratio below 0.95 has an optimum under 2.5×10^{-3}.

dataset	arm	certified n	arm valid	at optimum %	within 5 %	median ratio	max ratio
adult	DASH	184	184	76.6	94.6	1.001	1.162
adult	DASH + data seed	184	184	76.1	94.6	1.001	1.162
adult	Wachter	184	178	4.3	6.5	5.346	47.124
adult	DiCE	184	183	0.0	0.5	1.351	71.445
adult	Greedy*	184	184	0.0	1.1	1.327	142.891
adult	GS*	184	86	0.0	1.1	1.278	6.897
adult	PGD-bisect*	184	184	0.0	0.0	7.444	2020.784
adult	FMN*	184	0	0.0	0.0	NA	NA
adult	NICE	184	184	0.0	0.0	4.175	626.971
adult	FAB	184	0	0.0	0.0	NA	NA
adult	DeepFool	184	0	0.0	0.0	NA	NA
compas	DASH	310	310	83.2	96.1	1.002	1.158
compas	DASH + data seed	310	310	82.6	95.8	1.002	1.158
compas	Wachter	310	291	2.6	1.9	15.074	361.026
compas	DiCE	310	308	1.3	12.3	1.237	166.667
compas	Greedy*	310	302	0.0	3.5	1.598	333.334
compas	GS*	310	199	0.3	8.4	1.168	5.715
compas	PGD-bisect*	310	291	0.0	0.3	18.918	4714.050
compas	FMN*	310	0	0.0	0.0	NA	NA
compas	NICE	310	308	5.5	37.4	1.078	9.149
compas	FAB	310	1	0.3	0.3	0.858	0.858
compas	DeepFool	310	0	0.0	0.0	NA	NA
gmc	DASH	409	409	90.5	95.8	1.001	1.153
gmc	DASH + data seed	409	409	91.2	96.1	1.001	1.153
gmc	Wachter	409	409	97.1	98.3	1.000	2.220
gmc	DiCE	409	405	0.0	2.9	2.266	524.464
gmc	Greedy*	409	404	0.0	1.7	4.484	1048.928
gmc	GS*	409	409	14.4	20.5	1.124	34.253
gmc	PGD-bisect*	409	409	78.7	79.7	1.013	1.650
gmc	FMN*	409	1	0.2	0.2	1.000	1.000
gmc	NICE	409	402	26.7	22.5	1.119	24.268
gmc	FAB	409	7	1.7	1.7	0.291	0.443
gmc	DeepFool	409	2	0.5	0.0	1.251	1.383
heloc	DASH	119	119	79.0	97.5	1.002	1.080
heloc	DASH + data seed	119	119	79.0	96.6	1.002	1.080
heloc	Wachter	119	119	78.2	98.3	1.001	1.108
heloc	DiCE	119	119	0.0	0.0	1.454	76.923
heloc	Greedy*	119	119	0.0	0.0	2.523	153.847
heloc	GS*	119	119	0.0	0.0	1.783	46.314
heloc	PGD-bisect*	119	119	100.0	100.0	1.000	1.008
heloc	FMN*	119	17	14.3	14.3	1.000	1.001
heloc	NICE	119	119	0.0	0.0	1.990	97.371
heloc	FAB	119	119	89.9	100.0	1.001	1.013
heloc	DeepFool	119	2	0.8	0.8	1.061	1.092
all	DASH	1022	1022	84.4	95.9	1.001	1.162
all	DASH + data seed	1022	1022	84.4	95.8	1.001	1.162
all	Wachter	1022	997	49.5	52.5	1.013	361.026
all	DiCE	1022	1015	0.4	5.0	1.451	524.464
all	Greedy*	1022	1009	0.0	2.0	2.207	1048.928
all	GS*	1022	813	5.9	11.0	1.199	46.314
all	PGD-bisect*	1022	1003	43.2	43.6	1.096	4714.050
all	FMN*	1022	18	1.8	1.8	1.000	1.001
all	NICE	1022	1013	12.3	20.4	1.188	626.971
all	FAB	1022	127	11.3	12.4	1.000	1.013
all	DeepFool	1022	4	0.3	0.1	1.105	1.383

Table 18: MIP-EXP against DASH by depth, ℓ_2 objective, same queries, pooled over datasets and seeds. Median times are over all of these queries, solved or not (MIP-EXP stops at its 60 s budget).

depth	queries	MIP-EXP valid %	MIP-EXP certified %	MIP-EXP median time (s)	DASH valid %	DASH median time (s)
2	600	96.5	82.0	1.924	99.7	0.035
4	600	76.5	48.7	57.127	99.5	0.056
8	600	38.0	31.5	60.000	99.2	0.076

Table 19: Best-known counterfactual, ℓ_2 objective: population is every tabular query of this norm (the queries column); best = smallest valid distance among DASH and the nine prior methods on that query (the exact solvers and the ablation are not in the reference); a query counts for an arm at τ iff the arm is valid and its distance $\leq (1+\tau)\times$best; column best uses the absolute tie band best $+10^{-3}$ instead, which is wider than 1% where best is below 0.1; invalid rows never count, no penalty constant; 95% CI = percentile cluster bootstrap over trained models (dataset $\times$ depth $\times$ seed), 2,000 resamples. Sorted by within 5%; bold = best per column.

method	queries	valid %	best	within 1 %	within 5 %	95 % CI (within 5 %)	within 10 %
DASH	3000	**99.4**	**82.9**	**83.0**	**95.9**	94.1–97.3	**98.0**
Wachter	3000	92.6	36.6	36.4	48.9	37.2–60.4	50.6
PGD-bisect*	3000	96.2	37.7	31.5	41.3	30.3–52.9	45.2
FAB	3000	24.3	18.8	22.0	23.8	13.7–33.9	24.1
NICE	3000	98.9	8.2	4.7	15.7	11.1–20.6	24.7
FMN*	3000	15.8	11.4	11.8	12.3	7.0–18.2	12.4
GS*	3000	70.4	3.8	1.1	8.4	5.8–11.4	16.2
DiCE	3000	95.0	1.0	1.4	5.8	3.6–8.3	12.8
Greedy*	3000	93.6	0.1	0.3	1.8	1.1–2.6	5.4
DeepFool	3000	0.5	0.2	0.0	0.2	0.0–0.3	0.3

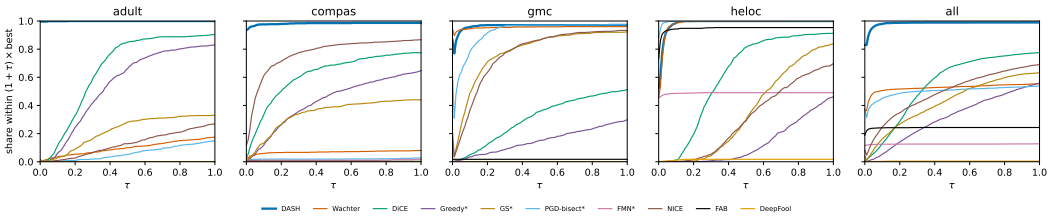

Figure 10: Performance profile, ℓ_2 objective (see Figure 6 for the definition).

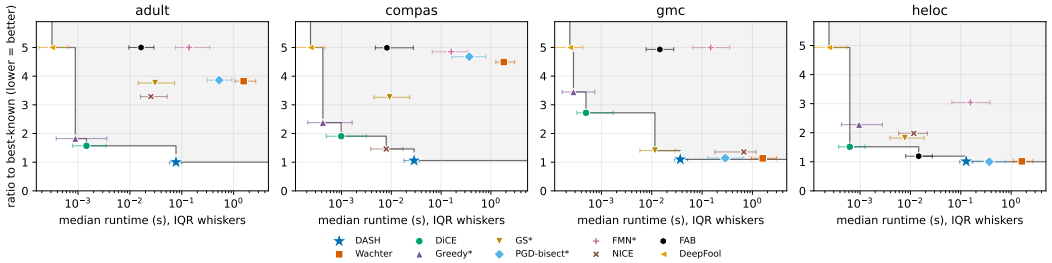

Figure 11: Runtime-quality trade-off, ℓ_2 objective (see Figure 7 for the definition).

Table 20: DASH against each arm on ℓ_2, paired per factual over the instances where both are valid: DASH wins = DASH is closer by more than 10^{-3}, losses = the other arm is closer; cells won = (depth, seed) cells ("all" row: dataset, depth, seed cells) with ≥ 1 pair, where the median paired difference favours DASH by more than 10^{-3}, printed as k/m; median ratio = median of DASH distance / other distance (below 1 favours DASH); two-sided Wilcoxon signed-rank on the non-tied pairs. Pairs within a cell share one trained model, so the pooled p-value overstates the evidence; cells won counts model-level agreement.

dataset	vs	n paired	DASH wins	ties	DASH losses	cells won	median ratio	Wilcoxon p
adult	DASH + data seed	750	31	696	23	0/15	1.000	0.77
adult	Wachter	645	633	10	2	15/15	0.245	1.2e-105
adult	DiCE	703	703	0	0	15/15	0.798	9.3e-117
adult	Greedy*	685	685	0	0	15/15	0.754	8e-114
adult	GS*	261	259	2	0	15/15	0.761	3.1e-44
adult	PGD-bisect*	717	716	0	1	15/15	0.229	4.9e-119
adult	NICE	750	750	0	0	15/15	0.312	2e-124
adult	MIP-EXP	286	83	154	49	3/8	1.000	1.2e-11
adult	MIP-OBJ	299	149	118	32	3/6	1.000	9.3e-28
compas	DASH + data seed	740	16	718	6	0/15	1.000	0.26
compas	Wachter	647	614	30	3	15/15	0.102	1.5e-102
compas	DiCE	675	639	22	14	15/15	0.865	1.2e-104
compas	Greedy*	653	648	4	1	15/15	0.711	2.1e-107
compas	GS*	358	345	11	2	15/15	0.857	1.4e-58
compas	PGD-bisect*	676	673	3	0	15/15	0.084	7.3e-112
compas	FMN*	104	103	0	1	9/9	0.223	9.3e-19
compas	NICE	728	640	58	30	15/15	0.943	4.2e-93
compas	FAB	2	0	2	0	0/2	1.030	NA
compas	MIP-EXP	362	43	265	54	0/9	1.000	0.018
compas	MIP-OBJ	277	106	142	29	3/6	1.000	1e-19
gmc	DASH + data seed	742	8	724	10	0/15	1.000	0.73
gmc	Wachter	736	22	625	89	0/15	1.001	8.3e-05
gmc	DiCE	722	719	2	1	15/15	0.545	4.9e-119
gmc	Greedy*	720	717	2	1	15/15	0.288	4.3e-119
gmc	GS*	742	605	117	20	15/15	0.912	1.5e-89
gmc	PGD-bisect*	742	233	471	38	4/15	0.985	1.4e-24
gmc	FMN*	1	0	0	1	0/1	1.123	NA
gmc	NICE	735	565	166	4	15/15	0.893	1.2e-92
gmc	FAB	13	0	13	0	0/9	3.666	NA
gmc	DeepFool	3	0	3	0	0/3	0.474	NA
gmc	MIP-EXP	435	25	371	39	0/9	0.999	0.17
gmc	MIP-OBJ	296	50	236	10	1/8	1.000	1.4e-09
heloc	DASH + data seed	750	90	600	60	0/15	1.000	0.03
heloc	Wachter	750	264	311	175	0/15	0.999	4.5e-08
heloc	DiCE	749	749	0	0	15/15	0.775	2.9e-124
heloc	Greedy*	749	749	0	0	15/15	0.493	2.9e-124
heloc	GS*	750	750	0	0	15/15	0.627	2e-124
heloc	PGD-bisect*	750	28	390	332	0/15	1.005	9.5e-43
heloc	FMN*	368	18	138	212	0/15	1.007	1.1e-25
heloc	NICE	750	750	0	0	15/15	0.596	2e-124
heloc	FAB	715	51	394	270	0/15	1.003	4e-25
heloc	DeepFool	13	8	5	0	6/9	0.920	0.0078
heloc	MIP-EXP	183	93	78	12	4/9	0.990	3.8e-16
heloc	MIP-OBJ	297	146	108	43	3/6	1.000	3.6e-26
all	DASH + data seed	2982	145	2738	99	0/60	1.000	0.022
all	Wachter	2778	1533	976	269	30/60	0.979	1.2e-245
all	DiCE	2849	2810	24	15	60/60	0.772	$< 10^{-300}$
all	Greedy*	2807	2799	6	2	60/60	0.572	$< 10^{-300}$
all	GS*	2111	1959	130	22	60/60	0.756	$< 10^{-300}$
all	PGD-bisect*	2885	1650	864	371	34/60	0.863	1.8e-261
all	FMN*	473	121	138	214	9/25	1.003	0.058
all	NICE	2963	2705	224	34	60/60	0.709	$< 10^{-300}$
all	FAB	730	51	409	270	0/26	1.003	4e-25
all	DeepFool	16	8	8	0	6/12	0.913	0.0078
all	MIP-EXP	1266	244	868	154	7/35	1.000	1.5e-25
all	MIP-OBJ	1169	451	604	114	10/26	1.000	1.6e-78

Table 21: Validity (%) per arm and cell, ℓ_2 objective.

dataset	depth	DASH	DASH + data seed	Wachter	DiCE	Greedy*	GS*	PGD-bisect*	FMN*	NICE	FAB	DeepFool	MIP-EXP	MIP-OBJ
adult	2	100	100	96	99	100	41	100	0	100	0	0	100	99
	4	100	100	88	97	100	57	100	0	100	0	0	88	100
	8	100	100	77	92	82	17	89	0	100	0	0	3	0
	16	100	100	87	91	91	48	97	0	100	0	0	–	–
	32	100	100	82	89	83	12	92	0	100	0	0	–	–
compas	2	99	99	88	98	97	69	88	0	98	0	0	99	99
	4	98	98	86	97	97	47	87	0	95	0	0	94	85
	8	98	98	88	93	85	37	89	11	98	1	0	48	0
	16	98	98	91	89	79	42	91	23	97	1	0	–	–
	32	100	100	78	73	77	44	95	35	99	0	0	–	–
gmc	2	99	99	99	99	99	99	99	0	97	1	0	99	97
	4	100	100	99	99	99	100	100	0	98	3	1	100	98
	8	99	99	97	91	91	99	99	1	98	1	1	91	2
	16	99	99	99	99	99	99	99	0	99	3	0	–	–
	32	98	98	96	93	93	98	98	0	98	1	1	–	–
heloc	2	100	100	100	100	100	100	100	24	100	100	1	87	99
	4	100	100	100	100	100	100	100	41	100	100	1	24	99
	8	100	100	100	99	99	100	100	51	100	96	1	11	0
	16	100	100	100	100	100	100	100	65	100	94	1	–	–
	32	100	100	100	100	100	100	100	64	100	87	4	–	–

Table 22: MNIST (3 vs 8, $n = 784$, depth 8, width 256, 150 factuals), ℓ_2 objective. Greedy* did not run at this scale; under ℓ_1 the norm-adapted GS* and PGD-bisect* runs cover 70 of the 150 factuals (validity over the arm's own rows). PBMC3k is reported in the main text, after the shared marker completion. DASH = the CPU implementation; NA = no valid counterfactual, or a metric the run did not record.

dataset	method	success %	valid %	ℓ_2	l0	redund.	yNN	yNN-dist	time (s)
mnist	DASH	**100**	**100**	1.196	473.0	71.5	0.000	5.139	74.1663
mnist	DASH + data seed	**100**	**100**	1.197	473.5	75.0	0.000	5.128	73.5537
mnist	Wachter	**100**	**100**	1.398	764.0	387.5	0.000	5.136	2.2161
mnist	DiCE	**100**	**100**	1.437	201.5	**1.0**	0.000	5.153	**0.0455**
mnist	GS*	**100**	**100**	3.375	37.0	24.0	0.000	6.109	2.2197
mnist	PGD-bisect*	**100**	**100**	1.102	478.0	62.0	0.000	**5.079**	5.2108
mnist	FMN*	55	55	1.277	473.0	34.0	0.000	5.163	0.8296
mnist	NICE	**100**	**100**	2.959	**22.0**	6.0	**0.200**	5.626	0.7318
mnist	FAB	**100**	**100**	**1.096**	476.0	32.0	0.000	**5.079**	0.1364
mnist	DeepFool	100	0	NA	NA	NA	NA	NA	NA

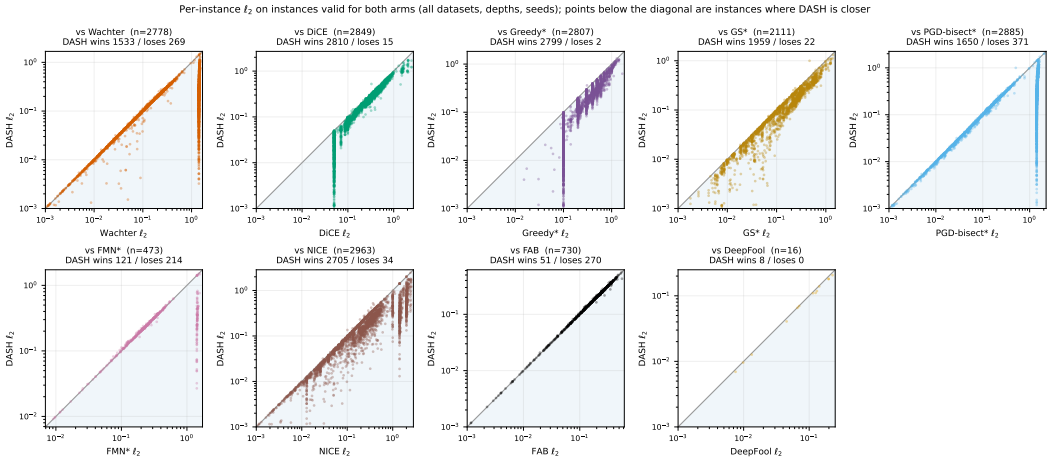

Figure 12: Per-instance paired comparison, ℓ_2 objective (see Figure 8 for the definition).

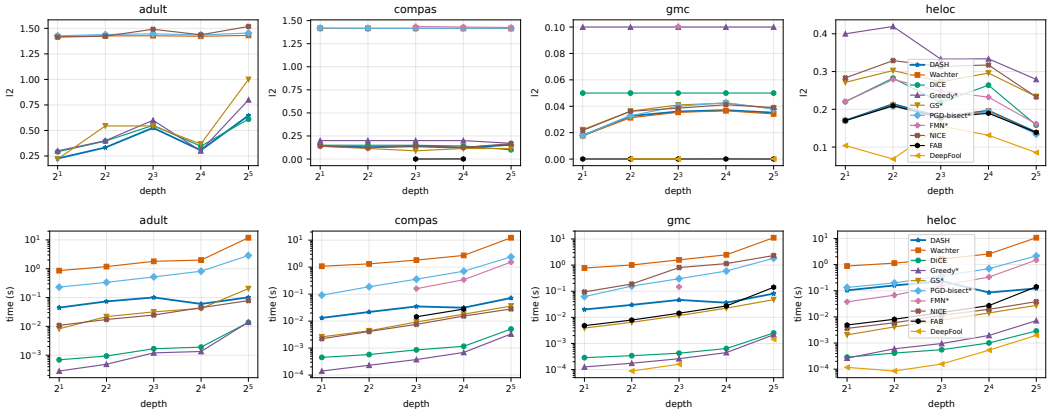

Figure 13: Median ℓ_2 and runtime versus depth, ℓ_2 objective.

Table 23: All tabular cells, ℓ_2 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 20) and the best-known table (Table 19).

dataset	depth	method	success %	valid %	ℓ_2	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_∞
adult	2	DASH	**100**	**100**	**0.225**	3.0	**0.0**	**0.000**	**0.333**	0.0448	0.287	0.216
adult	2	DASH + data seed	**100**	**100**	0.227	3.0	**0.0**	**0.000**	**0.333**	0.0502	**0.282**	0.218
adult	2	Wachter	97	96	1.420	5.0	3.0	**0.000**	1.425	0.8565	2.224	1.000
adult	2	DiCE	**100**	99	0.292	3.0	**0.0**	**0.000**	0.408	0.0007	0.500	**0.190**
adult	2	Greedy*	**100**	**100**	0.300	**1.0**	**0.0**	**0.000**	0.400	**0.0003**	0.300	0.300
adult	2	GS*	100	41	0.220	1.0	0.0	0.000	0.296	0.0082	0.220	0.220
adult	2	PGD-bisect*	**100**	**100**	1.427	4.0	2.0	**0.000**	1.429	0.2302	2.264	1.000
adult	2	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	2	NICE	**100**	**100**	1.414	2.0	**0.0**	**0.000**	0.589	0.0108	2.000	1.000
adult	2	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	2	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	2	MIP-EXP	**100**	**100**	**0.225**	3.0	**0.0**	**0.000**	**0.333**	4.4854	0.300	0.215
adult	2	MIP-OBJ	99	99	0.227	3.0	**0.0**	**0.000**	0.340	1.1407	0.306	0.217
adult	4	DASH	**100**	**100**	0.333	4.0	**0.0**	0.000	**0.537**	0.0734	0.503	0.298
adult	4	DASH + data seed	**100**	**100**	0.333	4.0	**0.0**	0.000	**0.537**	0.0831	0.503	0.298
adult	4	Wachter	91	88	1.426	5.0	3.0	0.000	1.432	1.1878	2.351	1.000
adult	4	DiCE	**100**	97	0.396	3.0	**0.0**	0.000	0.615	0.0009	0.747	**0.224**
adult	4	Greedy*	**100**	**100**	0.400	**1.0**	**0.0**	0.000	0.648	**0.0005**	**0.400**	0.400
adult	4	GS*	100	57	0.544	1.0	0.0	0.000	0.645	0.0218	0.544	0.544
adult	4	PGD-bisect*	**100**	**100**	1.440	5.0	3.0	0.000	1.446	0.3361	2.440	1.000
adult	4	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	4	NICE	**100**	**100**	1.424	2.0	**0.0**	0.200	0.983	0.0173	2.198	1.000
adult	4	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	4	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	4	MIP-EXP	88	88	**0.307**	4.0	1.0	0.000	0.683	60.0004	0.516	0.282
adult	4	MIP-OBJ	**100**	**100**	3.162	8.0	7.0	**0.400**	2.105	60.0061	10.045	1.000
adult	8	DASH	**100**	**100**	**0.526**	4.0	**0.0**	0.000	0.904	0.1010	0.930	0.406
adult	8	DASH + data seed	**100**	**100**	**0.526**	4.0	**0.0**	0.000	0.903	0.1046	0.926	0.406
adult	8	Wachter	79	77	1.428	5.0	3.0	0.000	1.430	1.8042	2.393	1.000
adult	8	DiCE	99	92	0.550	4.0	**0.0**	0.000	**0.853**	0.0017	1.096	**0.287**
adult	8	Greedy*	**100**	82	0.600	**1.0**	**0.0**	0.000	0.933	**0.0012**	**0.600**	0.600
adult	8	GS*	100	17	0.544	2.0	0.0	0.000	0.941	0.0319	0.596	0.525
adult	8	PGD-bisect*	89	89	1.447	5.0	3.0	0.000	1.448	0.5200	2.529	1.000
adult	8	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	8	NICE	**100**	**100**	1.491	3.0	**0.0**	**0.200**	0.897	0.0247	2.670	1.000
adult	8	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	8	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	8	MIP-EXP	3	3	0.016	4.0	0.0	0.200	0.066	14.4452	0.028	0.013
adult	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	16	DASH	**100**	**100**	0.305	4.0	**0.0**	0.000	0.637	0.0600	0.459	0.260
adult	16	DASH + data seed	**100**	**100**	0.305	4.0	**0.0**	0.000	0.637	0.0663	0.451	0.261
adult	16	Wachter	92	87	1.423	5.0	3.0	0.000	1.430	1.9960	2.307	1.000
adult	16	DiCE	**100**	91	0.346	3.0	**0.0**	0.000	**0.584**	0.0019	0.598	**0.191**
adult	16	Greedy*	**100**	91	**0.300**	**1.0**	**0.0**	0.000	0.607	**0.0013**	**0.300**	0.300
adult	16	GS*	100	48	0.368	1.0	0.0	0.000	0.601	0.0420	0.491	0.323
adult	16	PGD-bisect*	97	97	1.436	5.0	3.0	0.000	1.445	0.8194	2.378	1.000
adult	16	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	16	NICE	**100**	**100**	1.437	3.0	**0.0**	**0.200**	0.857	0.0445	2.292	1.000
adult	16	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	32	DASH	**100**	**100**	0.646	4.0	**0.0**	0.000	1.041	0.1015	1.134	0.474
adult	32	DASH + data seed	**100**	**100**	0.646	4.0	**0.0**	0.000	1.045	0.1078	1.134	0.474
adult	32	Wachter	82	82	1.431	5.0	3.0	0.000	1.431	11.8893	2.440	1.000
adult	32	DiCE	97	89	**0.611**	4.0	**0.0**	0.000	0.926	**0.0137**	1.221	**0.324**
adult	32	Greedy*	**100**	83	0.800	**1.0**	**0.0**	0.000	1.026	0.0140	**0.800**	0.800
adult	32	GS*	100	12	1.000	2.0	0.0	0.000	1.149	0.2037	1.000	1.000
adult	32	PGD-bisect*	92	92	1.454	5.0	3.0	0.000	1.450	2.8793	2.618	1.000
adult	32	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	32	NICE	**100**	**100**	1.519	3.0	**0.0**	**0.200**	**0.785**	0.0782	2.709	1.000
adult	32	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	32	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	DASH	99	**99**	**0.142**	2.0	**0.0**	0.400	**0.028**	0.0131	0.157	**0.139**
compas	2	DASH + data seed	99	**99**	**0.142**	2.0	**0.0**	0.400	**0.028**	0.0141	0.157	**0.139**
compas	2	Wachter	88	88	1.418	4.0	2.0	**0.800**	0.060	1.0800	2.127	1.000
compas	2	DiCE	98	98	0.150	**1.0**	**0.0**	**0.800**	**0.029**	0.0005	0.185	0.150
compas	2	Greedy*	**100**	97	0.200	**1.0**	**0.0**	**0.800**	0.037	**0.0001**	0.200	0.200
compas	2	GS*	100	69	0.139	1.0	0.0	0.600	0.027	0.0026	0.158	0.139
compas	2	PGD-bisect*	88	88	1.419	3.0	1.0	**0.800**	0.064	0.0914	2.130	1.000
compas	2	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	NICE	98	98	0.142	**1.0**	**0.0**	0.600	**0.028**	0.0022	**0.156**	0.141

Continued on next page

Table 23: All tabular cells, ℓ_2 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 20) and the best-known table (Table 19).

dataset	depth	method	success %	valid %	ℓ_2	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_∞
compas	2	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	DeepFool	99	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	MIP-EXP	99	**99**	**0.141**	2.0	**0.0**	0.400	**0.028**	0.9779	0.158	**0.139**
compas	2	MIP-OBJ	99	**99**	**0.141**	2.0	**0.0**	0.400	**0.028**	1.3533	0.157	**0.139**
compas	4	DASH	**98**	**98**	0.128	**1.0**	**0.0**	0.600	0.026	0.0218	0.140	0.125
compas	4	DASH + data seed	**98**	**98**	0.128	**1.0**	**0.0**	0.600	0.026	0.0221	0.140	0.125
compas	4	Wachter	86	86	1.417	4.0	2.0	**1.000**	0.058	1.3197	2.089	1.000
compas	4	DiCE	97	97	0.149	**1.0**	**0.0**	0.800	0.027	0.0006	0.150	0.149
compas	4	Greedy*	**98**	97	0.200	**1.0**	**0.0**	**1.000**	0.028	**0.0002**	0.200	0.200
compas	4	GS*	100	47	0.114	1.0	0.0	0.800	0.021	0.0044	0.137	0.114
compas	4	PGD-bisect*	87	87	1.417	3.0	1.0	**1.000**	0.055	0.1873	2.092	1.000
compas	4	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	NICE	95	95	0.122	**1.0**	**0.0**	0.800	**0.024**	0.0041	0.137	0.122
compas	4	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	MIP-EXP	94	94	**0.112**	2.0	**0.0**	0.600	0.026	10.1659	**0.134**	**0.112**
compas	4	MIP-OBJ	85	85	2.000	4.0	2.0	0.800	0.102	60.0057	4.000	1.000
compas	8	DASH	**98**	**98**	**0.138**	2.0	**0.0**	0.400	**0.025**	0.0349	**0.149**	**0.133**
compas	8	DASH + data seed	**98**	**98**	**0.138**	2.0	**0.0**	0.400	**0.026**	0.0364	**0.149**	**0.133**
compas	8	Wachter	88	88	1.417	4.0	2.0	0.800	0.117	1.8195	2.092	1.000
compas	8	DiCE	93	93	0.149	**1.0**	**0.0**	0.800	**0.026**	0.0009	**0.149**	0.149
compas	8	Greedy*	97	85	0.200	**1.0**	**0.0**	**1.000**	0.029	**0.0004**	0.200	0.200
compas	8	GS*	100	37	0.089	1.0	0.0	0.800	0.021	0.0092	0.101	0.080
compas	8	PGD-bisect*	89	89	1.417	3.0	2.0	0.800	0.077	0.3575	2.093	1.000
compas	8	FMN*	11	11	1.437	3.0	2.0	0.800	0.330	0.1593	2.275	1.000
compas	8	NICE	**98**	**98**	0.144	**1.0**	**0.0**	0.600	**0.026**	0.0076	0.156	0.141
compas	8	FAB	99	1	0.000	5.0	0.0	0.600	0.003	0.0145	0.000	0.000
compas	8	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	8	MIP-EXP	48	48	0.051	2.0	0.0	0.600	0.013	7.4927	0.053	0.051
compas	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	16	DASH	**98**	**98**	**0.122**	2.0	**0.0**	0.600	0.031	0.0314	**0.134**	0.117
compas	16	DASH + data seed	**98**	**98**	**0.122**	2.0	**0.0**	0.600	0.031	0.0331	**0.134**	0.117
compas	16	Wachter	91	91	1.416	4.0	2.0	**0.800**	0.129	2.7134	2.076	1.000
compas	16	DiCE	89	89	0.141	**1.0**	**0.0**	**0.800**	**0.027**	**0.0012**	0.143	0.141
compas	16	Greedy*	96	79	0.200	1.0	0.0	1.000	0.028	0.0007	0.200	0.200
compas	16	GS*	100	42	0.114	1.0	0.0	0.800	0.021	0.0180	0.114	0.114
compas	16	PGD-bisect*	91	91	1.417	3.0	2.0	**0.800**	0.114	0.7051	2.096	1.000
compas	16	FMN*	23	23	1.431	3.0	2.0	0.800	0.303	0.3404	2.225	1.000
compas	16	NICE	97	97	0.126	**1.0**	**0.0**	0.600	**0.027**	0.0154	0.139	**0.115**
compas	16	FAB	94	1	0.001	4.0	0.0	0.000	0.001	0.0279	0.002	0.001
compas	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	32	DASH	**100**	**100**	**0.155**	2.0	**0.0**	0.600	**0.030**	0.0717	**0.166**	**0.152**
compas	32	DASH + data seed	**100**	**100**	**0.155**	2.0	**0.0**	0.600	0.030	0.0742	**0.166**	0.155
compas	32	Wachter	78	78	1.415	4.0	2.0	0.800	0.091	12.2830	2.062	1.000
compas	32	DiCE	73	73	0.100	1.0	0.0	0.800	0.023	0.0051	0.100	0.100
compas	32	Greedy*	86	77	0.171	1.0	0.0	1.000	0.027	0.0033	0.200	0.150
compas	32	GS*	100	44	0.114	1.0	0.0	0.800	0.024	0.0361	0.114	0.114
compas	32	PGD-bisect*	95	95	1.417	3.0	2.0	**0.800**	0.173	2.3955	2.105	1.000
compas	32	FMN*	35	35	1.426	3.0	2.0	0.800	0.262	1.5389	2.218	1.000
compas	32	NICE	99	99	0.167	**1.0**	**0.0**	**0.800**	**0.029**	**0.0281**	0.168	0.154
compas	32	FAB	57	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	32	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	2	DASH	**99**	**99**	**0.018**	4.0	**0.0**	0.600	**0.034**	0.0200	**0.019**	**0.018**
gmc	2	DASH + data seed	**99**	**99**	**0.018**	4.0	**0.0**	0.600	**0.034**	0.0253	**0.019**	**0.018**
gmc	2	Wachter	**99**	**99**	**0.018**	4.0	**0.0**	0.600	**0.034**	0.7730	**0.020**	**0.018**
gmc	2	DiCE	**99**	**99**	0.050	**1.0**	**0.0**	0.800	0.057	0.0003	0.050	0.050
gmc	2	Greedy*	**99**	**99**	0.100	**1.0**	**0.0**	**1.000**	0.098	**0.0001**	0.100	0.100
gmc	2	GS*	**99**	**99**	0.022	**1.0**	**0.0**	0.600	0.036	0.0039	0.024	0.021
gmc	2	PGD-bisect*	**99**	**99**	**0.018**	4.0	1.0	0.600	**0.034**	0.0621	0.021	**0.017**
gmc	2	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	2	NICE	97	97	0.022	**1.0**	**0.0**	0.600	**0.035**	0.0941	0.022	0.019
gmc	2	FAB	100	1	0.000	6.0	0.0	0.800	0.016	0.0048	0.000	0.000
gmc	2	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	2	MIP-EXP	**99**	**99**	**0.018**	4.0	1.0	0.600	**0.034**	0.4926	**0.020**	**0.018**
gmc	2	MIP-OBJ	97	97	**0.018**	4.0	**0.0**	0.600	**0.034**	0.4722	**0.019**	**0.017**
gmc	4	DASH	**100**	**100**	0.033	3.0	**0.0**	0.600	**0.046**	0.0300	**0.037**	0.032
gmc	4	DASH + data seed	**100**	**100**	0.033	3.0	**0.0**	0.600	**0.046**	0.0361	**0.037**	0.032
gmc	4	Wachter	99	99	**0.031**	4.0	**0.0**	0.600	**0.045**	1.0089	**0.038**	**0.031**
gmc	4	DiCE	99	99	0.050	**1.0**	**0.0**	0.800	0.058	0.0003	0.050	0.050
gmc	4	Greedy*	99	99	0.100	**1.0**	**0.0**	**1.000**	0.098	**0.0002**	0.100	0.100

Continued on next page

Table 23: All tabular cells, ℓ_2 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 20) and the best-known table (Table 19).

dataset	depth	method	success %	valid %	ℓ_2	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_∞
gmc	4	GS*	**100**	**100**	0.036	**1.0**	**0.0**	0.600	0.046	0.0065	**0.038**	0.035
gmc	4	PGD-bisect*	**100**	**100**	0.033	4.0	1.0	0.600	**0.045**	0.1570	0.041	**0.031**
gmc	4	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	4	NICE	98	98	0.036	**1.0**	**0.0**	0.800	0.046	0.1889	0.038	0.035
gmc	4	FAB	100	3	0.000	7.0	0.0	0.800	0.015	0.0079	0.000	0.000
gmc	4	DeepFool	100	1	0.000	9.0	9.0	0.600	0.022	0.0001	0.000	0.000
gmc	4	MIP-EXP	**100**	**100**	0.032	3.0	1.0	0.600	**0.045**	1.6123	**0.037**	**0.031**
gmc	4	MIP-OBJ	98	98	0.045	3.0	**0.0**	0.800	0.059	60.0045	0.057	0.044
gmc	8	DASH	**99**	**99**	0.036	3.0	**0.0**	0.600	0.048	0.0465	0.043	0.035
gmc	8	DASH + data seed	**99**	**99**	0.036	3.0	**0.0**	0.600	0.048	0.0534	0.043	0.035
gmc	8	Wachter	97	97	0.035	3.0	**0.0**	0.600	0.047	1.5871	0.044	0.034
gmc	8	DiCE	91	91	0.050	**1.0**	**0.0**	0.800	0.056	0.0004	0.050	0.050
gmc	8	Greedy*	91	91	0.100	**1.0**	**0.0**	**1.000**	0.095	**0.0003**	0.100	0.100
gmc	8	GS*	**99**	**99**	0.041	**1.0**	**0.0**	0.800	0.049	0.0117	0.045	0.037
gmc	8	PGD-bisect*	**99**	**99**	0.040	3.0	1.0	0.800	0.047	0.3008	0.056	0.034
gmc	8	FMN*	1	1	0.100	4.0	1.0	0.800	0.080	0.1472	0.149	0.089
gmc	8	NICE	98	98	0.039	**1.0**	**0.0**	0.800	0.046	0.8054	0.049	0.038
gmc	8	FAB	100	1	0.000	5.0	0.0	0.900	0.017	0.0144	0.000	0.000
gmc	8	DeepFool	100	1	0.000	8.0	8.0	0.800	0.017	0.0002	0.000	0.000
gmc	8	MIP-EXP	91	91	**0.030**	3.0	1.0	0.800	**0.044**	5.1556	**0.041**	**0.028**
gmc	8	MIP-OBJ	2	2	0.333	2.0	0.0	0.000	0.337	1.7697	0.457	0.286
gmc	16	DASH	**99**	**99**	**0.037**	3.0	**0.0**	0.800	**0.049**	0.0363	**0.043**	0.037
gmc	16	DASH + data seed	**99**	**99**	**0.037**	3.0	**0.0**	0.800	**0.049**	0.0434	**0.043**	0.037
gmc	16	Wachter	**99**	**99**	**0.037**	3.0	**0.0**	0.800	**0.049**	2.4798	0.045	**0.036**
gmc	16	DiCE	**99**	**99**	0.050	**1.0**	**0.0**	0.800	0.059	0.0006	0.050	0.050
gmc	16	Greedy*	**99**	**99**	0.100	**1.0**	**0.0**	**1.000**	0.097	**0.0004**	0.100	0.100
gmc	16	GS*	**99**	**99**	0.042	**1.0**	**0.0**	0.800	0.052	0.0224	0.049	0.038
gmc	16	PGD-bisect*	**99**	**99**	0.043	3.0	1.0	0.800	0.051	0.5922	0.060	**0.036**
gmc	16	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	16	NICE	**99**	**99**	0.041	**1.0**	**0.0**	0.800	0.053	1.1421	0.047	0.040
gmc	16	FAB	100	3	0.000	7.0	0.0	0.900	0.027	0.0272	0.000	0.000
gmc	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	32	DASH	**98**	**98**	**0.035**	3.0	**0.0**	0.800	**0.046**	0.0809	0.041	0.035
gmc	32	DASH + data seed	**98**	**98**	**0.035**	3.0	**0.0**	0.800	**0.046**	0.0888	**0.040**	0.035
gmc	32	Wachter	96	96	**0.034**	4.0	**0.0**	0.800	**0.045**	11.1648	**0.040**	**0.034**
gmc	32	DiCE	93	93	0.050	**1.0**	**0.0**	0.800	0.055	0.0025	0.050	0.050
gmc	32	Greedy*	93	93	0.100	**1.0**	**0.0**	**1.000**	0.094	**0.0021**	0.100	0.100
gmc	32	GS*	**98**	**98**	0.038	**1.0**	**0.0**	0.800	0.048	0.0475	0.046	0.037
gmc	32	PGD-bisect*	**98**	**98**	0.038	3.0	1.0	0.800	0.046	1.8329	0.054	0.035
gmc	32	FMN*	0	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	32	NICE	**98**	**98**	0.039	**1.0**	**0.0**	0.800	0.047	2.3254	0.047	0.038
gmc	32	FAB	99	1	0.000	5.0	0.0	0.800	0.019	0.1414	0.000	0.000
gmc	32	DeepFool	100	1	0.000	9.0	9.0	0.600	0.018	0.0014	0.000	0.000
heloc	2	DASH	**100**	**100**	0.170	20.0	1.0	0.000	0.310	0.1036	0.603	0.087
heloc	2	DASH + data seed	**100**	**100**	0.170	20.0	1.0	0.000	0.310	0.1137	0.603	0.084
heloc	2	Wachter	**100**	**100**	0.170	22.0	2.0	0.000	0.309	0.8812	0.632	**0.075**
heloc	2	DiCE	**100**	**100**	0.220	8.0	**0.0**	0.200	0.326	0.0003	0.699	0.100
heloc	2	Greedy*	**100**	**100**	0.400	**2.0**	**0.0**	0.200	0.438	**0.0003**	**0.430**	0.350
heloc	2	GS*	**100**	**100**	0.271	7.0	1.0	0.200	0.338	0.0020	0.688	0.170
heloc	2	PGD-bisect*	**100**	**100**	0.170	21.0	**0.0**	0.000	0.307	0.1338	0.614	0.081
heloc	2	FMN*	24	24	0.220	21.0	0.0	0.000	0.308	0.0377	0.827	0.099
heloc	2	NICE	**100**	**100**	0.283	6.0	**0.0**	**0.400**	**0.286**	0.0036	0.576	0.195
heloc	2	FAB	**100**	**100**	0.170	21.0	**0.0**	0.000	0.308	0.0049	0.617	0.081
heloc	2	DeepFool	100	1	0.104	23.0	14.5	0.100	0.323	0.0001	0.402	0.050
heloc	2	MIP-EXP	87	87	**0.166**	21.0	8.0	0.200	0.310	60.0025	0.603	**0.074**
heloc	2	MIP-OBJ	99	99	0.169	21.0	**0.0**	0.000	0.306	35.0802	0.608	0.080
heloc	4	DASH	**100**	**100**	0.214	21.0	1.0	0.000	0.316	0.1595	0.789	0.094
heloc	4	DASH + data seed	**100**	**100**	0.214	21.0	1.0	0.000	0.317	0.1657	0.796	0.094
heloc	4	Wachter	**100**	**100**	0.211	23.0	2.0	0.000	0.316	1.1422	0.801	**0.088**
heloc	4	DiCE	**100**	**100**	0.283	9.0	**0.0**	0.100	0.333	**0.0004**	0.885	0.100
heloc	4	Greedy*	**100**	**100**	0.419	**3.0**	**0.0**	0.200	0.422	0.0006	**0.586**	0.400
heloc	4	GS*	**100**	**100**	0.303	8.0	**0.0**	0.200	0.358	0.0041	0.804	0.188
heloc	4	PGD-bisect*	**100**	**100**	**0.209**	21.0	**0.0**	0.000	0.315	0.1989	0.782	0.091
heloc	4	FMN*	41	41	0.279	21.0	0.0	0.000	0.324	0.0676	1.030	0.120
heloc	4	NICE	**100**	**100**	0.329	7.0	1.0	**0.400**	**0.284**	0.0060	0.723	0.212
heloc	4	FAB	**100**	**100**	**0.209**	21.0	**0.0**	0.000	0.314	0.0081	0.779	**0.089**
heloc	4	DeepFool	100	1	0.068	23.0	8.0	0.600	0.250	0.0001	0.275	0.025
heloc	4	MIP-EXP	24	24	0.086	21.0	14.5	0.400	0.263	60.0043	0.325	0.035
heloc	4	MIP-OBJ	99	99	0.822	21.0	10.0	0.200	0.662	60.0050	3.003	0.392
heloc	8	DASH	**100**	**100**	0.180	21.0	1.0	0.000	0.302	0.2332	0.666	0.080

Continued on next page

Table 23: All tabular cells, ℓ_2 objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 20) and the best-known table (Table 19).

dataset	depth	method	success %	valid %	ℓ_2	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_∞
heloc	8	DASH + data seed	**100**	**100**	0.180	21.0	1.0	0.000	0.302	0.2461	0.664	0.081
heloc	8	Wachter	**100**	**100**	**0.178**	22.0	1.0	0.000	0.303	1.6352	0.675	0.070
heloc	8	DiCE	99	99	0.218	9.0	**0.0**	0.200	0.318	**0.0006**	0.700	**0.050**
heloc	8	Greedy*	99	99	0.333	**3.0**	**0.0**	0.200	0.384	0.0010	**0.500**	0.300
heloc	8	GS*	**100**	**100**	0.273	7.0	**0.0**	0.200	0.357	0.0075	0.711	0.168
heloc	8	PGD-bisect*	**100**	**100**	**0.178**	21.0	**0.0**	0.000	0.301	0.3650	0.661	0.074
heloc	8	FMN*	51	51	0.248	21.0	0.0	0.000	0.323	0.1556	0.958	0.117
heloc	8	NICE	**100**	**100**	0.314	6.0	1.0	**0.400**	**0.294**	0.0114	0.627	0.208
heloc	8	FAB	98	96	**0.178**	21.0	**0.0**	0.000	0.302	0.0147	0.658	0.075
heloc	8	DeepFool	100	1	0.156	23.0	13.5	0.200	0.360	0.0002	0.629	0.062
heloc	8	MIP-EXP	11	11	0.021	20.0	4.0	0.400	0.252	4.6867	0.078	0.011
heloc	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
heloc	16	DASH	**100**	**100**	0.196	21.0	1.0	0.000	0.313	0.0854	0.739	0.088
heloc	16	DASH + data seed	**100**	**100**	0.198	20.0	1.0	0.000	0.307	0.0983	0.750	0.087
heloc	16	Wachter	**100**	**100**	0.195	22.0	2.0	0.000	0.306	2.5564	0.733	**0.076**
heloc	16	DiCE	**100**	**100**	0.264	9.0	**0.0**	0.200	0.328	**0.0010**	0.782	0.099
heloc	16	Greedy*	**100**	**100**	0.334	**3.0**	**0.0**	0.200	0.366	0.0019	**0.526**	0.300
heloc	16	GS*	**100**	**100**	0.296	8.0	**0.0**	0.200	0.349	0.0142	0.765	0.182
heloc	16	PGD-bisect*	**100**	**100**	0.193	21.0	**0.0**	0.000	0.307	0.7008	0.712	0.084
heloc	16	FMN*	65	65	0.232	21.0	0.0	0.000	0.322	0.3340	0.897	0.103
heloc	16	NICE	**100**	**100**	0.317	6.0	1.0	**0.400**	**0.291**	0.0195	0.661	0.205
heloc	16	FAB	97	94	**0.190**	21.0	1.0	0.000	0.305	0.0270	0.699	0.083
heloc	16	DeepFool	100	1	0.131	23.0	2.0	0.400	0.425	0.0005	0.537	0.058
heloc	32	DASH	**100**	**100**	0.139	21.0	1.0	0.000	0.299	0.1226	0.502	0.059
heloc	32	DASH + data seed	**100**	**100**	0.140	21.0	1.0	0.000	0.299	0.1407	0.508	0.059
heloc	32	Wachter	**100**	**100**	0.137	22.0	2.0	0.100	0.300	10.6565	0.514	0.057
heloc	32	DiCE	**100**	**100**	0.158	8.0	**0.0**	0.200	0.316	**0.0029**	0.500	**0.050**
heloc	32	Greedy*	**100**	**100**	0.279	**3.0**	**0.0**	0.200	0.354	0.0072	**0.400**	0.200
heloc	32	GS*	**100**	**100**	0.234	7.0	**0.0**	0.200	0.333	0.0275	0.547	0.131
heloc	32	PGD-bisect*	**100**	**100**	**0.134**	21.0	**0.0**	0.000	0.301	2.1554	0.493	0.057
heloc	32	FMN*	64	64	0.161	21.0	0.0	0.000	0.314	1.4984	0.619	0.072
heloc	32	NICE	**100**	**100**	0.233	6.0	1.0	**0.400**	**0.275**	0.0383	0.482	0.176
heloc	32	FAB	**100**	87	0.139	21.0	**0.0**	0.000	0.294	0.1418	0.491	0.061
heloc	32	DeepFool	99	4	0.085	23.0	15.5	0.500	0.331	0.0020	0.336	0.034

D.4 ℓ_∞ OBJECTIVE

Table 24: CARLA axes at depth 8 (ℓ_∞ objective, basic actionability; seeds 17/29/43).

dataset	method	success %	valid %	ℓ_∞	l0	redund.	yNN	yNN-dist	time (s)
adult	DASH	**100**	**100**	0.291	4.0	**0.0**	0.000	0.950	0.0997
	DASH + data seed	**100**	**100**	0.291	4.0	**0.0**	0.000	0.950	0.1052
	Wachter	78	77	1.000	5.0	3.0	0.000	1.430	1.2503
	DiCE	**100**	83	**0.276**	4.0	**0.0**	0.000	**0.765**	0.0017
	Greedy*	**100**	82	0.600	**1.0**	**0.0**	0.000	0.933	**0.0012**
	GS*	100	25	0.321	2.0	0.0	0.000	0.484	0.0086
	PGD-bisect*	73	73	1.000	5.0	4.0	0.000	1.429	0.3121
	FMN*	70	70	1.000	5.0	4.0	0.000	1.429	0.1756
	NICE	**100**	**100**	1.000	3.0	**0.0**	**0.600**	1.417	0.0253
	FAB	100	0	NA	NA	NA	NA	NA	NA
	DeepFool	99	0	NA	NA	NA	NA	NA	NA
	MIP-EXP	7	7	0.042	4.0	0.0	0.100	0.117	42.0766
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA
compas	DASH	**98**	**98**	**0.129**	2.0	**0.0**	0.400	**0.023**	0.0329
	DASH + data seed	**98**	**98**	**0.129**	2.0	**0.0**	0.400	**0.023**	0.0344
	Wachter	89	89	1.000	4.0	2.0	0.800	0.072	1.6603
	DiCE	**98**	93	0.149	**1.0**	**0.0**	0.800	0.026	0.0009
	Greedy*	97	85	0.200	**1.0**	**0.0**	**1.000**	0.029	**0.0004**
	GS*	100	35	0.101	1.0	0.0	1.000	0.014	0.0137
	PGD-bisect*	87	87	1.000	3.0	2.0	0.800	0.161	0.2235
	FMN*	85	85	1.000	3.0	2.0	0.800	0.161	0.1702
	NICE	**98**	**98**	0.141	**1.0**	**0.0**	0.600	0.025	0.0076
	FAB	99	0	NA	NA	NA	NA	NA	NA
	DeepFool	100	0	NA	NA	NA	NA	NA	NA
	MIP-EXP	48	48	0.050	2.0	0.0	0.600	0.009	6.6603
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA
gmc	DASH	**99**	**99**	0.030	3.0	**0.0**	0.800	0.045	0.0459
	DASH + data seed	**99**	**99**	0.030	3.0	**0.0**	0.800	0.045	0.0520
	Wachter	97	97	0.035	3.5	1.0	0.600	0.048	1.2837
	DiCE	92	92	0.050	**1.0**	**0.0**	0.800	0.056	0.0005
	Greedy*	91	91	0.100	**1.0**	**0.0**	**1.000**	0.095	**0.0003**
	GS*	**99**	**99**	0.035	**1.0**	**0.0**	0.800	0.049	0.0139
	PGD-bisect*	**99**	**99**	0.030	3.0	**0.0**	0.800	0.045	0.1754
	FMN*	64	64	0.026	3.0	0.0	0.800	0.045	0.1625
	NICE	98	98	0.036	**1.0**	**0.0**	0.800	0.047	0.8128
	FAB	100	1	0.000	8.5	0.0	0.900	0.017	0.0143
	DeepFool	100	0	NA	NA	NA	NA	NA	NA
	MIP-EXP	90	90	**0.023**	3.0	1.0	0.800	**0.041**	3.1355
	MIP-OBJ	2	2	0.243	2.0	0.0	0.000	0.347	1.7662
heloc	DASH	**100**	**100**	0.047	21.0	**0.0**	0.000	0.324	0.2311
	DASH + data seed	**100**	**100**	0.047	21.0	**0.0**	0.000	0.322	0.2411
	Wachter	**100**	**100**	0.054	22.0	4.0	0.000	**0.309**	1.3847
	DiCE	99	99	0.050	9.0	**0.0**	**0.200**	0.317	**0.0006**
	Greedy*	99	99	0.300	**3.0**	**0.0**	**0.200**	0.384	0.0010
	GS*	**100**	**100**	0.178	5.5	1.0	**0.200**	0.395	0.0095
	PGD-bisect*	**100**	**100**	**0.047**	21.0	**0.0**	0.000	0.321	0.2554
	FMN*	**100**	**100**	**0.047**	21.0	**0.0**	0.000	0.325	0.1682
	NICE	**100**	**100**	0.184	5.0	**0.0**	**0.200**	0.311	0.0107
	FAB	98	96	**0.046**	21.0	**0.0**	0.000	0.322	0.0147
	DeepFool	99	1	0.039	23.0	7.5	0.100	0.355	0.0002
	MIP-EXP	59	59	0.031	21.0	0.0	0.200	0.301	11.9084
	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA

success = the returned point crosses the threshold (CARLA's success rate); valid = it also satisfies the actionability box and one-hot constraints. Bold = best per column within a dataset among arms with validity $\geq$ 80%. Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 28) and the best-known table (Table 27).

Table 25: ℓ_∞ relative to the certified optimum (MIP-EXP or MIP-OBJ; Gurobi, 4 threads, 60 s budget per factual, absolute optimality gap 10^{-4}, depths 2/4/8, same 50 factuals per cell as the main run). certified n = instances on which either exact formulation proved an optimum in that dataset; arm valid = how many of those the arm returns a valid point on; at optimum = arm valid and distance $\leq$ ref $+10^{-3}$ (one-sided: below the solver's incumbent is inside its gap); within 5% = arm valid and distance $\leq 1.05\times$ref (purely relative: where the optimum is below 0.02 the absolute tie band is the wider of the two, so the columns are not nested); both percentages over certified n; ratio columns over the arm-valid subset. A valid point can lie up to 6×10^{-4} below the certified optimum (the solver's absolute gap and numerical tolerances), so ratios below 1 occur where the optimum itself is small: every ratio below 0.95 has an optimum under 2.5×10^{-3}.

dataset	arm	certified n	arm valid	at optimum %	within 5 %	median ratio	max ratio
adult	DASH	279	279	93.2	99.3	1.000	1.106
adult	DASH + data seed	279	279	93.5	99.3	1.000	1.106
adult	Wachter	279	259	0.4	0.0	4.419	13.879
adult	DiCE	279	275	1.1	20.8	1.122	130.887
adult	Greedy*	279	279	0.0	0.0	1.957	261.774
adult	GS*	279	191	0.0	0.0	1.849	376.044
adult	PGD-bisect*	279	250	0.0	0.0	6.637	2617.742
adult	FMN*	279	227	0.0	0.0	6.427	2617.742
adult	NICE	279	279	0.0	0.0	4.688	1148.601
adult	FAB	279	0	0.0	0.0	NA	NA
adult	DeepFool	279	0	0.0	0.0	NA	NA
compas	DASH	326	326	85.0	95.1	1.001	1.217
compas	DASH + data seed	326	326	85.0	95.1	1.002	1.217
compas	Wachter	326	306	22.1	23.0	3.190	67.114
compas	DiCE	326	323	0.6	13.5	1.240	200.000
compas	Greedy*	326	318	0.0	4.0	1.667	400.000
compas	GS*	326	205	0.3	3.7	1.446	50.621
compas	PGD-bisect*	326	302	0.3	0.3	13.611	4000.000
compas	FMN*	326	275	0.0	0.0	14.255	4000.000
compas	NICE	326	324	3.1	35.3	1.088	11.172
compas	FAB	326	0	0.0	0.0	NA	NA
compas	DeepFool	326	0	0.0	0.0	NA	NA
gmc	DASH	419	419	82.8	82.3	1.001	1.615
gmc	DASH + data seed	419	419	83.5	82.3	1.001	1.615
gmc	Wachter	419	419	28.6	22.0	1.114	1.736
gmc	DiCE	419	414	0.0	1.7	2.681	583.733
gmc	Greedy*	419	412	0.0	0.7	5.340	1167.466
gmc	GS*	419	419	3.1	6.2	1.395	91.191
gmc	PGD-bisect*	419	419	99.0	98.3	1.001	1.358
gmc	FMN*	419	306	69.2	70.4	1.000	1.169
gmc	NICE	419	412	16.0	4.3	1.223	25.589
gmc	FAB	419	7	1.7	1.7	0.205	0.280
gmc	DeepFool	419	0	0.0	0.0	NA	NA
heloc	DASH	331	331	88.2	96.1	1.002	1.134
heloc	DASH + data seed	331	331	87.0	95.8	1.002	1.134
heloc	Wachter	331	331	0.6	0.3	1.266	6.272
heloc	DiCE	331	331	0.6	1.8	1.622	250.000
heloc	Greedy*	331	331	0.0	0.0	7.707	500.000
heloc	GS*	331	331	0.0	0.0	4.312	352.063
heloc	PGD-bisect*	331	331	100.0	99.7	1.001	1.221
heloc	FMN*	331	330	96.4	99.7	1.000	1.048
heloc	NICE	331	331	0.0	0.0	4.241	316.456
heloc	FAB	331	330	95.8	99.4	1.001	1.051
heloc	DeepFool	331	7	1.2	0.9	1.056	1.112
all	DASH	1355	1355	86.8	92.3	1.001	1.615
all	DASH + data seed	1355	1355	86.8	92.2	1.001	1.615
all	Wachter	1355	1315	14.4	12.4	1.281	67.114
all	DiCE	1355	1343	0.5	8.5	1.502	583.733
all	Greedy*	1355	1340	0.0	1.2	3.653	1167.466
all	GS*	1355	1146	1.0	2.8	2.027	376.044
all	PGD-bisect*	1355	1302	55.1	54.8	1.005	4000.000
all	FMN*	1355	1138	44.9	46.1	1.013	4000.000
all	NICE	1355	1346	5.7	9.8	1.645	1148.601
all	FAB	1355	337	23.9	24.8	1.001	1.051
all	DeepFool	1355	7	0.3	0.2	1.056	1.112

Table 26: MIP-EXP against DASH by depth, ℓ_∞ objective, same queries, pooled over datasets and seeds. Median times are over all of these queries, solved or not (MIP-EXP stops at its 60 s budget).

depth	queries	MIP-EXP valid %	MIP-EXP certified %	MIP-EXP median time (s)	DASH valid %	DASH median time (s)
2	600	99.7	99.7	1.007	99.7	0.031
4	600	96.3	66.5	9.036	99.5	0.059
8	600	51.0	45.0	60.000	99.2	0.073

Table 27: Best-known counterfactual, ℓ_∞ objective: population is every tabular query of this norm (the queries column); best = smallest valid distance among DASH and the nine prior methods on that query (the exact solvers and the ablation are not in the reference); a query counts for an arm at τ iff the arm is valid and its distance $\leq (1+\tau)\times$best; column best uses the absolute tie band best $+10^{-3}$ instead, which is wider than 1% where best is below 0.1; invalid rows never count, no penalty constant; 95% CI = percentile cluster bootstrap over trained models (dataset $\times$ depth $\times$ seed), 2,000 resamples. Sorted by within 5%; bold = best per column.

method	queries	valid %	best	within 1 %	within 5 %	95 % CI (within 5 %)	within 10 %
DASH	3000	**99.4**	**88.6**	**81.4**	**92.3**	89.6–94.7	**95.8**
PGD-bisect*	3000	92.3	49.7	47.2	49.3	36.9–62.0	49.4
FMN*	3000	81.9	39.8	35.9	41.0	30.3–51.4	41.7
FAB	3000	24.4	21.5	19.4	23.9	13.6–34.1	24.3
NICE	3000	98.9	6.1	4.0	11.9	7.6–16.5	17.7
DiCE	3000	93.6	1.2	2.1	11.3	8.6–14.2	20.5
Wachter	3000	92.7	10.1	4.6	10.1	7.2–13.1	17.6
GS*	3000	72.2	0.9	0.5	2.8	1.7–4.0	6.9
Greedy*	3000	93.6	0.1	0.2	1.4	0.7–2.2	2.9
DeepFool	3000	0.4	0.2	0.0	0.1	0.0–0.3	0.3

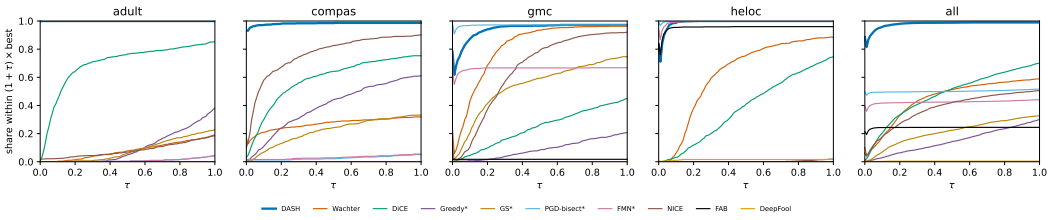

Figure 14: Performance profile, ℓ_∞ objective (see Figure 6 for the definition).

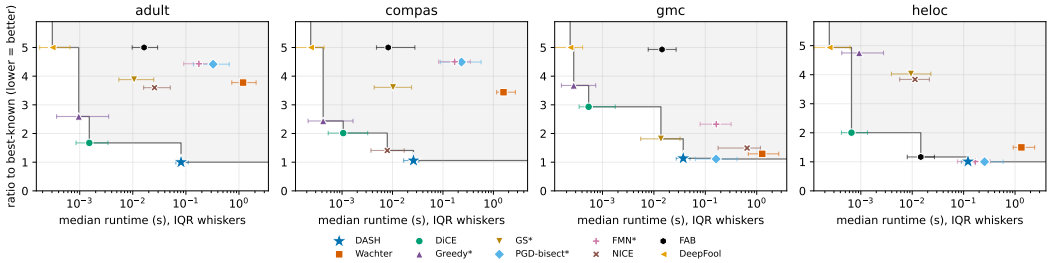

Figure 15: Runtime-quality trade-off, ℓ_∞ objective (see Figure 7 for the definition).

Table 28: DASH against each arm on ℓ_∞, paired per factual over the instances where both are valid: DASH wins = DASH is closer by more than 10^{-3}, losses = the other arm is closer; cells won = (depth, seed) cells ("all" row: dataset, depth, seed cells) with ≥ 1 pair, where the median paired difference favours DASH by more than 10^{-3}, printed as k/m; median ratio = median of DASH distance / other distance (below 1 favours DASH); two-sided Wilcoxon signed-rank on the non-tied pairs. Pairs within a cell share one trained model, so the pooled p-value overstates the evidence; cells won counts model-level agreement.

dataset	vs	n paired	DASH wins	ties	DASH losses	cells won	median ratio	Wilcoxon p
adult	DASH + data seed	750	4	741	5	0/15	1.000	0.73
adult	Wachter	646	643	3	0	15/15	0.256	5.7e-107
adult	DiCE	677	670	6	1	15/15	0.907	1.6e-111
adult	Greedy*	685	685	0	0	15/15	0.468	8e-114
adult	GS*	338	338	0	0	15/15	0.501	3.7e-57
adult	PGD-bisect*	612	612	0	0	15/15	0.185	6.6e-102
adult	FMN*	573	573	0	0	15/15	0.189	1.5e-95
adult	NICE	750	736	14	0	15/15	0.260	3.8e-122
adult	MIP-EXP	310	71	220	19	3/9	1.000	1.8e-11
adult	MIP-OBJ	298	11	269	18	0/6	1.000	0.34
compas	DASH + data seed	740	6	731	3	0/15	1.000	0.73
compas	Wachter	650	522	121	7	15/15	0.289	1.5e-87
compas	DiCE	661	631	19	11	15/15	0.855	1.5e-104
compas	Greedy*	653	648	4	1	15/15	0.672	2.5e-107
compas	GS*	336	334	2	0	15/15	0.704	1.7e-56
compas	PGD-bisect*	664	654	10	0	15/15	0.112	9.2e-109
compas	FMN*	618	609	8	1	15/15	0.110	1.4e-101
compas	NICE	728	651	46	31	15/15	0.936	8.5e-94
compas	MIP-EXP	359	41	272	46	0/9	1.000	0.0048
compas	MIP-OBJ	293	6	241	46	0/6	1.001	9e-05
gmc	DASH + data seed	742	7	722	13	0/15	1.000	0.09
gmc	Wachter	736	474	216	46	14/15	0.918	7.2e-52
gmc	DiCE	722	715	5	2	15/15	0.468	5.6e-119
gmc	Greedy*	720	717	2	1	15/15	0.246	3.5e-119
gmc	GS*	742	649	49	44	15/15	0.794	1e-84
gmc	PGD-bisect*	742	4	585	153	0/15	1.000	3.1e-25
gmc	FMN*	514	12	432	70	0/15	1.001	5.3e-10
gmc	NICE	735	605	117	13	15/15	0.819	2.4e-95
gmc	FAB	13	0	13	0	0/9	5.132	NA
gmc	MIP-EXP	434	10	349	75	0/9	1.000	2.1e-06
gmc	MIP-OBJ	293	3	243	47	0/8	1.002	1.3e-07
heloc	DASH + data seed	750	30	689	31	0/15	1.000	0.92
heloc	Wachter	750	747	2	1	15/15	0.808	4.3e-124
heloc	DiCE	749	747	2	0	15/15	0.625	6.1e-124
heloc	Greedy*	749	749	0	0	15/15	0.155	2.9e-124
heloc	GS*	750	750	0	0	15/15	0.258	2e-124
heloc	PGD-bisect*	750	3	625	122	0/15	1.002	3.5e-19
heloc	FMN*	749	22	629	98	0/15	1.001	1.9e-07
heloc	NICE	750	750	0	0	15/15	0.267	2e-124
heloc	FAB	720	53	590	77	0/15	1.000	0.19
heloc	DeepFool	12	7	5	0	6/9	0.924	0.016
heloc	MIP-EXP	379	37	300	42	0/9	1.000	0.0063
heloc	MIP-OBJ	297	71	193	33	2/6	1.001	4.4e-11
all	DASH + data seed	2982	47	2883	52	0/60	1.000	0.57
all	Wachter	2782	2386	342	54	59/60	0.745	$< 10^{-300}$
all	DiCE	2809	2763	32	14	60/60	0.724	$< 10^{-300}$
all	Greedy*	2807	2799	6	2	60/60	0.339	$< 10^{-300}$
all	GS*	2166	2071	51	44	60/60	0.437	$< 10^{-300}$
all	PGD-bisect*	2768	1273	1220	275	30/60	0.995	5.5e-223
all	FMN*	2454	1216	1069	169	30/60	0.982	4.5e-212
all	NICE	2963	2742	177	44	60/60	0.537	$< 10^{-300}$
all	FAB	733	53	603	77	0/24	1.000	0.19
all	DeepFool	12	7	5	0	6/9	0.924	0.016
all	MIP-EXP	1482	159	1141	182	3/36	1.000	1.4e-06
all	MIP-OBJ	1181	91	946	144	2/26	1.000	0.054

Table 29: Validity (%) per arm and cell, ℓ_∞ objective.

dataset	depth	DASH	DASH + data seed	Wachter	DiCE	Greedy*	GS*	PGD-bisect*	FMN*	NICE	FAB	DeepFool	MIP-EXP	MIP-OBJ
adult	2	100	100	95	99	100	71	92	84	100	0	0	100	99
	4	100	100	88	97	100	61	83	76	100	0	0	100	100
	8	100	100	77	83	82	25	73	70	100	0	0	7	0
	16	100	100	89	91	91	46	85	83	100	0	0	–	–
	32	100	100	82	81	83	23	74	69	100	0	0	–	–
compas	2	99	99	89	97	97	61	87	79	98	0	0	99	99
	4	98	98	87	96	97	51	85	74	95	0	0	92	96
	8	98	98	89	93	85	35	87	85	98	0	0	48	0
	16	98	98	91	93	79	36	90	89	97	0	0	–	–
	32	100	100	79	62	77	41	93	87	99	0	0	–	–
gmc	2	99	99	99	99	99	99	99	73	97	1	0	99	97
	4	100	100	99	99	99	100	100	76	98	3	0	100	97
	8	99	99	97	92	91	99	99	64	98	1	0	90	2
	16	99	99	99	99	99	99	99	63	99	3	0	–	–
	32	98	98	96	92	93	98	98	66	98	1	0	–	–
heloc	2	100	100	100	100	100	100	100	99	100	100	2	100	100
	4	100	100	100	100	100	100	100	100	100	100	1	93	98
	8	100	100	100	99	99	100	100	100	100	96	1	59	0
	16	100	100	100	100	100	100	100	100	100	96	1	–	–
	32	100	100	100	100	100	100	100	100	100	88	2	–	–

Table 30: MNIST (3 vs 8, $n = 784$, depth 8, width 256, 150 factuals), ℓ_∞ objective. Greedy* did not run at this scale; under ℓ_1 the norm-adapted GS* and PGD-bisect* runs cover 70 of the 150 factuals (validity over the arm's own rows). PBMC3k is reported in the main text, after the shared marker completion. DASH = the CPU implementation; NA = no valid counterfactual, or a metric the run did not record.

dataset	method	success %	valid %	ℓ_∞	l0	redund.	yNN	yNN-dist	time (s)
mnist	DASH	**100**	**100**	0.075	477.0	59.5	0.000	5.194	74.2436
mnist	DASH + data seed	**100**	**100**	0.075	480.0	59.5	0.000	5.176	74.8974
mnist	Wachter	**100**	**100**	0.105	766.5	407.5	0.000	NA	2.0280
mnist	DiCE	77	77	0.100	161.5	0.0	0.000	NA	0.0372
mnist	GS*	**100**	**100**	0.838	33.0	23.5	0.000	6.251	0.2176
mnist	PGD-bisect*	**100**	**100**	**0.069**	477.0	27.0	0.000	5.127	3.0292
mnist	FMN*	**100**	**100**	**0.069**	476.0	70.0	0.000	NA	0.8799
mnist	NICE	**100**	**100**	0.992	**22.0**	**6.0**	**0.200**	NA	0.3853
mnist	FAB	**100**	**100**	0.070	479.0	35.0	0.000	**5.117**	**0.1361**
mnist	DeepFool	100	0	NA	NA	NA	NA	NA	NA

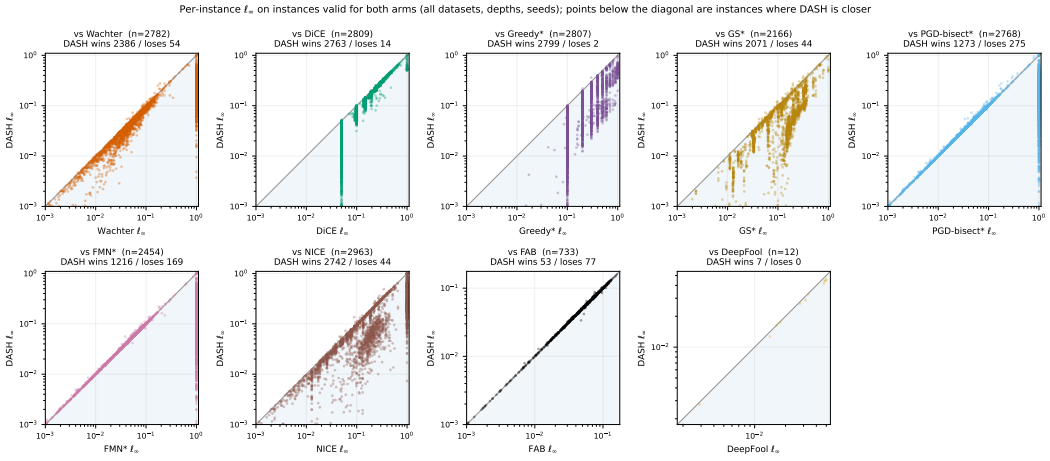

Figure 16: Per-instance paired comparison, ℓ_∞ objective (see Figure 8 for the definition).

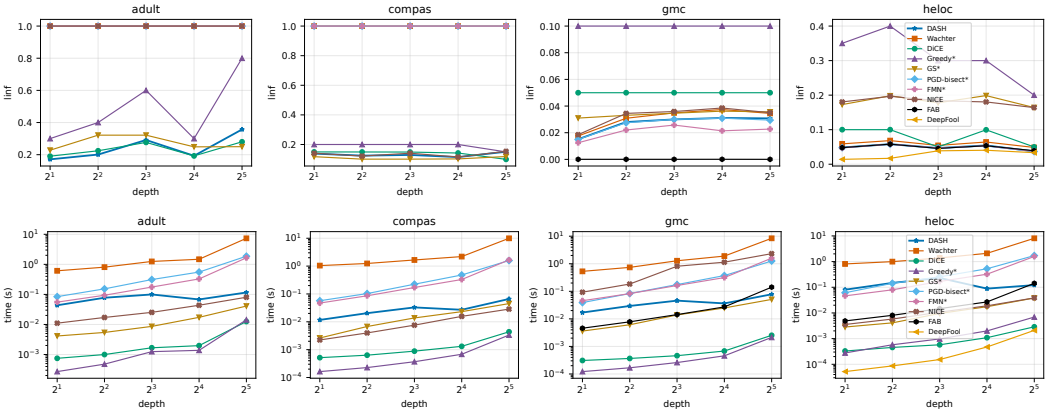

Figure 17: Median ℓ_∞ and runtime versus depth, ℓ_∞ objective.

Table 31: All tabular cells, ℓ_∞ objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 28) and the best-known table (Table 27).

dataset	depth	method	success %	valid %	ℓ_∞	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_2
adult	2	DASH	**100**	**100**	0.171	3.0	**0.0**	0.000	0.407	0.0435	0.553	0.300
adult	2	DASH + data seed	**100**	**100**	0.171	3.0	**0.0**	0.000	0.407	0.0486	0.552	0.300
adult	2	Wachter	96	95	1.000	5.0	3.0	0.000	1.425	0.6108	2.258	1.421
adult	2	DiCE	**100**	99	0.191	3.0	**0.0**	0.000	0.411	0.0007	0.451	**0.294**
adult	2	Greedy*	**100**	**100**	0.300	**1.0**	**0.0**	0.000	0.400	**0.0003**	**0.300**	0.300
adult	2	GS*	100	71	0.228	1.0	0.0	0.000	0.339	0.0041	0.251	0.251
adult	2	PGD-bisect*	92	92	1.000	5.0	3.0	0.000	1.425	0.0842	2.256	1.421
adult	2	FMN*	84	84	1.000	5.0	2.5	0.000	1.425	0.0554	2.260	1.421
adult	2	NICE	**100**	**100**	1.000	2.0	**0.0**	**0.400**	1.418	0.0110	4.000	2.000
adult	2	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	2	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	2	MIP-EXP	**100**	**100**	0.168	3.0	**0.0**	0.000	**0.398**	2.6996	0.568	0.300
adult	2	MIP-OBJ	99	99	**0.167**	3.0	**0.0**	0.000	**0.398**	0.6615	0.564	0.297
adult	4	DASH	**100**	**100**	**0.201**	4.0	**0.0**	0.000	0.646	0.0774	0.802	0.401
adult	4	DASH + data seed	**100**	**100**	**0.201**	4.0	**0.0**	0.000	0.646	0.0864	0.802	0.401
adult	4	Wachter	90	88	1.000	5.0	2.0	0.000	1.432	0.8069	2.376	1.427
adult	4	DiCE	**100**	97	0.224	3.0	**0.0**	0.000	**0.623**	0.0010	0.718	**0.396**
adult	4	Greedy*	**100**	**100**	0.400	**1.0**	**0.0**	0.000	0.648	**0.0005**	**0.400**	0.400
adult	4	GS*	100	61	0.321	1.0	0.0	0.000	0.471	0.0054	0.321	0.321
adult	4	PGD-bisect*	83	83	1.000	5.0	3.0	0.000	1.431	0.1530	2.351	1.425
adult	4	FMN*	76	76	1.000	5.0	3.0	0.000	1.431	0.0938	2.368	1.426
adult	4	NICE	**100**	**100**	1.000	3.0	**0.0**	**0.400**	1.417	0.0171	4.228	2.011
adult	4	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	4	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	4	MIP-EXP	**100**	**100**	**0.202**	4.0	**0.0**	0.000	0.674	60.0003	0.804	0.402
adult	4	MIP-OBJ	**100**	**100**	**0.201**	4.0	**0.0**	0.000	0.700	38.8316	0.804	0.402
adult	8	DASH	**100**	**100**	0.291	4.0	**0.0**	0.000	0.950	0.0997	1.134	0.567
adult	8	DASH + data seed	**100**	**100**	0.291	4.0	**0.0**	0.000	0.950	0.1052	1.134	0.567
adult	8	Wachter	78	77	1.000	5.0	3.0	0.000	1.430	1.2503	2.423	1.430
adult	8	DiCE	**100**	83	**0.276**	4.0	**0.0**	0.000	**0.765**	0.0017	1.064	**0.532**
adult	8	Greedy*	**100**	82	0.600	**1.0**	**0.0**	0.000	0.933	**0.0012**	**0.600**	0.600
adult	8	GS*	100	25	0.321	2.0	0.0	0.000	0.484	0.0086	0.501	0.354
adult	8	PGD-bisect*	73	73	1.000	5.0	4.0	0.000	1.429	0.3121	2.392	1.428
adult	8	FMN*	70	70	1.000	5.0	4.0	0.000	1.429	0.1756	2.389	1.428
adult	8	NICE	**100**	**100**	1.000	3.0	**0.0**	**0.600**	1.417	0.0253	4.118	2.004
adult	8	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	8	DeepFool	99	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	8	MIP-EXP	7	7	0.042	4.0	0.0	0.100	0.117	42.0766	0.166	0.083
adult	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	16	DASH	**100**	**100**	**0.193**	4.0	**0.0**	0.000	0.683	0.0680	0.761	0.382
adult	16	DASH + data seed	**100**	**100**	**0.193**	4.0	**0.0**	0.000	0.683	0.0743	0.761	0.382
adult	16	Wachter	91	89	1.000	5.0	3.0	0.000	1.433	1.4773	2.360	1.426
adult	16	DiCE	**100**	91	**0.192**	3.0	**0.0**	0.000	**0.591**	0.0020	0.596	0.343
adult	16	Greedy*	**100**	91	0.300	**1.0**	**0.0**	0.000	0.607	**0.0014**	**0.300**	**0.300**
adult	16	GS*	100	46	0.249	1.0	0.0	0.000	0.471	0.0173	0.321	0.274
adult	16	PGD-bisect*	85	85	1.000	5.0	3.0	0.000	1.430	0.5541	2.318	1.423
adult	16	FMN*	83	83	1.000	5.0	3.0	0.000	1.430	0.3273	2.324	1.423
adult	16	NICE	**100**	**100**	1.000	3.0	**0.0**	**0.400**	1.420	0.0436	4.077	2.001
adult	16	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	32	DASH	**100**	**100**	0.357	4.0	**0.0**	0.000	1.108	0.1159	1.428	0.714
adult	32	DASH + data seed	**100**	**100**	0.357	4.0	**0.0**	0.000	1.108	0.1242	1.428	0.714
adult	32	Wachter	82	82	1.000	5.0	3.0	0.000	1.431	7.3611	2.471	1.434
adult	32	DiCE	**100**	81	**0.280**	4.0	**0.0**	0.000	**0.866**	**0.0125**	1.112	**0.556**
adult	32	Greedy*	**100**	83	0.800	**1.0**	**0.0**	0.000	1.026	0.0143	**0.800**	0.800
adult	32	GS*	100	23	0.250	2.0	0.0	0.000	0.550	0.0410	0.491	0.347
adult	32	PGD-bisect*	74	74	1.000	5.0	4.0	0.000	1.426	1.8650	2.419	1.430
adult	32	FMN*	69	69	1.000	5.0	4.0	0.000	1.427	1.6500	2.455	1.432
adult	32	NICE	**100**	**100**	1.000	3.0	**0.0**	**0.600**	1.416	0.0805	4.049	2.001
adult	32	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
adult	32	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	DASH	99	**99**	**0.138**	2.0	**0.0**	0.400	**0.026**	0.0117	**0.159**	**0.141**
compas	2	DASH + data seed	99	**99**	**0.138**	2.0	**0.0**	0.400	**0.026**	0.0124	**0.159**	**0.141**
compas	2	Wachter	89	89	1.000	4.0	2.0	0.600	0.041	1.0425	2.092	1.417
compas	2	DiCE	99	97	0.150	**1.0**	**0.0**	**0.800**	0.028	0.0005	0.169	0.150
compas	2	Greedy*	**100**	97	0.200	**1.0**	**0.0**	**0.800**	0.037	**0.0002**	0.200	0.200
compas	2	GS*	100	61	0.118	1.0	0.0	0.800	0.024	0.0026	0.118	0.118
compas	2	PGD-bisect*	87	87	1.000	3.0	2.0	**0.800**	0.110	0.0570	2.106	1.417
compas	2	FMN*	79	79	1.000	3.0	2.0	0.800	0.134	0.0471	2.101	1.416
compas	2	NICE	98	98	0.141	**1.0**	**0.0**	0.600	0.027	0.0022	0.167	0.142

Continued on next page

Table 31: All tabular cells, ℓ_∞ objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 28) and the best-known table (Table 27).

dataset	depth	method	success %	valid %	ℓ_∞	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_2
compas	2	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	2	MIP-EXP	99	**99**	**0.138**	2.0	**0.0**	0.400	**0.026**	0.9411	**0.160**	**0.141**
compas	2	MIP-OBJ	99	**99**	**0.138**	2.0	**0.0**	0.400	**0.026**	0.7204	**0.160**	**0.141**
compas	4	DASH	**98**	**98**	0.125	2.0	**0.0**	0.600	0.026	0.0204	0.155	0.128
compas	4	DASH + data seed	**98**	**98**	0.125	2.0	**0.0**	0.600	0.026	0.0209	0.155	0.128
compas	4	Wachter	87	87	1.000	4.0	2.0	0.800	0.040	1.2427	2.091	1.416
compas	4	DiCE	**98**	96	0.149	**1.0**	**0.0**	0.800	0.027	0.0006	0.149	0.149
compas	4	Greedy*	**98**	97	0.200	**1.0**	**0.0**	**1.000**	0.028	**0.0002**	0.200	0.200
compas	4	GS*	100	51	0.101	1.0	0.0	1.000	0.013	0.0068	0.101	0.101
compas	4	PGD-bisect*	85	85	1.000	3.0	2.0	0.800	0.160	0.1031	2.070	1.416
compas	4	FMN*	74	74	1.000	3.0	2.0	0.800	0.161	0.0859	2.069	1.415
compas	4	NICE	95	95	0.122	**1.0**	**0.0**	0.800	**0.024**	0.0040	0.138	0.124
compas	4	FAB	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	4	MIP-EXP	92	92	**0.108**	2.0	**0.0**	0.600	**0.025**	10.9424	**0.137**	**0.111**
compas	4	MIP-OBJ	96	96	0.119	2.0	**0.0**	0.600	0.026	33.9119	0.156	0.130
compas	8	DASH	**98**	**98**	**0.129**	2.0	**0.0**	0.400	**0.023**	0.0329	0.152	**0.138**
compas	8	DASH + data seed	**98**	**98**	**0.129**	2.0	**0.0**	0.400	**0.023**	0.0344	0.152	**0.138**
compas	8	Wachter	89	89	1.000	4.0	2.0	0.800	0.072	1.6603	2.088	1.416
compas	8	DiCE	**98**	93	0.149	**1.0**	**0.0**	0.800	0.026	0.0009	**0.149**	0.149
compas	8	Greedy*	97	85	0.200	**1.0**	**0.0**	**1.000**	0.029	**0.0004**	0.200	0.200
compas	8	GS*	100	35	0.101	1.0	0.0	1.000	0.014	0.0137	0.101	0.101
compas	8	PGD-bisect*	87	87	1.000	3.0	2.0	0.800	0.161	0.2235	2.072	1.416
compas	8	FMN*	85	85	1.000	3.0	2.0	0.800	0.161	0.1702	2.071	1.416
compas	8	NICE	**98**	**98**	0.141	**1.0**	**0.0**	0.600	0.025	0.0076	0.162	0.154
compas	8	FAB	99	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	8	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	8	MIP-EXP	48	48	0.050	2.0	0.0	0.600	0.009	6.6603	0.056	0.053
compas	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	16	DASH	**98**	**98**	0.115	2.0	**0.0**	0.400	**0.026**	0.0270	**0.137**	**0.124**
compas	16	DASH + data seed	**98**	**98**	**0.112**	2.0	**0.0**	0.400	**0.027**	0.0295	**0.137**	**0.123**
compas	16	Wachter	91	91	1.000	4.0	2.0	**0.800**	0.103	2.1968	2.066	1.415
compas	16	DiCE	93	93	0.142	**1.0**	**0.0**	**0.800**	0.028	**0.0013**	0.143	0.143
compas	16	Greedy*	96	79	0.200	1.0	0.0	1.000	0.028	0.0007	0.200	0.200
compas	16	GS*	100	36	0.103	1.0	0.0	1.000	0.016	0.0231	0.103	0.103
compas	16	PGD-bisect*	90	90	1.000	3.0	2.0	**0.800**	0.163	0.4794	2.074	1.416
compas	16	FMN*	89	89	1.000	3.0	2.0	**0.800**	0.171	0.3299	2.068	1.415
compas	16	NICE	97	97	0.115	**1.0**	**0.0**	0.600	**0.027**	0.0157	0.141	0.126
compas	16	FAB	94	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	32	DASH	**100**	**100**	**0.151**	2.0	**0.0**	0.600	**0.030**	0.0661	**0.166**	**0.154**
compas	32	DASH + data seed	**100**	**100**	**0.152**	2.0	**0.0**	0.600	**0.029**	0.0694	**0.166**	**0.155**
compas	32	Wachter	79	79	1.000	4.0	2.0	0.800	0.089	9.9577	2.050	1.415
compas	32	DiCE	63	62	0.100	1.0	0.0	0.800	0.021	0.0044	0.100	0.100
compas	32	Greedy*	86	77	0.150	1.0	0.0	1.000	0.027	0.0033	0.200	0.171
compas	32	GS*	100	41	0.119	1.0	0.0	1.000	0.022	0.0450	0.119	0.119
compas	32	PGD-bisect*	93	93	1.000	3.0	2.0	**0.800**	0.187	1.5866	2.092	1.417
compas	32	FMN*	87	87	1.000	3.0	2.0	**0.800**	0.176	1.6704	2.087	1.416
compas	32	NICE	99	99	0.154	**1.0**	**0.0**	**0.800**	**0.029**	**0.0285**	0.168	0.167
compas	32	FAB	57	0	NA	NA	NA	NA	NA	NA	NA	NA
compas	32	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	2	DASH	**99**	**99**	0.015	4.0	**0.0**	0.600	**0.034**	0.0170	0.037	0.023
gmc	2	DASH + data seed	**99**	**99**	0.015	4.0	**0.0**	0.600	0.034	0.0238	0.037	0.023
gmc	2	Wachter	**99**	**99**	0.018	4.0	1.0	0.600	0.034	0.5301	**0.021**	**0.018**
gmc	2	DiCE	**99**	**99**	0.050	**1.0**	**0.0**	0.800	0.057	0.0003	0.050	0.050
gmc	2	Greedy*	**99**	**99**	0.100	**1.0**	**0.0**	**1.000**	0.098	**0.0001**	0.100	0.100
gmc	2	GS*	**99**	**99**	0.031	**1.0**	**0.0**	0.600	0.041	0.0036	0.033	0.031
gmc	2	PGD-bisect*	**99**	**99**	**0.015**	4.0	1.0	0.600	**0.034**	0.0389	0.040	0.024
gmc	2	FMN*	73	73	0.012	4.0	0.0	0.600	0.031	0.0456	0.036	0.021
gmc	2	NICE	97	97	0.019	**1.0**	**0.0**	0.600	0.034	0.0930	**0.021**	0.021
gmc	2	FAB	100	1	0.000	7.0	0.0	0.800	0.016	0.0046	0.000	0.000
gmc	2	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	2	MIP-EXP	**99**	**99**	**0.015**	4.0	1.0	0.600	**0.034**	0.4145	0.041	0.025
gmc	2	MIP-OBJ	97	97	**0.014**	4.0	**0.0**	0.600	**0.033**	0.2084	0.039	0.024
gmc	4	DASH	**100**	**100**	0.028	3.0	**0.0**	0.800	**0.043**	0.0294	0.066	0.042
gmc	4	DASH + data seed	**100**	**100**	0.028	3.0	**0.0**	0.800	**0.043**	0.0342	0.066	0.042
gmc	4	Wachter	99	99	0.031	4.0	1.0	0.600	0.046	0.7389	**0.037**	**0.033**
gmc	4	DiCE	99	99	0.050	**1.0**	**0.0**	0.800	0.058	0.0004	0.050	0.050
gmc	4	Greedy*	99	99	0.100	**1.0**	**0.0**	**1.000**	0.098	**0.0002**	0.100	0.100

Continued on next page

Table 31: All tabular cells, ℓ_∞ objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 28) and the best-known table (Table 27).

dataset	depth	method	success %	valid %	ℓ_∞	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_2
gmc	4	GS*	**100**	**100**	0.033	**1.0**	**0.0**	0.800	0.047	0.0061	0.041	0.040
gmc	4	PGD-bisect*	**100**	**100**	0.027	4.0	1.0	0.800	0.044	0.0830	0.076	0.044
gmc	4	FMN*	76	76	0.022	4.0	0.0	0.800	0.041	0.0826	0.057	0.035
gmc	4	NICE	98	98	0.034	**1.0**	**0.0**	0.800	0.045	0.1838	0.044	0.038
gmc	4	FAB	100	3	0.000	9.0	0.0	0.800	0.015	0.0078	0.000	0.000
gmc	4	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	4	MIP-EXP	**100**	**100**	0.028	4.0	1.0	0.800	0.045	1.1700	0.076	0.045
gmc	4	MIP-OBJ	97	97	**0.026**	4.0	**0.0**	0.800	**0.043**	14.3552	0.075	0.044
gmc	8	DASH	**99**	**99**	0.030	3.0	**0.0**	0.800	0.045	0.0459	0.062	0.043
gmc	8	DASH + data seed	**99**	**99**	0.030	3.0	**0.0**	0.800	0.045	0.0520	0.062	0.043
gmc	8	Wachter	97	97	0.035	3.5	1.0	0.600	0.048	1.2837	**0.044**	0.038
gmc	8	DiCE	92	92	0.050	**1.0**	**0.0**	0.800	0.056	0.0005	0.050	0.050
gmc	8	Greedy*	91	91	0.100	**1.0**	**0.0**	**1.000**	0.095	**0.0003**	0.100	0.100
gmc	8	GS*	**99**	**99**	0.035	**1.0**	**0.0**	0.800	0.049	0.0139	0.048	0.041
gmc	8	PGD-bisect*	**99**	**99**	0.030	3.0	**0.0**	0.800	0.045	0.1754	0.074	0.047
gmc	8	FMN*	64	64	0.026	3.0	0.0	0.800	0.045	0.1625	0.069	0.041
gmc	8	NICE	98	98	0.036	**1.0**	**0.0**	0.800	0.047	0.8128	0.056	0.043
gmc	8	FAB	100	1	0.000	8.5	0.0	0.900	0.017	0.0143	0.000	0.000
gmc	8	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	8	MIP-EXP	90	90	**0.023**	3.0	1.0	0.800	**0.041**	3.1355	0.062	**0.036**
gmc	8	MIP-OBJ	2	2	0.243	2.0	0.0	0.000	0.347	1.7662	0.486	0.344
gmc	16	DASH	**99**	**99**	**0.031**	3.0	**0.0**	0.800	**0.048**	0.0364	0.077	0.046
gmc	16	DASH + data seed	**99**	**99**	**0.031**	3.0	**0.0**	0.800	**0.048**	0.0433	0.077	0.046
gmc	16	Wachter	**99**	**99**	0.037	3.0	1.0	0.800	0.049	1.8976	**0.044**	**0.039**
gmc	16	DiCE	**99**	**99**	0.050	**1.0**	**0.0**	0.800	0.059	0.0007	0.050	0.050
gmc	16	Greedy*	**99**	**99**	0.100	**1.0**	**0.0**	**1.000**	0.097	**0.0005**	0.100	0.100
gmc	16	GS*	**99**	**99**	0.036	**1.0**	**0.0**	0.800	0.050	0.0249	0.059	0.048
gmc	16	PGD-bisect*	**99**	**99**	**0.031**	3.0	**0.0**	0.800	**0.048**	0.3725	0.079	0.048
gmc	16	FMN*	63	63	0.021	3.0	0.0	0.800	0.040	0.3148	0.056	0.034
gmc	16	NICE	**99**	**99**	0.039	**1.0**	**0.0**	0.800	0.051	1.1385	0.053	0.046
gmc	16	FAB	100	3	0.000	8.5	0.0	0.900	0.027	0.0271	0.000	0.000
gmc	16	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
gmc	32	DASH	**98**	**98**	0.031	3.0	1.0	0.800	**0.046**	0.0779	0.065	0.042
gmc	32	DASH + data seed	**98**	**98**	**0.030**	3.0	**0.0**	0.800	**0.046**	0.0862	0.063	0.044
gmc	32	Wachter	96	96	0.034	3.0	1.0	0.800	**0.046**	8.3873	**0.037**	**0.035**
gmc	32	DiCE	92	92	0.050	**1.0**	**0.0**	0.800	0.055	0.0025	0.050	0.050
gmc	32	Greedy*	93	93	0.100	**1.0**	**0.0**	**1.000**	0.094	**0.0021**	0.100	0.100
gmc	32	GS*	**98**	**98**	0.036	2.0	**0.0**	0.800	0.049	0.0508	0.050	0.041
gmc	32	PGD-bisect*	**98**	**98**	**0.029**	3.0	1.0	0.800	**0.046**	1.2592	0.069	0.044
gmc	32	FMN*	66	66	0.023	3.0	0.0	0.800	0.037	1.5946	0.057	0.036
gmc	32	NICE	**98**	**98**	0.035	**1.0**	**0.0**	0.800	0.049	2.3263	0.051	0.039
gmc	32	FAB	100	1	0.000	7.5	0.0	0.800	0.019	0.1415	0.000	0.000
gmc	32	DeepFool	100	0	NA	NA	NA	NA	NA	NA	NA	NA
heloc	2	DASH	**100**	**100**	**0.048**	20.0	**0.0**	0.100	0.323	0.0803	0.962	0.213
heloc	2	DASH + data seed	**100**	**100**	**0.048**	20.0	**0.0**	0.000	0.323	0.0897	0.962	0.212
heloc	2	Wachter	**100**	**100**	0.059	22.0	5.0	0.000	0.313	0.8103	0.721	**0.178**
heloc	2	DiCE	**100**	**100**	0.100	8.0	**0.0**	**0.200**	0.325	0.0003	0.697	0.220
heloc	2	Greedy*	**100**	**100**	0.350	**2.0**	**0.0**	**0.200**	0.438	**0.0003**	**0.430**	0.400
heloc	2	GS*	**100**	**100**	0.173	5.5	1.0	**0.200**	0.389	0.0028	0.786	0.348
heloc	2	PGD-bisect*	**100**	**100**	**0.048**	20.5	**0.0**	0.000	0.324	0.0612	0.921	0.206
heloc	2	FMN*	99	99	**0.048**	21.0	**0.0**	**0.200**	0.324	0.0463	0.926	0.210
heloc	2	NICE	**100**	**100**	0.180	5.0	**0.0**	**0.200**	**0.305**	0.0036	0.574	0.276
heloc	2	FAB	**100**	**100**	**0.048**	21.0	**0.0**	**0.200**	0.325	0.0049	0.913	0.205
heloc	2	DeepFool	100	2	0.014	23.0	13.0	0.200	0.422	0.0001	0.325	0.068
heloc	2	MIP-EXP	**100**	**100**	**0.048**	20.0	**0.0**	0.100	0.324	1.2474	0.927	0.209
heloc	2	MIP-OBJ	**100**	**100**	**0.048**	20.0	**0.0**	0.100	0.324	2.8335	0.926	0.209
heloc	4	DASH	**100**	**100**	0.058	21.0	**0.0**	0.000	0.335	0.1478	1.145	0.254
heloc	4	DASH + data seed	**100**	**100**	0.058	21.0	**0.0**	0.000	0.334	0.1522	1.168	0.257
heloc	4	Wachter	**100**	**100**	0.068	22.0	4.0	0.000	0.324	0.9962	0.968	**0.226**
heloc	4	DiCE	**100**	**100**	0.100	9.5	**0.0**	0.200	0.333	**0.0005**	0.897	0.290
heloc	4	Greedy*	**100**	**100**	0.400	**3.0**	**0.0**	0.200	0.422	0.0006	**0.586**	0.419
heloc	4	GS*	**100**	**100**	0.198	7.0	1.0	0.200	0.394	0.0041	0.978	0.391
heloc	4	PGD-bisect*	**100**	**100**	0.058	21.0	**0.0**	0.000	0.334	0.1424	1.130	0.253
heloc	4	FMN*	**100**	**100**	0.058	21.0	**0.0**	0.000	0.334	0.0795	1.149	0.256
heloc	4	NICE	**100**	**100**	0.197	7.0	**0.0**	**0.300**	**0.308**	0.0058	0.709	0.316
heloc	4	FAB	**100**	**100**	0.058	21.0	**0.0**	0.000	0.334	0.0081	1.097	0.249
heloc	4	DeepFool	100	1	0.017	23.0	4.0	0.600	0.249	0.0001	0.392	0.082
heloc	4	MIP-EXP	93	93	**0.055**	21.0	**0.0**	0.200	0.336	13.1694	1.074	0.239
heloc	4	MIP-OBJ	98	98	0.069	21.0	**0.0**	0.000	0.362	60.0045	1.310	0.291
heloc	8	DASH	**100**	**100**	0.047	21.0	**0.0**	0.000	0.324	0.2311	0.965	0.211

Continued on next page

Table 31: All tabular cells, ℓ_∞ objective (medians over valid rows; bold = best per column within a cell among arms with validity $\geq 80\%$). Medians are over each arm's own valid rows; an arm with low validity is summarised on an easier subset, so compare arms with different validity through the paired table (Table 28) and the best-known table (Table 27).

dataset	depth	method	success %	valid %	ℓ_∞	l0	redund.	yNN	yNN-dist	time (s)	ℓ_1	ℓ_2
heloc	8	DASH + data seed	**100**	**100**	0.047	21.0	**0.0**	0.000	0.322	0.2411	0.966	0.213
heloc	8	Wachter	**100**	**100**	0.054	22.0	4.0	0.000	**0.309**	1.3847	0.768	**0.185**
heloc	8	DiCE	99	99	0.050	9.0	**0.0**	**0.200**	0.317	**0.0006**	0.700	0.218
heloc	8	Greedy*	99	99	0.300	**3.0**	**0.0**	**0.200**	0.384	0.0010	**0.500**	0.333
heloc	8	GS*	**100**	**100**	0.178	5.5	1.0	**0.200**	0.395	0.0095	0.799	0.337
heloc	8	PGD-bisect*	**100**	**100**	**0.047**	21.0	**0.0**	0.000	0.321	0.2554	0.910	0.206
heloc	8	FMN*	**100**	**100**	**0.047**	21.0	**0.0**	0.000	0.325	0.1682	0.924	0.212
heloc	8	NICE	**100**	**100**	0.184	5.0	**0.0**	**0.200**	0.311	0.0107	0.615	0.284
heloc	8	FAB	98	96	**0.046**	21.0	**0.0**	0.000	0.322	0.0147	0.900	0.205
heloc	8	DeepFool	99	1	0.039	23.0	7.5	0.100	0.355	0.0002	0.892	0.186
heloc	8	MIP-EXP	59	59	0.031	21.0	0.0	0.200	0.301	11.9084	0.643	0.140
heloc	8	MIP-OBJ	0	0	NA	NA	NA	NA	NA	NA	NA	NA
heloc	16	DASH	**100**	**100**	0.054	21.0	1.0	0.000	0.331	0.0885	1.081	0.241
heloc	16	DASH + data seed	**100**	**100**	0.054	21.0	**0.0**	0.000	0.329	0.0995	1.072	0.239
heloc	16	Wachter	**100**	**100**	0.064	22.5	4.0	0.000	0.321	2.0969	0.892	**0.206**
heloc	16	DiCE	**100**	**100**	0.100	9.0	**0.0**	**0.200**	0.329	**0.0011**	0.780	0.264
heloc	16	Greedy*	**100**	**100**	0.300	**3.0**	**0.0**	**0.200**	0.366	0.0020	**0.526**	0.334
heloc	16	GS*	**100**	**100**	0.199	6.0	1.0	**0.200**	0.403	0.0175	0.877	0.381
heloc	16	PGD-bisect*	**100**	**100**	**0.053**	21.0	**0.0**	0.000	0.327	0.5253	1.028	0.227
heloc	16	FMN*	**100**	**100**	**0.053**	21.0	**0.0**	0.000	0.327	0.3226	1.030	0.233
heloc	16	NICE	**100**	**100**	0.180	6.0	1.0	**0.200**	**0.315**	0.0190	0.673	0.310
heloc	16	FAB	97	96	**0.053**	21.0	**0.0**	0.000	0.327	0.0268	0.989	0.223
heloc	16	DeepFool	100	1	0.040	23.0	14.5	0.200	0.368	0.0005	0.862	0.184
heloc	32	DASH	**100**	**100**	**0.037**	21.0	**0.0**	0.000	0.314	0.1214	0.753	0.164
heloc	32	DASH + data seed	**100**	**100**	**0.037**	21.0	**0.0**	0.000	0.311	0.1347	0.753	0.164
heloc	32	Wachter	**100**	**100**	0.049	22.0	5.0	0.200	0.304	8.0712	0.574	**0.146**
heloc	32	DiCE	**100**	**100**	0.050	8.0	**0.0**	0.200	0.318	**0.0029**	0.500	0.160
heloc	32	Greedy*	**100**	**100**	0.200	**3.0**	**0.0**	0.200	0.354	0.0069	**0.400**	0.279
heloc	32	GS*	**100**	**100**	0.164	5.0	1.0	0.200	0.370	0.0382	0.613	0.295
heloc	32	PGD-bisect*	**100**	**100**	**0.037**	21.0	**0.0**	0.200	0.310	1.7129	0.718	0.158
heloc	32	FMN*	**100**	**100**	**0.037**	21.0	**0.0**	0.200	0.314	1.5847	0.721	0.159
heloc	32	NICE	**100**	**100**	0.164	5.0	0.5	**0.400**	**0.299**	0.0385	0.485	0.233
heloc	32	FAB	**100**	88	0.039	21.0	**0.0**	0.000	0.306	0.1426	0.748	0.167
heloc	32	DeepFool	99	2	0.033	23.0	23.0	0.400	0.325	0.0021	0.769	0.160