Q-LOSS: A TRAINING OBJECTIVE FOR DIRECT CATE OPTIMIZATION ACROSS META-LEARNER FAMILIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional average treatment effect (CATE) estimation increasingly relies on model-agnostic meta-learners that decompose the problem into standard supervised-learning subproblems. These learners, however, are trained on surrogate losses only loosely aligned with treatment-effect quality, and a recent large-scale benchmark finds that 62% of fitted CATE estimators are degenerate—no better than predicting no effect at all. The Q-statistic of Yu & Sun (2025) provides an unbiased estimate of CATE risk that requires no counterfactuals and has been used to evaluate and rank estimators. We develop a differentiable, end-to-end algorithm for training meta-learners directly on this statistic—with learned feature gates that select treatment-effect modifiers—and instantiate it across the S-learner, X-learner, and DML-Learner families. On real and synthetic randomized-trial benchmarks, we then characterize when this helps. The resulting learners beat the baseline on most benchmarks for standard S-Learner baseline (best variant, paired Wilcoxon). In an ablation, the Q-loss objective and the learned gates each independently recover a plain network from 45% degeneracy to 0%. The apparent failure of the raw inverse-propensity target is largely a matter of propensity misspecification, which estimating the propensity and adopting a doubly-robust target resolves; and the resulting gains are *non-circular*—tracking ground-truth PEHE ($\rho = 0.93$) and an independent held-out policy-value metric rather than the training statistic itself. Together, these analyses clarify when direct optimization of the Q-statistic is preferable to standard baselines.

1 INTRODUCTION

Estimating the Conditional Average Treatment Effect $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ from observational data is central to personalized medicine, policy design, and business experimentation—wherever the right action depends on subgroup characteristics (which patients benefit from a treatment, which customers respond to a promotion).

Despite a decade of methodological progress, a troubling empirical reality has emerged. Yu & Sun (2025) evaluated 16 contemporary CATE estimators across 43,200 dataset variants derived from 12 real-world randomized controlled trials, and found that 62% of fitted CATE estimators are degenerate, they perform worse than the trivial predictor τ(x) = 0 that predicts no treatment effect for anyone. This is not a failure of model capacity or data volume. It is a failure of objective alignment.

The root cause is the proxy-loss problem. Every major family optimizes a surrogate for CATE quality: outcome-prediction learners (S/T/X) minimize MSE on the observed Y (rewarding baseline risk, not effect heterogeneity); semi-parametric learners (R-Learner, DML) minimize R-loss, a control-variate-adjusted Q-statistic variant that still depends on nuisance quality; and DR-Learners inherit both misalignments. Models that excel at outcome prediction thus routinely fail at effect estimation, because minimizing MSE on Y does not minimize MSE on $\tau(x)$.

Yu & Sun (2025) proposed the Q-statistic—an unbiased estimator of CATE error computable without counterfactual outcomes—as an **evaluation** metric to diagnose this failure, leaving open whether it can be optimized *during training*. **We answer yes, and show it works across the full taxonomy of CATE meta-learner families.**

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.

Used as a training loss, the Q-statistic of Yu & Sun (2025) defines what we call **Q-Loss**[1]:

$$\hat{Q}(\hat{\tau}) = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\tau}(x_i)^2 - 2\cdot\hat{\tau}(x_i)\cdot\eta(x_i,t_i,y_i)\right], \tag{1}$$

where η is the Horvitz-Thompson transformation satisfying $\mathbb{E}[\eta \mid X = x] = \tau(x)$. We prove that Q-Loss is convex in $\hat{\tau}$, has bounded gradients under propensity clipping, and is consistent: minimizing $\hat{Q}$ converges to the true CATE in expectation (Propositions 1–2). We further argue that a feature that predicts outcomes but does not modify the treatment effect adds only variance, so under the doubly-robust target the paired feature gates drive its weight toward zero while retaining effect modifiers (Proposition 3; verified in Appendix 8.6). Minimizing this loss is least-squares regression on the pseudo-outcome η (completing the square); with the IPW or doubly-robust choice of η it coincides with the propensity-weighted and doubly-robust learners (Curth & van der Schaar, 2021; Kennedy, 2023).

We instantiate Q-Loss across the three major meta-learning strategies—outcome prediction (Q-S, Q-X), semi-parametric regression (Q-R, Q-DML), and post-hoc selection—and find a consistent pattern: **Q-Loss improves CATE estimation in proportion to how misaligned the base learner's loss is with CATE quality**, with the largest gains for outcome-prediction learners and negligible gains for the already-orthogonalized semi-parametric ones.

Contributions:

1. **End-to-end differentiable optimization of the Q-objective** (Sections 3.1, 3.3): We turn the Q-statistic—equivalently pseudo-outcome regression, i.e. the propensity-weighted and doubly-robust learners (Curth & van der Schaar, 2021; Kennedy, 2023)—into a differentiable training loss and optimize it end-to-end, instantiated as two neural estimators (Q-S-Learner, Q-X-Learner) with Gumbel-Softmax feature gates for the balanced and imbalanced regimes, and applied as a selection criterion for the semiparametric DML/R-Learners.
2. **A feature-selection guarantee under the doubly-robust target** (Section 3.2): We prove that under the doubly-robust target the paired feature gates prune purely prognostic (non-modifying) covariates while retaining effect modifiers (Proposition 3), verified on synthetic data with planted feature roles. (The objective is a convex quadratic in the predicted CATE and inherits the consistency of Yu & Sun (2025)—minimizing it minimizes PEHE in expectation.)
3. **A systematic cross-family study** (Section 4): The first comprehensive study of Q-Loss across the S, X, DML, T, and R architectures, characterizing gains as a function of how misaligned the base learner's loss is with CATE quality, and verifying that the Q-trained improvements transfer to ground-truth PEHE (Spearman $\rho = 0.93$) and to an independent downstream policy-value metric (Section 4.4).
4. **Empirical characterization of when Q-Loss helps, and the observational setting** (Sections 4.5, 4.3.1): Empirical conditions under which Q-Loss gives large gains (outcome-prediction learners, genuine heterogeneity, $n \geq 5$k), small gains (semi-parametric learners), or none (extreme imbalance, sparse outcomes). We further show that applied observationally the propensity must be *estimated* rather than plugged in as a constant: a naive constant-propensity IPW target degenerates, while a cross-fitted $\hat{e}(x)$ removes most of it and a doubly-robust target removes the residual.

2 Background

2.1 Problem Setup

We adopt the potential outcomes framework (Rubin, 1974). Let (X, T, Y) be a tuple where $X \in \mathbb{R}^d$ are pretreatment co-variates, $T \in \{0, 1\}$ is binary treatment, and $Y = Y(T)$ is the observed

[1]The "Q" refers to the Q-statistic of Yu & Sun (2025); it is unrelated to Q-learning and the action-value function Q in reinforcement learning.

outcome. Each unit has two potential outcomes $Y(0)$ and $Y(1)$, but only the factual outcome $Y = Y(T)$ is observed. The CATE is defined as:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]. \tag{2}$$

Under the standard ignorability assumptions (consistency, un-confoundedness, overlap), CATE is identifiable from observational data. Evaluation typically uses the Precision in Estimating Heterogeneous Effects (PEHE):

$$P(\hat{\tau}) = \mathbb{E}_X\left[(\tau(X) - \hat{\tau}(X))^2\right], \tag{3}$$

which requires knowing the true CATE – only possible on synthetic or semi-synthetic datasets.

2.2 Meta-Learners for CATE

Meta-learners decompose CATE estimation into standard supervised learning subproblems. The two main strategies are:

Outcome prediction (S-Learner, T-Learner, X-Learner): Fit models for $\mathbb{E}[Y \mid X, T]$ and take differences. These methods optimize MSE on outcome Y—a proxy loss with no direct connection to CATE quality.

Semi-parametric regression (R-Learner, DML): Residual Y and T against their conditional means, then regress the residuals. These methods optimize R-loss, which equals $\hat{Q}/4$ plus a constant independent of $\hat{\tau}$ (Yu & Sun, 2025, Propositions 3.10–3.11) – they are already Q-statistic variants with control variates.

Thus outcome-prediction methods optimize losses far from CATE quality, while semi-parametric methods are already Q-statistic variants—predicting the empirical pattern of Section 4.2: Q-loss helps most for S/X-Learners on real RCTs where trees cannot exploit the heterogeneity, and little where the baseline already captures the signal (large structured or synthetic data).

2.3 The Q-Statistic

Yu & Sun (2025) decompose PEHE as:

$$P(\hat{\tau}) = \mathbb{E}[\tau^2(X)] + \mathbb{E}[\hat{\tau}^2(X)] - 2\,\mathbb{E}[\tau(X)\hat{\tau}(X)]. \tag{4}$$

The first term is a constant independent of the estimator. The remaining terms define:

$$Q(\hat{\tau}) = \mathbb{E}[\hat{\tau}^2(X)] - 2\,\mathbb{E}[\tau(X)\hat{\tau}(X)] = P(\hat{\tau}) - \mathbb{E}[\tau^2(X)]. \tag{5}$$

The sample estimator uses the Horvitz-Thompson (IPW) transformation (Horvitz & Thompson, 1952) $\eta(x,t,y) = \dfrac{y\,(t - e(x))}{e(x)\,(1 - e(x))}$, which satisfies $\mathbb{E}[\eta \mid X = x] = \tau(x)$ when propensity $e(x)$ is known:

$$\hat{Q}(\hat{\tau}) = \frac{1}{N}\sum_{i=1}^{N}\Big[\hat{\tau}(x_i)^2 - 2\,\hat{\tau}(x_i)\,\eta(x_i, t_i, y_i)\Big]. \tag{6}$$

Lemma 1 (Yu & Sun 2025, Lemma 3.1)**.** *When $e(x)$ is known, $\mathbb{E}[\hat{Q}(\hat{\tau})] = Q(\hat{\tau})$, making $\hat{Q}$ an unbiased estimator of CATE MSE up to the dataset-level constant $\mathbb{E}[\tau^2(X)]$.*

Degeneracy: A model is degenerate if $\hat{Q}(\hat{\tau}) \geq 0$, meaning it is worse than the zero predictor $\hat{\tau}_0(x) = 0$. Yu & Sun (2025) find that 62% of contemporary CATE estimators are degenerate on real-world benchmarks.

3 Q-Learner Framework

Figure 1 contrasts the two: both regress the same pseudo-outcome η, but we optimize it end-to-end with learned gates rather than as a frozen second-stage target.

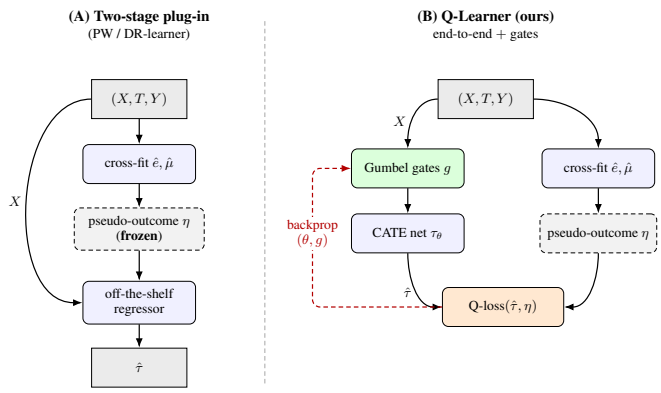

Figure 1: Q-Learner (B) versus the standard two-stage plug-in (A). **(A)** Propensity-weighted / doubly-robust learners (Curth & van der Schaar, 2021; Kennedy, 2023) construct the pseudo-outcome η once, freeze it, and regress it with an off-the-shelf base learner. **(B)** The Q-Learner regresses the *same* η as a differentiable loss: Gumbel-Softmax feature gates and the CATE network are trained *end-to-end* on the Q-loss, so the representation and feature selection co-adapt to the treatment-effect objective.

3.1 Q-Loss as a Training Objective

We propose using $\hat{Q}(\hat{\tau})$ directly as a training loss. The gradient with respect to model parameters θ (via the chain rule through $\hat{\tau}(x;\theta)$) is:

$$\frac{\partial \hat{Q}}{\partial \theta} = \frac{2}{N}\sum_{i=1}^{N}\Big[\hat{\tau}(x_i;\theta) - \eta(x_i, t_i, y_i)\Big] \cdot \frac{\partial \hat{\tau}(x_i;\theta)}{\partial \theta}. \tag{7}$$

Proposition 1 (Well-posedness)**.** *Q-loss is quadratic and convex in $\hat{\tau}$. For any neural network $\hat{\tau}(x;\theta)$ with Lipschitz-continuous output, the gradient $\partial\hat{Q}/\partial\theta$ is bounded when propensity scores are clipped to $e \in [\varepsilon, 1-\varepsilon]$ for $\varepsilon > 0$. The minimum of $\hat{Q}(\hat{\tau})$ over all measurable functions $\hat{\tau}$ is attained at $\hat{\tau}^*(x) = \tau(x)$.*

Proofs of Propositions 1 and 2 are given in Appendix 7.2.

Proposition 2 (Consistency)**.** *Minimizing $\hat{Q}(\hat{\tau})$ over a sufficiently rich function class converges to the true CATE $\tau(x)$ as $N \to \infty$, under the conditions of Lemma 1 (known propensity) and standard regularity conditions on the data-generating process.*

Remark 1 (Relationship to R-loss and DR-loss)**.** *R-loss equals $\hat{Q}(r_R)/4$ plus a constant and DR-loss equals $\hat{Q}(r_{DR})$ plus a constant (Yu & Sun, 2025, Propositions 3.10–3.11), where r_R and r_{DR} are variance-reducing control variates. Our Q-loss is the "raw" $\hat{Q}$ without control variates—higher variance but no dependence on nuisance model estimates. This nuisance-freedom is what distinguishes it from simply switching to DML (which already approximates Q), while retaining theoretical consistency.*

3.1.1 Target Variants: IPW and AIPW Pseudo-Outcomes

The Q-loss regresses $\hat{\tau}$ onto a pseudo-outcome η with the defining property $\mathbb{E}[\eta \mid X] = \tau(X)$; any such η yields the same population minimizer $\hat{\tau}^*(x) = \tau(x)$ (Proposition 1), so the *choice of η is a variance-reduction decision, not a bias one*. We consider two targets:

IPW target (default). The raw Horvitz-Thompson transform $\eta_{\text{IPW}} = \dfrac{Y(T-e)}{e(1-e)}$ requires no nuisance models—it is the nuisance-free objective inherited directly from Yu & Sun (2025). Its price is variance: every outcome is divided by $e(1-e)$ and sign-flipped by $(T-e)$, so the gradient signal-to-noise ratio, $\sigma_{\text{CATE}}/\sigma_\eta$, can be as low as 5–8% on datasets where the true effect is small relative to outcome noise.

AIPW target (extension). Substituting the augmented (doubly-robust) pseudo-outcome

$$\eta_{\mathrm{DR}} = \big(\hat{\mu}_1(x) - \hat{\mu}_0(x)\big) + \frac{T - e}{e(1-e)}\big(Y - \hat{\mu}_T(x)\big), \tag{8}$$

leaves the population minimizer unchanged while reducing gradient variance: the outcome models $\hat{\mu}_0, \hat{\mu}_1$ explain the prognostic component of Y, so only the residual $(Y - \hat{\mu}_T)$ passes through the IPW weight. This is the DR-loss control variate of Remark 1—i.e. **Q-loss with the AIPW target equals the DR-loss up to a constant**—used here as a *training target* rather than an evaluation metric. In an RCT the propensity e is known, so η_{DR} is unbiased regardless of $\hat{\mu}$ quality (the outcome models affect variance, not bias), while evaluation still uses the nuisance-free IPW Q-statistic; we cross-fit $\hat{\mu}_0, \hat{\mu}_1$ (2-fold) to avoid overfitting.

The AIPW gain is finite-sample: it lowers variance only when the nuisances are estimable, and on very small samples the outcome-model error can dominate. We therefore treat IPW as the default Q-Learner and AIPW as an extension, reporting the controlled head-to-head (Section 4.3).

3.2 Differentiable Feature Selection via Gumbel-Softmax Gates

Unlike standard feature selection, which ranks features by correlation with Y and cannot separate effect modifiers from purely prognostic covariates, the Q-loss rewards features that improve treatment-effect prediction—enabling causally-grounded selection.

We implement this via learnable Gumbel-Softmax gates (Jang et al., 2017). For each feature j, we maintain a logit $g_j \in \mathbb{R}$. The gate mask is:

$$m = \sigma\left(\frac{g + \varepsilon}{\beta}\right), \qquad \varepsilon \sim \mathrm{Gumbel}(0, 1), \tag{9}$$

where β is a temperature parameter annealed from $\beta_{\mathrm{init}} = 1.0$ to $\beta_{\mathrm{final}} = 0.1$ during training. The masked features $\tilde{x} = m \odot x$ are passed to the CATE network. The full objective is:

$$\mathcal{L}(\theta, g) = \hat{Q}\big(\hat{\tau}(\tilde{x}; \theta)\big) + \lambda \cdot \frac{1}{d}\sum_{j=1}^{d} \sigma(g_j), \tag{10}$$

where λ controls sparsity. As $\beta \to 0$, gates converge to hard binary selection; at $\beta = 1$ they provide smooth gradients.

Proposition 3 (Feature selection under the AIPW target)**.** *Let Z be a covariate that does not modify the treatment effect ($\tau(x)$ does not depend on Z) and is non-confounding ($Z \perp T \mid X_{-Z}$). Then Z is irrelevant to the population minimizer of the Q-loss, so including it cannot reduce the population objective and can only increase the finite-sample variance of $\hat{\tau}$. Under the variance-reduced AIPW target, the paired feature gate drives Z's weight toward zero while retaining genuine effect modifiers.*

The argument—including a non-modifier is a variance-only decision, realized only under the low-variance AIPW target—is given in Appendix 7.2; we verify it on synthetic data with planted feature roles (Appendix 8.6), where only the AIPW target drives the prognostic feature's gate to zero, even at $\lambda = 0$.

Temperature annealing ensures convergence: at $\beta \to 0$, the Gumbel relaxation is exact (Jang et al., 2017), so the continuous-relaxation solution converges to the discrete feature selection solution.

3.3 Architecture Variants

The Q-Learner framework is instantiated in three architectures:

Q-S-Learner: A single neural network $\hat{\tau}(\tilde{x}; \theta)$ trained end-to-end with Q-loss. Simple architecture; appropriate when the treatment-control split is balanced and the S-Learner's single-model assumption is reasonable.

Q-X-Learner: Follows the X-Learner (Künzel et al., 2019) pipeline (outcome models $\hat{\mu}_0, \hat{\mu}_1$, then CATE networks $\hat{\tau}_0, \hat{\tau}_1$ fine-tuned with Q-loss on the imputed effects), combined by the propensity-weighted $\hat{\tau}(x) = e(x)\hat{\tau}_0(x) + (1 - e(x))\hat{\tau}_1(x)$, which handles imbalance by trusting the larger group's model.

Q-DML-Learner: Follows the DML (Chernozhukov et al., 2018) pipeline—cross-fitted Robinson residuals—then fits the CATE network on the orthogonalized target and selects by held-out $\hat{Q}$. This target is *already* the variance-reduced form of $\hat{Q}$ (Remark 1), so the IPW/AIPW distinction collapses; Q-DML is the one variant that stays non-degenerate on synthetic and large-scale data (Section 7.5.4).

Framework boundaries: Q-loss attaches end-to-end only to architectures with a single unified $\hat{\tau}(x)$ to backpropagate through. It does not train end-to-end on per-arm or tree models (T-Learner, Causal Forest) or on multi-objective architectures (DR-Learner, DragonNet); for residualization-first learners (R-Learner, Q-T-Learner) it applies post-hoc as a selection criterion rather than end-to-end.

The full training loop—nuisance cross-fitting, pseudo-outcome construction, and joint gate/network updates on the Q-loss—is given in Algorithm 1 (Appendix 8.7).

4 Experiments

4.1 Experimental Setup

Evaluation protocol: We follow the observational sampling protocol of Yu & Sun (2025). Starting from RCT data with known propensity, we introduce selection bias via a biasing function $G(t, x)$ to create a confounded training set D_{est}, while retaining an unbiased RCT holdout D_{eval}. Models are trained on D_{est} and evaluated on D_{eval} using $\hat{Q}(\hat{\tau})$ with the known RCT propensity. This protocol allows evaluation without counterfactual outcomes while controlling for confounding via the known propensity.

Baselines, models, and metrics: We compare against six baseline configurations across five model families from the Yu & Sun (2025) benchmark—S/T/X-Learner (XGB), R-Learner (Ridge), DML (XGB, Linear)—and our Q-S/Q-X-Learners (sparse/medium/dense sparsity). We report mean $\hat{Q}\pm$std, degeneracy rate ($\hat{Q} \geq 0$), and head-to-head win rate over 20 iterations.

Implementation: MLPs with hidden dims $[64, 32]$, Adam (lr $= 10^{-3}$); Q-S 30 epochs (temperature-annealed), Q-X 20 pretrain + 15 finetune; propensity clipped to $[0.05, 0.95]$; PyTorch (Paszke et al., 2019).

4.2 Main Results

Table 1 compares the two AIPW-target Q-Learners—Q-S and Q-X—against the `s.xgb` baseline across six datasets, each run for 20 iterations of the observational sampling protocol. Cells report mean $\hat{Q}\pm$std (lower is better) and the degeneracy rate (the fraction of iterations with $\hat{Q} \geq 0$). Q-DML and the raw-IPW variants are analyzed separately in Sections 7.5.2–7.5.4.

Reading this table We compare the two Q-loss learners against `s.xgb`, the canonical S-Learner baseline of Yu & Sun (2025), on two axes. *Reliability:* on every dataset except Sandercock both Q-loss learners are non-degenerate (degeneracy 0–25%, mean $\hat{Q} < 0$), on par with the baseline; Sandercock is a near-null-heterogeneity RCT on which all three learners here (Q-S, Q-X, `s.xgb`) degenerate (90–95%). *Competitiveness:* at matched (zero) degeneracy the Q-loss learners achieve a lower mean $\hat{Q}$ than the baseline on the large balanced settings (Hillstrom -0.0037 vs. -0.0021; Megafon -0.0295 vs. -0.0200), where the tree captures less of the available heterogeneity. Because the architecture-to-regime heuristic fixes *a priori* which learner applies (Q-S balanced, Q-X imbalanced; Section 3.3), this is evidence the objective transfers across architectures, not a post-hoc pick of the better of two.

Scope. We do *not* claim the lowest $\hat{Q}$ among all estimators; stronger, separately tuned baselines (regularized DML/R-Learner variants) exist. The fair test of the *objective* is the paired, same-backbone comparison—the identical learner trained with vs. without Q-loss—reported per family in Appendix 7.6 and, for the neural backbone, in the ablation (Section 4.3). Per-dataset commentary is in Appendix 7.5.

Holding the architecture fixed, the per-family results and the raw-IPW-vs-AIPW target results for Q-S, Q-X, and Q-DML are reported in Appendix 7.5.

Table 1: Q-loss meta-learners (Q-S, Q-X; AIPW target) against the standard S-Learner XGBoost baseline `s.xgb` across six datasets (20 iterations of the observational-sampling protocol, bias_strength 3.0). We report mean $\hat{Q}\pm$std (lower is better) and the degeneracy rate ($\hat{Q} \geq 0$, i.e. worse than the zero-effect predictor). This table asks whether training on the Q-loss objective yields *reliable, competitive* CATE estimates relative to the standard practitioner baseline; it is not a leaderboard over all estimators. The isolated, same-backbone value of the objective (vanilla vs. Q-loss) is reported in Appendix 7.6 (Tables 10, 11) and the ablation (Section 4.3). **Bold** marks the winning arm in each row—the Q-learner(s) that beat `s.xgb`, or `s.xgb` where it wins; * marks $p < 10^{-5}$ (two-sided paired Wilcoxon, exact null, $n = 20$), the strongest at the exact floor 1.9×10^{-6}. Every Q-learner–vs–`s.xgb` difference is significant at $p < 0.05$; the direction of each is given by the sign of the mean $\hat{Q}$. Because both arms see the same 20 sampling draws, the reported $\pm$std is common-mode across-iteration dispersion that largely cancels in the paired test, so interval overlap is not the comparison of record.

				`q_s.aipw` (ours)		`qx.aipw` (ours)		`s.xgb` (baseline)	
Dataset	Type	n	Treat%	mean $\hat{Q}\pm$std	Degen	mean $\hat{Q}\pm$std	Degen	mean $\hat{Q}\pm$std	Degen
Hillstrom	Real RCT	64k	67%	$\mathbf{-0.0037 \pm 0.0005}^{*}$	0%	$\mathbf{-0.0035 \pm 0.0006}^{*}$	0%	-0.0021 ± 0.0007	0%
Megafon	Synth RCT	600k	50%	$\mathbf{-0.0295 \pm 0.0034}^{*}$	0%	$\mathbf{-0.0213 \pm 0.0011}$	0%	-0.0200 ± 0.0008	0%
SynthUplift	Synthetic	50k	70%	$-0.0040 \pm 0.0023^{*}$	5%	-0.0078 ± 0.0014	0%	$\mathbf{-0.0084 \pm 0.0016}$	0%
Criteo	Real RCT	14M	85%	$\mathbf{-0.0008 \pm 0.0001}^{*}$	25%	$\mathbf{-0.0009 \pm 0.0002}$	0%	-0.0007 ± 0.0002	0%
RetailHero	Real RCT	200k	50%	$\mathbf{-0.0009 \pm 0.0002}^{*}$	0%	$\mathbf{-0.0009 \pm 0.0005}$	10%	-0.0005 ± 0.0003	0%
Sandercock	Real RCT	19k	50%	$+0.0071 \pm 0.0019^{*}$	95%	$+0.0013 \pm 0.0008$	95%	$\mathbf{+0.0007 \pm 0.0005}$	90%

4.3 Ablation Studies

We decompose Q-Learner improvement into its components via a 2×2 factorial design (Q-loss $\times$ feature gates). We report Mean $\hat{Q}\pm$ std over $N = 20$ seeds; lower $\hat{Q}$ is better, lower degeneracy is better, higher win-rate is better.

What the ablation establishes. The finding is that *each component is independently contributing to the results*: a plain neural net under MSE is 45% degenerate and sits at Mean $\hat{Q} \approx 0$ (indistinguishable from the zero predictor), and adding **either** Q-loss **or** gates alone drops degeneracy to 0% while moving Mean $\hat{Q}$ to ≈ -0.020—a shift of roughly 5 std, and past the XGBoost reference (-0.0132). Neither the neural architecture alone nor either component alone accounts for the improvement; each mechanism independently recovers the model from degeneracy.

On the imbalanced SynthUplift dataset the two mechanisms are complementary rather than redundant: on the two-arm X-Learner backbone, gates alone reach mean $\hat{Q} = -0.2518$ and Q-loss alone -0.1480, but only the full model combines them to -0.2783 at 0% degeneracy and a 60% head-to-head win rate—the clearest case where both components are jointly required. The full 2×2 ablation tables for both backbones (Q-S on Hillstrom, Q-X on SynthUplift) are in Appendix 7.3.

Hyperparameter sensitivity analyses and the full paired IPW-vs-AIPW target comparison (which the propensity analysis below subsumes) are in Appendix 7.4.

4.3.1 Is the degeneracy variance, or propensity misspecification?

The comparison above holds the propensity *known* (the constant RCT value). But the observational-sampling protocol injects confounding, so on the confounded training split the true assignment probability is x-dependent; plugging in the constant RCT value biases the IPW target toward the prognostic surface $\mu_1 + \mu_0$ rather than τ. We therefore ask whether the raw-IPW degeneracy is intrinsic gradient variance or an artifact of this misspecified propensity. We re-train Q-S with a *cross-fitted, data-driven* propensity $\hat{e}(x)$ estimated on the training split (gradient-boosted, 3-fold out-of-fold), evaluating—as everywhere—with the known RCT propensity. Table 2 reports the result.

Estimating $\hat{e}(x)$ alone removes most of the degeneracy on five of six datasets (all but near-null Sandercock; Table 2). So the raw-IPW collapse largely reflects *propensity misspecification*—the sampling protocol makes treatment assignment x-dependent on the training split, so the constant RCT propensity (correct for the RCT evaluation set) is misspecified there—rather than intrinsic gradient variance. The doubly-robust (AIPW) target then removes the residual (e.g. RetailHero $70\% \to 5\%$, Megafon $\to 0\%$): the two mechanisms compound—correct the propensity first, and

Table 2: Q-S under three training targets: raw IPW with the known constant propensity (as in Table 7); IPW with a cross-fitted estimated propensity $\hat{e}(x)$; and AIPW with $\hat{e}(x)$. Estimating $\hat{e}(x)$ alone removes most of the degeneracy (all but near-null Sandercock); AIPW removes the residual. Evaluation always uses the known RCT propensity. Mean $\hat{Q}\pm$std (lower better) and degeneracy rate; **bold** marks the lowest mean $\hat{Q}$ per row.

	IPW, known e		IPW, est. $\hat{e}(x)$		AIPW, est. $\hat{e}(x)$	
Dataset	mean $\hat{Q}\pm$std	degen	mean $\hat{Q}\pm$std	degen	mean $\hat{Q}\pm$std	degen
Hillstrom	$+0.0267 \pm 0.0116$	95%	$\mathbf{-0.0034 \pm 0.0008}$	0%	-0.0030 ± 0.0012	5%
SynthUplift	$+0.0290 \pm 0.0084$	100%	$+0.0001 \pm 0.0028$	40%	$\mathbf{-0.0003 \pm 0.0025}$	45%
Criteo	$+0.0785 \pm 0.0414$	100%	$+0.0002 \pm 0.0012$	35%	$\mathbf{-0.0001 \pm 0.0000}$	10%
RetailHero	$+0.0801 \pm 0.0173$	100%	$+0.0019 \pm 0.0035$	70%	$\mathbf{-0.0009 \pm 0.0004}$	5%
Megafon	$+0.0647 \pm 0.0102$	100%	-0.0069 ± 0.0082	15%	$\mathbf{-0.0383 \pm 0.0024}$	0%
Sandercock	$+0.0324 \pm 0.0087$	100%	$+0.0199 \pm 0.0102$	100%	$\mathbf{+0.0191 \pm 0.0057}$	100%

AIPW mops up the remaining variance. This delimits the variance framing of Section 3.1.1 and shows the method extends to the observational setting where the propensity must be estimated rather than known.

4.4 Validation: Q-Statistic Is Not Circular

A central concern with training and evaluating on the same metric is self-serving bias: a model optimized for $\hat{Q}$ might score well on $\hat{Q}$ without genuinely improving CATE estimation. Two analyses rule this out. **(a) Agreement with ground-truth PEHE.** On SynthUplift, where true τ is known, the $\hat{Q}$ ranking of all models in Table 13 agrees closely with the PEHE ranking (Spearman $\rho = 0.93$, Pearson $r = 0.99$; Table 13)—so optimizing $\hat{Q}$, which needs no counterfactuals, selects models genuinely closer to the true CATE. **(b) Transfer to an independent policy-value metric.** On held-out RCT data we compute policy value $V(\pi) = \mathbb{E}[Y(\pi(X))]$ under the policy $\pi(x) = \mathbb{1}\{\hat{\tau}(x) > 0\}$—a downstream decision-quality metric structurally independent of $\hat{Q}$ (Appendix Table 21).

Within-model ρ is the Spearman correlation between Q-stat and policy-value improvements across iterations. On Hillstrom, Q-stat improvements track policy-value improvements ($\rho = 0.726$, $p < 0.001$); agreement is weaker on near-null Sandercock ($\rho = 0.235$), as expected where no method recovers signal.

Theoretical grounding. Propositions 1–2 establish that minimizing Q-loss minimizes PEHE in expectation—equivalent to training and evaluating with MSE, standard practice in supervised learning; the novelty is that $\hat{Q}$ is computable *without counterfactuals*, enabling direct CATE optimization from observational data.

Scope of the degeneracy test: the criterion ($\hat{Q} \geq 0$) is MSE-based and hence scale-sensitive, so it does not by itself certify the scale-invariant who-to-treat ranking; we therefore report ranking metrics (Qini, policy value) alongside it (Appendix 7.1).

4.5 When Does Q-Learner Help?

Q-Loss provides the largest benefit when four conditions hold. **(i) Genuine heterogeneity exists**: on near-null Sandercock both Q-S and the tree baseline stay degenerate (95%/90%) rather than fabricating effects—the conservative behaviour we want on signal-free data. **(ii) The tree baseline has not already captured the signal**: the operative factor is exploitable signal, not the treatment rate—the AIPW target makes Q-S competitive on heavily imbalanced Criteo (85% treated), whereas dense-signal RetailHero (50%) keeps trees strong; the architecture-to-imbalance match (Q-S balanced, Q-X imbalanced) is a useful default heuristic, not a rule. **(iii) Sample size is sufficient**: Q-X reaches 0% degeneracy at $n \geq 5$k. **(iv) The outcome is not extremely sparse**: at Criteo's 0.43% conversion the IPW signal from $Y = 1$ observations is diluted and recovery is only partial. On target choice, for the raw-IPW learners (Q-S, Q-X) prefer AIPW—it cures degeneracy at no significant cost (Appendix 7.4) and widens the viable imbalance range; for the already-orthogonalized Q-DML the choice is immaterial.

5 Related Work

CATE meta-learners. S/T/X-Learners (Künzel et al., 2019), the R-Learner (Nie & Wager, 2020), the DR-Learner (Kennedy, 2023), and double machine learning (Chernozhukov et al., 2018) define the baseline family we build on. Bayesian Causal Forests (Hahn et al., 2020) introduce regularization and shrinkage priors tailored to the treatment-effect component; we instead target heterogeneity-driving features directly through the Q-statistic. Most closely, Curth & van der Schaar (2021) frame these meta-learners as two-step *pseudo-outcome regression*: a second stage regresses a pseudo-outcome—the Horvitz–Thompson (IPW) transform (Horvitz & Thompson, 1952) for the PW-learner, and its doubly-robust augmentation for the DR-learner (Kennedy, 2023)—on the covariates with an arbitrary base learner. Our Q-loss regresses onto the *same* transforms, but as a differentiable end-to-end objective with learned feature gates rather than a plug-in second stage; the objective is thus not new—our contribution is the end-to-end training method (with learned feature selection) and its cross-family study. The variance benefit of doubly-robust scores is itself well established—in excess-risk terms (Foster & Syrgkanis, 2023) and as variance reduction for stochastic gradients (Lee et al., 2018)—which we exploit as a *training target* for CATE.

CATE evaluation. Our selection criterion builds on the Q-statistic evaluation from the large-scale CATE benchmark of Yu & Sun (2025). Related model-ranking metrics include Qini curves and AUUC for uplift modeling (Radcliffe, 2007; Diemert et al., 2021) and PEHE for semi-synthetic evaluation (Hill, 2011).

Feature selection for causal inference. Prior work selects covariates to deconfound and to minimize the asymptotic variance of *average* treatment-effect estimates—e.g., adjustment-set selection that separates confounders and precision variables from instruments via a reinforcement-learned feature mask (Wang et al., 2023)—or applies shrinkage to the outcome model (Hahn et al., 2020). Our approach differs on both objective and target: we use the Q-statistic to select features that drive treatment-effect *heterogeneity* for CATE estimation, rather than features that are outcome-predictive or variance-optimal for the ATE. Closer to our aim, doubly-robust adaptive-LASSO selection (Bahamyirou et al., 2022) and treatment-effect variable-importance measures (Hines et al., 2025) identify effect modifiers using doubly-robust machinery, but as post-hoc selection or inference on a fitted CATE; we instead learn the selection *in training* through differentiable gates coupled to the Q-loss, so the estimator itself uses only the retained covariates.

Differentiable feature selection. Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) and L_0 regularization via continuous relaxations (Louizos et al., 2017) enable end-to-end learnable feature selection, and are well established outside causal inference (Balın et al., 2019; Singh et al., 2023; Strypsteen & Bertrand, 2021; 2024). Within causal inference, Gumbel-Softmax has been used for *structure* learning—sampling discrete graph edges in differentiable causal discovery (Brouillard et al., 2020; Lopez et al., 2022; Gong et al., 2022)—rather than for selecting input features to a treatment-effect estimator. To our knowledge, ours is the first application of Gumbel-Softmax-based feature selection to identifying heterogeneity-driving covariates within a CATE meta-learner.

Concurrent work. Ma et al. (2026) propose prior-data fitted networks that perform amortized, in-context CATE estimation from synthetic priors; we instead learn directly from each target dataset via the Q-loss.

6 Conclusion

We introduced Q-Learner, meta-learners that optimize the Q-statistic (Yu & Sun, 2025) as a training loss across the S-, X-, and DML-Learner families. The objective reduces to pseudo-outcome regression, coinciding with known propensity-weighted and doubly-robust learners; our contribution is a training method—estimating the propensity, adopting a doubly-robust target to remove the raw-IPW degeneracy, and Gumbel-Softmax gates that select effect modifiers—that makes it non-degenerate in practice. Under observational sampling the resulting estimates are competitive with the standard S-Learner baseline on most datasets, while tree baselines remain preferable on near-null-heterogeneity settings (Section 4.2).

USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools to polish grammatical errors and writing, used the tools to find related work in different areas of the paper such as a feature selection, meta-learner families and to reformat tabulated results to fit into the ICLR conference format. We additionally used generative AI tools to draft and edit portions of the manuscript for concision and readability, to format references and tables, to suggest document structure We did not use generative AI tools to generate synthetic datasets, for research ideation or execution, for proving mathematical claims. We have reviewed all AI-assisted work—verifying every reported number against its source, manually checking all citations and their characterizations against the cited papers, and confirming the correctness of AI-written analysis code. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Asma Bahamyirou, Mireille E. Schnitzer, Edward H. Kennedy, Lucie Blais, and Yi Yang. Doubly robust adaptive lasso for effect modifier discovery. *The International Journal of Biostatistics*, 18 (2):307–327, 2022. doi: doi:10.1515/ijb-2020-0073. URL https://doi.org/10.1515/ijb-2020-0073.

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pp. 444–453. PMLR, 2019.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 2021.

Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Théophane Gregoir, and Thibaud Rahier. A large scale benchmark for individual treatment effect prediction and uplift modeling. *arXiv preprint arXiv:2111.10106*, 2021.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning, 2023. URL https://arxiv.org/abs/1901.09036.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965 – 2020, 2020. doi: 10.1214/19-BA1195. URL https://doi.org/10.1214/19-BA1195.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Oliver J Hines, Karla Diaz-Ordaz, and Stijn Vansteelandt. Variable importance measures for heterogeneous treatment effects. *Biometrics*, 81(4), October 2025. ISSN 1541-0420. doi: 10.1093/biomtc/ujaf140. URL http://dx.doi.org/10.1093/biomtc/ujaf140.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2280784.

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

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008 – 3049, 2023. doi: 10.1214/23-EJS2157. URL https://doi.org/10.1214/23-EJS2157.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1804597116.

Kanghoon Lee, Jihye Choi, Moonsu Cha, Jung-Kwon Lee, and Taeyoon Kim. Stochastic doubly robust gradient. *arXiv preprint arXiv:1812.08997*, 2018.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35: 19290–19303, 2022.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks, 2026. URL https://arxiv.org/abs/2506.10914.

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

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects, 2020. URL https://arxiv.org/abs/1712.04912.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pp. 14–21, 2007.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. 66, 1974. doi: 10.1037/h0037350.

Dinesh Singh, Héctor Climente-González, Mathis Petrovich, Eiryo Kawakami, and Makoto Yamada. Fsnet: Feature selection network on high-dimensional biological data. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2023.

Thomas Strypsteen and Alexander Bertrand. End-to-end learnable EEG channel selection for deep neural networks with Gumbel-softmax. *Journal of Neural Engineering*, 18(4):0460a9, 2021. doi: 10.1088/1741-2552/ac115d.

Thomas Strypsteen and Alexander Bertrand. Conditional Gumbel-softmax for constrained feature selection with application to node selection in wireless sensor networks, 2024. URL https://arxiv.org/abs/2406.01162.

Haotian Wang, Kun Kuang, Haoang Chi, Longqi Yang, Mingyang Geng, Wanrong Huang, and Wenjing Yang. Treatment effect estimation with adjustment feature selection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 2290–2301, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599531. URL https://doi.org/10.1145/3580305.3599531.

Haining Yu and Yizhou Sun. Do contemporary causal inference models capture real-world heterogeneity? findings from a large-scale benchmark, 2025. URL https://arxiv.org/abs/2410.07021.

7 Appendix

7.1 Scope of the Degeneracy Test

The degeneracy criterion ($\hat{Q} \geq 0$) is, by construction, a statement about the MSE of $\hat{\tau}$, since $\hat{Q}$ equals MSE minus a constant. It is therefore sensitive to the scale and calibration of $\hat{\tau}$ and does not by itself certify the ranking of who-to-treat, which is scale-invariant: an estimator that orders units correctly but is mis-scaled can be flagged degenerate, while one shrunk toward zero can pass. This is why we report ranking-based metrics (Qini, policy value) alongside degeneracy, and read our results as evidence about accuracy under a fixed protocol rather than a universal ranking of estimators.

7.2 Proofs

Proposition 1 (Well-posedness).

Proof sketch. $\hat{Q}$ is a sum of terms $\hat{\tau}(x_i)^2 - 2\hat{\tau}(x_i)\eta_i$, each a convex quadratic in $\hat{\tau}(x_i)$. The gradient $\partial\hat{Q}/\partial\hat{\tau}(x) = 2(\hat{\tau}(x) - \eta)$ is bounded because $|\eta| \leq |Y|/\big(\varepsilon(1-\varepsilon)\big)$ under clipping. Taking expectations,

$$\frac{\partial \mathbb{E}[\hat{Q}]}{\partial \hat{\tau}(x)} = 2\big(\hat{\tau}(x) - \mathbb{E}[\eta \mid X = x]\big) = 2\big(\hat{\tau}(x) - \tau(x)\big),$$

which equals zero at $\hat{\tau} = \tau$. □

Proposition 2 (Consistency).

Proof. Follows directly from Lemma 1 (Yu & Sun, 2025), which establishes $\mathbb{E}[\hat{Q}(\hat{\tau})] = Q(\hat{\tau}) = P(\hat{\tau}) - C$, where C is a constant. Therefore minimizing $\hat{Q}$ in expectation minimizes PEHE. Consistency of the sample minimizer follows from standard M-estimation theory under boundedness conditions. □

Proposition 3 (Feature selection under the AIPW target).

Argument. Since $\tau(x)$ does not depend on Z, the population minimizer $\hat{\tau}^\star(x) = \tau(x)$ (Proposition 1) is constant in Z, so adding Z to the input leaves the population Q-value unchanged—this is a variance-only decision (cf. Section 3.1.1), *not* a population-gradient one. Fitting a function that depends on Z therefore adds estimation variance without reducing bias, and the sparsity-regularized gate is penalized for opening it. Whether this force is realized depends on the gradient signal-to-noise ratio: under the raw IPW target the pseudo-outcome variance is large enough to mask it, whereas under the doubly-robust AIPW target the gate reliably keeps effect modifiers and prunes prognostic non-modifiers. We verify this on synthetic data with planted feature roles (Appendix 8.6): only the AIPW target drives the prognostic feature's gate to zero—even at $\lambda = 0$, so it is not an artifact of the L_1 penalty—while MSE and IPW do not. □

Table 3: 2×2 ablation: Q-loss × feature gates (Hillstrom, S-Learner architecture, $N = 20$).

Configuration	Q-loss	Gates	Mean $\hat{Q}\downarrow$	Degen↓	Win%↑
S-Learner XGBoost (reference)	×	×	-0.0132 ± 0.0013	0%	0%
S-Learner Neural (MSE)	×	×	-0.0003 ± 0.0012	45%	0%
S-Learner Neural + Gates	×	✓	-0.0208 ± 0.0033	0%	35%
Q-S-Learner, no gates	✓	×	-0.0219 ± 0.0036	0%	20%
Q-S-Learner (full)	✓	✓	$\mathbf{-0.0297 \pm 0.0048}$	0%	45%

Table 4: Q-X-Learner 2×2 ablation (Q-loss × feature gates) under the AIPW target on SynthUplift (70% treated, $N = 20$). Mean $\hat{Q}\pm$std (lower better), degeneracy, and head-to-head win rate.

Configuration	Q-loss	Gates	Mean $\hat{Q}\pm$std↓	Degen↓	Win%↑
X-Learner XGBoost	×	×	-0.0854 ± 0.0085	0%	0%
S-Learner XGBoost	×	×	-0.0859 ± 0.0081	0%	10%
X-Learner Neural (MSE)	×	×	-0.1901 ± 0.1402	10%	0%
X-Learner Neural + Gates	×	✓	-0.2518 ± 0.1279	15%	0%
Q-X-Learner, no gates	✓	×	-0.1480 ± 0.1497	0%	20%
Q-X-Learner (full)	✓	✓	$\mathbf{-0.2783 \pm 0.1103}$	0%	**60%**

7.3 Ablation Details

7.3.1 Q-S Ablation Studies

In the Q-S ablation we hold the S-Learner backbone fixed and toggle the two mechanisms—the Q-loss objective (vs. plain outcome MSE) and the Gumbel feature gates—across the full 2×2 grid on Hillstrom, isolating how much each contributes on top of the neural architecture alone (Table 3).

7.3.2 Q-X Ablation Studies

In the Q-X ablation we repeat the same 2×2 design on the two-arm X-Learner backbone under the AIPW target, on the imbalanced SynthUplift dataset (70% treated) where the Q-X architecture is expected to be most useful (Table 4). Here the two mechanisms are complementary rather than redundant: gates alone lower mean $\hat{Q}$ from -0.1901 to -0.2518 and Q-loss alone reaches -0.1480, but only the full model combines them to -0.2783 at 0% degeneracy and a 60% head-to-head win rate against the strongest baseline—the one setting where both components are jointly required.

7.4 Sensitivity and Target-Comparison Details

7.4.1 Sensitivity Analyses

Beyond the 2×2 decomposition, we vary one hyperparameter at a time around the default configuration to check that the Q-Learner results are not an artifact of a particular setting:

Sparsity weight (λ): Performance is robust across $\lambda \in [0.0001, 0.5]$ on Hillstrom—all values beat XGBoost with 0% degeneracy. Slightly higher sparsity (0.05–0.5) gives marginally better Mean $\hat{Q}$, suggesting aggressive feature pruning helps.

Epoch count: Below 20 epochs, Q-S-Learner degenerates (20–30% rate). The sweet spot is 20–50 epochs (0% degeneracy, best Mean $\hat{Q}$). At 100 epochs, slight overfitting returns (10% degeneracy). We use 30 epochs as default.

Sample size: Q-X-Learner achieves 0% degeneracy and wins at all tested sample sizes (n = 1K to 50K) on imbalanced data, with performance improving monotonically to n = 25K then slightly decreasing at n = 50K (likely due to longer training needed for larger data).

Table 5: AIPW vs. IPW target, Q-S-Learner (paired, 20 iterations; $\Delta Q > 0$ favors AIPW).

Dataset	n	Treat%	Degen IPW→AIPW	Mean ΔQ (IPW−AIPW)	Wilcoxon p	Winner
Megafon	600k	50%	100% → **0%**	+0.0764	< 0.001 ***	**AIPW**
SynthUplift	50k	70%	100% → **0%**	+0.0263	< 0.001 ***	**AIPW**
Criteo	14M	85%	85% → 25%	+0.0175	< 0.001 ***	**AIPW**
RetailHero	200k	50%	80% → 30%	+0.0083	0.03 *	**AIPW**
Hillstrom	64k	67%	35% → **0%**	+0.0036	< 0.001 ***	**AIPW**
Sandercock	19k	50%	95% → 95%	+0.0009	0.04 *	**AIPW**

7.4.2 Training Target: IPW vs. AIPW

We evaluate the doubly-robust (AIPW) target as a drop-in replacement for the raw IPW target (Section 3.1.1), holding architecture, sparsity, and epochs fixed. Table 5 reports the paired comparison: since both variants see the identical biased/eval split each iteration, their per-iteration $\hat{Q}$ values are paired, and we test the difference $\Delta Q = \hat{Q}_{\text{IPW}} - \hat{Q}_{\text{AIPW}}$ ($\Delta Q > 0$ favors AIPW) with a Wilcoxon signed-rank test over 20 iterations.

The AIPW target is a near-strict improvement for Q-S: it improves the expected $\hat{Q}$ on all six datasets, significantly on every one (Megafon, SynthUplift, Criteo, RetailHero, Hillstrom, Sandercock), and is *never significantly worse*. The three most dramatic effects are the degeneracy cures on the datasets where the raw IPW target collapses entirely: Megafon ($100\% \to 0\%$, the largest effect size in the table), SynthUplift ($100\% \to 0\%$), and Criteo ($85\% \to 25\%$), all highly significant.

The benefit is **architecture-specific**. Applying the same AIPW target to Q-X-Learner produces a significant improvement on Megafon, SynthUplift, and Criteo (all $p < 0.001$, the synthetic/imbalanced degeneracy cures); on the remaining five datasets the paired difference is non-significant and mixed in sign. Even where the target comparison is significant, Q-X does not win outright against the best baseline—on Megafon, for instance, `qx.aipw` cures its own degeneracy ($100\% \to 0\%$).

AIPW does not stack on an already-orthogonalized learner. As a control, we test the same substitution on Q-DML, whose default pseudo-outcome is *already* a variance-reduced orthogonalized (Robinson) score $(Y - \hat{m}) \cdot (T - e)/(e(1-e))$ rather than the raw IPW η. Replacing it with the full doubly-robust score $(\hat{\mu}_1 - \hat{\mu}_0) + (T-e)/(e(1-e)) \cdot (Y - \hat{\mu}_T)$ leads to marked improvement: across the 8 datasets, AIPW is significantly better on 3, significantly worse on 2 (Sandercock, RetailHero), and tied on 3 (paired Wilcoxon). As expected, the AIPW target is valuable precisely for learners that train on the *raw IPW* pseudo-outcome (Q-S, and partly Q-X), where it cures degeneracy—but it offers no reliable gain once the target is already orthogonalized, empirically confirming Remark 1 that "DML already approximates the variance-reduced target."

Two honest caveats bound the claim. First, AIPW improves Q-S but does not overturn the regime findings of Section 4.2: tree baselines still win outright on large/structured/imbalanced data (RetailHero). Second, the improvement is on effect-*magnitude* metrics ($\hat{Q}$, and policy value; Section 4.4)—on the Qini ranking metric, tree baselines remain competitive or better on Hillstrom, indicating Q-S estimates effect sizes better than it ranks who-to-treat.

7.5 Per-Family Results and Target Sweeps

Per-dataset commentary (main-table results).

- **Megafon (largest margin)**: Under the raw IPW target Q-S is completely degenerate on Megafon (Table 7); the AIPW target reverses this, and Q-S improves on the baseline (-0.0295 vs. `s.xgb`'s -0.0200) at 0% degeneracy (Table 1)—the largest Q-S margin over the baseline of any dataset here, on a dataset trees were expected to own.
- **Hillstrom**: Q-S improves on the `s.xgb` baseline in mean $\hat{Q}$ (-0.0037 vs. -0.0021) at matched 0% degeneracy (Table 1), and the improvement is stable across iterations.
- **Sandercock**: A near-null-heterogeneity RCT on which no method recovers much signal. The AIPW target yields only a small (though significant) mean-$\hat{Q}$ improvement over IPW

Table 6: Paired within-architecture comparison on ground-truth datasets: the identical DML/R backbone trained/selected with the vanilla loss vs. the Q objective (`q_*`), over 20 iterations. Lower mean $\hat{Q}$ and lower PEHE are better; bold marks the better member of each pair. This isolates the effect of the Q objective holding the estimator family fixed.

		mean $\hat{Q}\pm$std (↓)		PEHE$\pm$std (↓)	
Dataset	Family	vanilla	+Q-loss	vanilla	+Q-loss
SynthUplift	DML (XGB)	-0.0029 ± 0.0020	$\mathbf{-0.0088 \pm 0.0010}$	0.0915 ± 0.0060	$\mathbf{0.0497 \pm 0.0055}$
	DML (Elastic)	$\mathbf{-0.0086 \pm 0.0009}$	-0.0085 ± 0.0010	0.0486 ± 0.0041	$\mathbf{0.0482 \pm 0.0041}$
	DML (Lasso)	-0.0086 ± 0.0009	$\mathbf{-0.0086 \pm 0.0010}$	0.0486 ± 0.0041	$\mathbf{0.0478 \pm 0.0042}$
	R (XGB)	-0.0055 ± 0.0014	$\mathbf{-0.0089 \pm 0.0009}$	0.0751 ± 0.0020	$\mathbf{0.0479 \pm 0.0039}$
	R (Ridge)	-0.0085 ± 0.0010	$\mathbf{-0.0086 \pm 0.0011}$	0.0486 ± 0.0025	$\mathbf{0.0476 \pm 0.0026}$

(Table 5), and Q-S stays ∼95% degenerate—appropriately declining to fabricate heterogeneity where there is essentially none (Section 4.5).

- **Criteo**: Criteo combines extreme imbalance (85% treated) with an extremely sparse outcome (0.43% conversion), the hardest regime for the raw IPW target, on which Q-S is fully degenerate. The AIPW target removes most of this degeneracy (from 85% to 25%; Table 5), and Q-S's mean $\hat{Q}$ (-0.0008) edges the `s.xgb` baseline (-0.0007)—but its 25% degeneracy makes this only a partial recovery. Q-X, by contrast, handles this imbalanced, sparse-outcome regime cleanly: it is non-degenerate and edges the baseline (-0.0009 at 0% degeneracy; Table 1), the case where the two-arm architecture is expected to help most.
- **Synthetic data failure under the generic protocol is informative**: On SynthUplift and Megafon, the CATE signal is programmatic (high SNR), so tree models exploit it directly under the default IPW target. Q-loss gradient SNR $\approx \sigma_{\mathrm{CATE}}/\sigma_\eta \approx$ 5–8% on synthetic vs. ≈13% on Hillstrom—insufficient for neural net convergence at the default settings. Section 4.3 shows the AIPW target directly addresses this SNR gap.
- **RetailHero**: The raw IPW target leaves Q-S degenerate on RetailHero, but the AIPW target recovers a non-degenerate estimate: Q-S reaches -0.0009 at 0% degeneracy, on par with the `s.xgb` baseline (-0.0005, also 0%; Table 1). RetailHero therefore follows the same IPW→AIPW recovery seen on Megafon, rather than being a regime where trees dominate.

7.5.1 Does the Q objective help, holding the architecture fixed?

To isolate the effect of the objective from the choice of estimator, we compare each learner against its Q-loss counterpart on the *same* backbone, on SynthUplift, which has known ground-truth CATE (Table 6). Holding the DML/R backbone fixed, training or selecting via the Q objective substantially improves the flexible (gradient-boosted) learners—roughly halving PEHE (DML-XGB $0.0915 \to 0.0497$; R-XGB $0.0751 \to 0.0479$)—while being neutral on the already-linear backbones. The objective itself, not a particular architecture, therefore drives the improvement where exploitable signal exists.

7.5.2 Q-S Results

Table 7 isolates the training-target choice for Q-S. Under the raw IPW target, Q-S is degenerate ($\hat{Q} \geq 0$) on all six datasets (Hillstrom, Sandercock, SynthUplift, Criteo, RetailHero, and Megafon). Switching to the AIPW target moves mean $\hat{Q}$ below zero on every dataset and drives degeneracy to 0% on the balanced and large structured datasets (Hillstrom, Megafon, and RetailHero). Residual degeneracy remains only on the hardest cases—Criteo (25%, extreme outcome sparsity) and especially Sandercock (95%, near-null heterogeneity), with a trace on SynthUplift (5%). The mean-$\hat{Q}$ improvement is largest precisely where the raw IPW variance is worst—RetailHero ($+0.5134 \to -0.0009$), Megafon ($+0.0647 \to -0.0295$), and Criteo ($+0.0785 \to -0.0008$). We read this as strong evidence that Q-S degeneracy is driven mainly by the high variance of the raw IPW target rather than by the Q-loss objective itself: holding the architecture fixed and changing only the pseudo-outcome removes the degeneracy on most datasets, with the residual failures confined to extreme outcome sparsity or near-null heterogeneity.

Table 7: Q-S robustness sweep: mean $\hat{Q}\pm$std and degeneracy at the baseline configuration ($[64, 32]$, lr 10^{-3}, sparsity 0.01, 20 iterations) under the raw IPW vs. doubly-robust AIPW training target. Lower $\hat{Q}$ is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor. *deg* is the degeneracy rate.

				IPW target		AIPW target	
Dataset	Type	n	Treat%	mean $\hat{Q}\pm$std	deg	mean $\hat{Q}\pm$std	deg
Hillstrom	Real RCT	64k	67%	$+0.0267 \pm 0.0116$	95	-0.0037 ± 0.0005	0
Sandercock	Real RCT	19k	50%	$+0.0324 \pm 0.0087$	100	$+0.0071 \pm 0.0019$	95
SynthUplift	Synthetic	50k	70%	$+0.0290 \pm 0.0084$	100	-0.0040 ± 0.0023	5
Criteo	Real RCT	14M	85%	$+0.0785 \pm 0.0414$	100	-0.0008 ± 0.0001	25
RetailHero	Real RCT	200k	50%	$+0.5134 \pm 0.1273$	100	-0.0009 ± 0.0002	0
Megafon	Synth RCT	600k	50%	$+0.0647 \pm 0.0102$	100	-0.0295 ± 0.0034	0

Table 8: Q-X robustness sweep: mean $\hat{Q}\pm$std and degeneracy at the baseline configuration ($[64, 32]$, lr 10^{-3}, sparsity 0.01, 20 iterations) under the raw IPW vs. doubly-robust AIPW training target. Lower $\hat{Q}$ is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor. *deg* is the degeneracy rate.

				IPW target		AIPW target	
Dataset	Type	n	Treat%	mean $\hat{Q}\pm$std	deg	mean $\hat{Q}\pm$std	deg
Hillstrom	Real RCT	64k	67%	$+0.0303 \pm 0.0113$	100	-0.0035 ± 0.0006	0
Sandercock	Real RCT	19k	50%	$+0.0129 \pm 0.0036$	100	-0.007 ± 0.0001	0
SynthUplift	Synthetic	50k	70%	$+0.0224 \pm 0.0052$	100	-0.0080 ± 0.0016	0
Criteo	Real RCT	14M	85%	$+0.0539 \pm 0.0321$	100	-0.0006 ± 0.0005	20
RetailHero	Real RCT	200k	50%	$+0.4978 \pm 0.1242$	100	-0.0009 ± 0.0005	10
Megafon	Synth RCT	600k	50%	$+0.0513 \pm 0.0095$	100	-0.0227 ± 0.0011	0

7.5.3 Q-X Results

Table 8 repeats the target comparison for the two-arm Q-X architecture. The IPW pattern mirrors Q-S—degenerate on the same six datasets—and AIPW again moves every mean $\hat{Q}$ below zero. As with Q-S, some residual degeneracy survives the switch: Q-X retains 20% on Criteo and 10% on RetailHero, the imbalanced and sparse-outcome cases where the two-arm pretraining is least stable, while reaching 0% on the remaining datasets. The AIPW target is therefore *necessary* but not always *sufficient* for Q-X, just as for Q-S. We present Q-S and Q-X as two complementary instantiations of Q-loss matched to different data regimes—the single-model balanced setting and the two-arm imbalanced setting (Section 3.3)—rather than ranking one above the other; their degeneracy profiles differ across datasets, with neither uniformly lower.

7.5.4 Q-DML Results

Table 9 reports Q-DML, which trains on the orthogonal Robinson target and therefore has no IPW/AIPW split. Its behaviour is qualitatively different from Q-S and Q-X: rather than degenerating under IPW and being rescued by AIPW, it is *already* non-degenerate on the large, structured datasets (0% degeneracy on Hillstrom, SynthUplift, RetailHero, and Megafon), consistent with the orthogonalization being built into its objective. It wins the majority of head-to-head iterations against `s.xgb` on Hillstrom and Megafon (both 100%), Sandercock (60%), and RetailHero (40%), but not on the high-SNR synthetic and sparse-outcome datasets SynthUplift and Criteo (0%), where trees exploit the programmatic signal directly (Section 4.2). One dataset remains hard for every method: Sandercock (85% degenerate, near-null heterogeneity)—its 60% win rate reflects Q-DML being the less-degenerate of two weak estimators rather than a genuine success. Overall, Q-DML is the most stable Q-variant on large data but adds little over a strong tree baseline where that baseline already captures the signal.

Table 9: Q-DML robustness sweep: mean $\hat{Q}\pm$std, degeneracy, and head-to-head win rate vs. `s.xgb` at the baseline configuration ($[64, 32]$, lr 10^{-3}, sparsity 0.01, 20 iterations). Q-DML uses the Robinson/R-loss target (orthogonal by construction; no IPW/AIPW split).

Dataset	Type	n	Treat%	mean $\hat{Q}\pm$std	deg	win vs `s.xgb`
Hillstrom	Real RCT	64k	67%	-0.0032 ± 0.0005	0	100%
Sandercock	Real RCT	19k	50%	$+0.0007 \pm 0.0009$	85	60%
SynthUplift	Synthetic	50k	70%	-0.0028 ± 0.0006	0	0%
Criteo	Real RCT	14M	85%	-0.0001 ± 0.0000	10	0%
RetailHero	Real RCT	200k	50%	-0.0009 ± 0.0002	0	40%
Megafon	Synth RCT	600k	50%	-0.0351 ± 0.0033	0	100%

Table 10: R-Learner: best vanilla vs. best Q-loss-augmented variant per dataset.

Dataset	Best vanilla R	Win%	Degen	Best Q-R	Win%	Degen
Hillstrom	`r.ridge` (-0.0035)	30%	0%	`q_r.xgb` (-0.0030)	0%	0%
Sandercock	`r.ridge` ($+0.0006$)	5%	95%	`q_r.ridge` ($+0.0010$)	5%	95%
SynthUplift	`r.ridge` (-0.0085)	10%	0%	`q_r.xgb` (-0.0088)	35%	0%
RetailHero	`r.ridge` (-0.0009)	0%	0%	`q_r.xgb` (-0.0005)	10%	25%

7.6 Additional Experimental Results: Vanilla vs. Q-Loss-Augmented Baselines

This appendix reports the protocol-matched comparison of vanilla R-Learner and DML baselines against their Q-loss-augmented counterparts. All results use the identical observational sampling protocol as Table 1 (20 iterations, bias strength 3.0).

`q_r.xgb` selects its final configuration by Q-statistic on a held-out split while fitting the same R-loss-weighted target as vanilla R-Learner; `q_r.ridge` instead refines the R-Learner-Ridge estimate with an additional neural network trained directly on the raw-IPW Q-loss (Remark 1).

Each Q-DML variant (`q_dml.xgb`, `q_dml.elastic`, `q_dml.lasso`) selects its configuration by Q-statistic on a held-out split while fitting the same orthogonalized (Robinson) target as its vanilla counterpart (Remark 1).

7.7 Feature-Gate Impact

Table 19 isolates the Gumbel feature gate by sweeping its sparsity weight (the L_1 penalty on gate probabilities) from 0.001 to 0.1 for Q-S-AIPW; larger weights prune more covariates. Performance is largely insensitive to gate strength on most datasets—the gate delivers interpretable feature selection at little accuracy cost—with modest gains from stronger pruning on SynthUplift and Megafon and a reduction in degeneracy on Sandercock.

8 Detailed Derivations

This appendix expands the results stated in Section 3. Throughout, $e(x) = \Pr(T = 1 \mid X = x)$ is the (known, in an RCT) propensity, $\mu_t(x) = \mathbb{E}[Y \mid T = t, X = x]$, and $\tau(x) = \mu_1(x) - \mu_0(x)$.

8.1 The Q-parameter as shifted CATE-MSE

The precision in estimating heterogeneous effects (PEHE) is

$$P(\hat{\tau}) = \mathbb{E}_X\left[(\tau(X) - \hat{\tau}(X))^2\right] = \underbrace{\mathbb{E}_X[\tau(X)^2]}_{P_1} + \underbrace{\mathbb{E}_X[\hat{\tau}(X)^2]}_{P_2} - 2\underbrace{\mathbb{E}_X[\tau(X)\hat{\tau}(X)]}_{P_3}. \tag{11}$$

Table 11: DML: best vanilla vs. best Q-loss-augmented variant per dataset.

Dataset	Best vanilla DML	Win%	Degen	Best Q-DML	Win%	Degen
Hillstrom	`dml.lasso` (−0.0035)	20%	0%	`q_dml.elastic` (−0.0033)	0%	0%
Sandercock	`dml.elastic` (+0.0001)	50%	65%	`q_dml.lasso` (+0.0014)	0%	95%
SynthUplift	`dml.elastic` (−0.0086)	0%	0%	`q_dml.xgb` (−0.0087)	20%	0%
RetailHero	`dml.elastic` (−0.0010)	10%	0%	`q_dml.xgb` (−0.0008)	25%	20%

Table 12: Per-model results on Hillstrom (20 iterations, observational sampling, bias_strength 3.0). Sorted by mean $\hat{Q}$ (lower is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor). † = Q-loss learner.

Model	mean $\hat{Q}\pm$std	degen%	Qini	PV
`q_s.aipw`†	-0.0037 ± 0.0005	0	2.3551 ± 19.7549	0.1669 ± 0.0028
`qx.aipw`†	-0.0035 ± 0.0006	0	42.7773 ± 18.4398	0.1664 ± 0.0028
`q_dml.elastic`†	-0.0033 ± 0.0004	0	4.1870 ± 22.5411	0.1669 ± 0.0028
`q_dml.lasso`†	-0.0033 ± 0.0004	0	2.3551 ± 19.7549	0.1669 ± 0.0028
`q_r.xgb`†	-0.0031 ± 0.0006	0	27.9101 ± 17.9407	0.1652 ± 0.0035
`q_dml.xgb`†	-0.0030 ± 0.0005	0	24.1515 ± 26.7547	0.1651 ± 0.0030
`q_r.ridge`†	-0.0029 ± 0.0007	0	25.3856 ± 34.0283	0.1613 ± 0.0043
`s.xgb`	-0.0021 ± 0.0007	0	27.8901 ± 18.0961	0.1630 ± 0.0033
`s.ext.xgb`	-0.0012 ± 0.0010	10	32.6079 ± 22.3963	0.1617 ± 0.0033
`x.xgb`	$+0.0004 \pm 0.0010$	60	18.6780 ± 19.7179	0.1589 ± 0.0037
`q_s.sparse`†	$+0.0019 \pm 0.0105$	20	-4.3090 ± 20.7838	0.1604 ± 0.0138
`t.xgb`	$+0.0025 \pm 0.0013$	100	26.7713 ± 17.6274	0.1565 ± 0.0036
`qx.dense`†	$+0.0255 \pm 0.0088$	100	-1.1636 ± 39.7190	0.1407 ± 0.0075
`qx.sparse`†	$+0.0257 \pm 0.0096$	95	1.3632 ± 38.0919	0.1406 ± 0.0067
`qx.medium`†	$+0.0266 \pm 0.0099$	100	-0.8518 ± 38.9214	0.1405 ± 0.0069
`q_s.medium`†	$+0.0284 \pm 0.0131$	95	0.8778 ± 40.7513	0.1416 ± 0.0064
`q_s.dense`†	$+0.0295 \pm 0.0134$	100	-1.6054 ± 39.0264	0.1413 ± 0.0066

The term P_1 does not depend on $\hat{\tau}$. Defining the *Q-parameter* $Q(\hat{\tau}) = P_2(\hat{\tau}) - 2P_3(\hat{\tau}) = P(\hat{\tau}) - P_1$, minimizing Q is equivalent to minimizing PEHE, and $Q(\hat{\tau}) \geq 0$ certifies that $\hat{\tau}$ is worse than the zero-effect predictor $\hat{\tau}_0 \equiv 0$ (for which $Q = 0$).

8.2 Unbiasedness of the IPW pseudo-outcome

Let $\eta(x, t, y) = \dfrac{y\,(t - e(x))}{e(x)(1 - e(x))}$. Under unconfoundedness and known $e(x)$,

$$\mathbb{E}[\eta \mid X = x] = \frac{e(x)\,\mu_1(x)\,(1 - e(x)) + (1 - e(x))\,\mu_0(x)\,(-e(x))}{e(x)(1 - e(x))} \tag{12}$$

$$= \mu_1(x) - \mu_0(x) = \tau(x). \tag{13}$$

Hence $\mathbb{E}_X[\hat{\tau}(X)\,\eta] = \mathbb{E}_X[\hat{\tau}(X)\tau(X)] = P_3$, and the sample statistic $\hat{Q}(\hat{\tau}) = \frac{1}{N}\sum_i [\hat{\tau}(x_i)^2 - 2\hat{\tau}(x_i)\eta_i]$ satisfies $\mathbb{E}[\hat{Q}(\hat{\tau})] = Q(\hat{\tau})$ (Lemma 1).

8.3 Q-loss is regression on the pseudo-outcome

Completing the square,

$$\hat{Q}(\hat{\tau}) = \frac{1}{N}\sum_{i=1}^{N} \big(\hat{\tau}(x_i) - \eta_i\big)^2 - \underbrace{\frac{1}{N}\sum_{i=1}^{N} \eta_i^2}_{\text{independent of } \hat{\tau}}. \tag{14}$$

Thus *minimizing $\hat{Q}$ is exactly least-squares regression of the pseudo-outcome η on x*, with the constant $\frac{1}{N}\sum_i \eta_i^2$ shifting the objective but not its minimizer. This is the sense in which the Q-statistic,

Table 13: Per-model results on SynthUplift (20 iterations, observational sampling, bias_strength 3.0). Sorted by mean $\hat{Q}$ (lower is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor). PEHE/regret shown (true τ known). † = Q-loss learner.

Model	mean $\hat{Q}\pm$std	degen%	PEHE	regret	Qini	PV
q_r.xgb†	-0.0089 ± 0.0009	0	0.0479 ± 0.0039	0.0022 ± 0.0007	244.3655 ± 12.1540	0.1632 ± 0.0057
q_dml.xgb†	-0.0088 ± 0.0010	0	0.0497 ± 0.0055	0.0025 ± 0.0011	237.0808 ± 11.6724	0.1628 ± 0.0050
q_dml.lasso†	-0.0086 ± 0.0010	0	0.0478 ± 0.0042	0.0044 ± 0.0009	228.8493 ± 11.0270	0.1601 ± 0.0043
q_r.ridge†	-0.0086 ± 0.0011	0	0.0476 ± 0.0026	0.0038 ± 0.0006	226.9382 ± 13.5162	0.1605 ± 0.0048
q_dml.elastic†	-0.0085 ± 0.0010	0	0.0482 ± 0.0041	0.0045 ± 0.0009	228.4165 ± 10.8910	0.1600 ± 0.0042
s.xgb	-0.0084 ± 0.0016	0	0.0495 ± 0.0029	0.0033 ± 0.0005	231.5099 ± 15.4755	0.1611 ± 0.0060
qx.aipw†	-0.0078 ± 0.0014	0	0.0532 ± 0.0039	0.0050 ± 0.0008	216.7531 ± 14.5934	0.1594 ± 0.0054
s.ext.xgb	-0.0071 ± 0.0017	0	0.0610 ± 0.0019	0.0046 ± 0.0006	220.2087 ± 15.8540	0.1602 ± 0.0056
x.xgb	-0.0048 ± 0.0020	0	0.0795 ± 0.0040	0.0069 ± 0.0007	205.2619 ± 11.8960	0.1577 ± 0.0047
q_s.aipw†	-0.0040 ± 0.0023	5	0.0807 ± 0.0075	0.0090 ± 0.0020	190.8531 ± 17.2421	0.1553 ± 0.0049
t.xgb	-0.0005 ± 0.0022	40	0.1004 ± 0.0038	0.0096 ± 0.0007	177.8794 ± 14.3797	0.1551 ± 0.0052
q_s.sparse†	$+0.0115 \pm 0.0099$	90	0.1480 ± 0.0330	0.0255 ± 0.0101	85.3719 ± 56.2652	0.1410 ± 0.0099
qx.dense†	$+0.0148 \pm 0.0051$	100	0.1592 ± 0.0139	0.0167 ± 0.0050	128.7901 ± 43.7553	0.1485 ± 0.0050
qx.sparse†	$+0.0154 \pm 0.0048$	100	0.1603 ± 0.0134	0.0168 ± 0.0050	128.6304 ± 43.3446	0.1484 ± 0.0051
qx.medium†	$+0.0163 \pm 0.0056$	100	0.1634 ± 0.0169	0.0170 ± 0.0048	127.0204 ± 47.9096	0.1480 ± 0.0053
q_s.medium†	$+0.0292 \pm 0.0069$	100	0.1990 ± 0.0177	0.0212 ± 0.0064	116.4952 ± 46.1848	0.1430 ± 0.0060
q_s.dense†	$+0.0322 \pm 0.0098$	100	0.2051 ± 0.0222	0.0209 ± 0.0051	116.9910 ± 44.2867	0.1442 ± 0.0055

Table 14: Per-model results on Sandercock (20 iterations, observational sampling, bias_strength 3.0). Sorted by mean $\hat{Q}$ (lower is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor). † = Q-loss learner.

Model	mean $\hat{Q}\pm$std	degen%	Qini	PV
s.xgb	$+0.0007 \pm 0.0005$	90	0.7013 ± 5.1701	0.0926 ± 0.0041
q_r.ridge†	$+0.0009 \pm 0.0005$	100	2.2779 ± 6.6753	0.0917 ± 0.0055
qx.aipw†	$+0.0013 \pm 0.0008$	100	-0.8274 ± 5.2273	0.0905 ± 0.0045
q_dml.lasso†	$+0.0015 \pm 0.0010$	90	1.0413 ± 6.8188	0.0917 ± 0.0043
q_dml.elastic†	$+0.0017 \pm 0.0011$	100	0.9836 ± 6.7968	0.0920 ± 0.0044
q_r.xgb†	$+0.0019 \pm 0.0011$	100	-1.3645 ± 6.2694	0.0915 ± 0.0034
q_dml.xgb†	$+0.0020 \pm 0.0012$	95	-2.4954 ± 7.3492	0.0910 ± 0.0041
s.ext.xgb	$+0.0046 \pm 0.0019$	100	0.6401 ± 6.9119	0.0919 ± 0.0033
qx.dense†	$+0.0056 \pm 0.0020$	100	0.7432 ± 4.7756	0.0923 ± 0.0038
qx.medium†	$+0.0064 \pm 0.0026$	100	-2.6552 ± 5.3709	0.0897 ± 0.0046
qx.sparse†	$+0.0067 \pm 0.0021$	100	-0.6018 ± 6.2349	0.0908 ± 0.0042
q_s.aipw†	$+0.0071 \pm 0.0019$	100	-0.3561 ± 7.7474	0.0917 ± 0.0062
x.xgb	$+0.0085 \pm 0.0033$	100	0.7270 ± 7.5468	0.0922 ± 0.0033
t.xgb	$+0.0156 \pm 0.0040$	100	1.0488 ± 7.0404	0.0920 ± 0.0030
q_s.sparse†	$+0.0304 \pm 0.0083$	100	-0.2417 ± 5.6634	0.0908 ± 0.0037
q_s.medium†	$+0.0324 \pm 0.0074$	100	0.0380 ± 6.5829	0.0909 ± 0.0052
q_s.dense†	$+0.0330 \pm 0.0087$	100	1.1506 ± 5.5529	0.0913 ± 0.0046

an *evaluation* metric, is repurposed as a *training* objective: any regressor $\hat{\tau}(x;\theta)$ trained to predict η minimizes $\hat{Q}$ in expectation, with gradient $\partial\hat{Q}/\partial\theta = \frac{2}{N}\sum_i(\hat{\tau}(x_i;\theta) - \eta_i)\,\partial_\theta\hat{\tau}(x_i;\theta)$.

8.4 Equivalence of the AIPW target to the DR-Learner second stage

Replace η by the augmented (doubly-robust) pseudo-outcome

$$\eta_{\mathrm{DR}}(x,t,y) = \big(\hat{\mu}_1(x) - \hat{\mu}_0(x)\big) + \frac{t - e(x)}{e(x)(1-e(x))}\big(y - \hat{\mu}_t(x)\big). \tag{15}$$

Taking conditional expectations and writing $\bar{\mu}_t$ for the estimand of $\hat{\mu}_t$,

$$\mathbb{E}[\eta_{\mathrm{DR}} \mid X = x] - \tau(x) = \Big(1 - \tfrac{\bar{e}(x)}{e(x)}\Big)\big(\mu_1(x) - \bar{\mu}_1(x)\big) - \Big(1 - \tfrac{1-\bar{e}(x)}{1-e(x)}\Big)\big(\mu_0(x) - \bar{\mu}_0(x)\big), \tag{16}$$

Table 15: Per-model results on Retailhero (20 iterations, observational sampling, bias_strength 3.0). Sorted by mean $\hat{Q}$ (lower is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor). † = Q-loss learner.

Model	mean $\hat{Q}\pm$std	degen%	Qini	PV
q_s.aipw†	-0.0009 ± 0.0002	0	-1.4292 ± 33.4384	0.6363 ± 0.0019
qx.aipw†	-0.0009 ± 0.0005	10	129.0435 ± 33.8855	0.6348 ± 0.0031
q_dml.xgb†	-0.0008 ± 0.0006	20	119.3283 ± 46.0716	0.6333 ± 0.0028
q_dml.lasso†	-0.0006 ± 0.0005	15	9.3152 ± 44.5840	0.6340 ± 0.0074
q_dml.elastic†	-0.0006 ± 0.0005	20	9.7902 ± 45.7699	0.6340 ± 0.0074
s.xgb	-0.0005 ± 0.0003	0	112.8454 ± 32.9944	0.6347 ± 0.0025
q_r.xgb†	-0.0005 ± 0.0008	25	107.9969 ± 46.2632	0.6327 ± 0.0030
s.ext.xgb	-0.0001 ± 0.0005	35	103.4176 ± 27.1922	0.6318 ± 0.0030
x.xgb	$+0.0006 \pm 0.0008$	80	81.1683 ± 33.0083	0.6305 ± 0.0031
t.xgb	$+0.0027 \pm 0.0008$	100	64.4341 ± 28.8507	0.6288 ± 0.0025
q_r.ridge†	$+0.0038 \pm 0.0019$	100	49.7085 ± 77.8454	0.6264 ± 0.0060
qx.sparse†	$+0.4826 \pm 0.1281$	100	23.0896 ± 89.2599	0.6227 ± 0.0066
qx.dense†	$+0.4834 \pm 0.1255$	100	24.1363 ± 89.7367	0.6226 ± 0.0069
qx.medium†	$+0.4838 \pm 0.1268$	100	22.6933 ± 87.6891	0.6227 ± 0.0065
q_s.sparse†	$+0.4840 \pm 0.1178$	100	29.9918 ± 82.5314	0.6230 ± 0.0070
q_s.medium†	$+0.4896 \pm 0.1360$	100	32.9757 ± 81.4302	0.6226 ± 0.0065
q_s.dense†	$+0.4945 \pm 0.1264$	100	31.0981 ± 82.1344	0.6224 ± 0.0067

Table 16: Per-model results on Criteo (20 iterations, observational sampling, bias_strength 3.0). Sorted by mean $\hat{Q}$ (lower is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor). † = Q-loss learner.

Model	mean $\hat{Q}\pm$std	degen%	Qini	PV
qx.aipw†	-0.0009 ± 0.0002	0	350.6313 ± 60.2811	0.0479 ± 0.0008
q_s.aipw†	-0.0008 ± 0.0001	25	16.9491 ± 52.3417	0.0469 ± 0.0038
s.xgb	-0.0007 ± 0.0002	0	345.6868 ± 74.9332	0.0481 ± 0.0008
s.ext.xgb	-0.0007 ± 0.0002	0	321.1674 ± 89.0679	0.0475 ± 0.0011
q_r.xgb†	-0.0005 ± 0.0002	0	306.5456 ± 77.0250	0.0476 ± 0.0007
q_dml.elastic†	-0.0001 ± 0.0003	20	103.9892 ± 161.7723	0.0474 ± 0.0028
q_dml.lasso†	-0.0000 ± 0.0004	25	92.2125 ± 151.8974	0.0470 ± 0.0032
x.xgb	$+0.0001 \pm 0.0005$	50	211.1112 ± 82.8221	0.0466 ± 0.0010
q_dml.xgb†	$+0.0002 \pm 0.0008$	50	319.3913 ± 88.9293	0.0479 ± 0.0010
q_r.ridge†	$+0.0007 \pm 0.0008$	80	96.9991 ± 170.5457	0.0452 ± 0.0019
t.xgb	$+0.0008 \pm 0.0005$	100	168.6173 ± 77.5729	0.0460 ± 0.0009
q_s.sparse†	$+0.0172 \pm 0.0252$	75	-93.6694 ± 255.1469	0.0439 ± 0.0042
qx.medium†	$+0.0641 \pm 0.0348$	100	-66.3695 ± 278.4329	0.0436 ± 0.0028
qx.dense†	$+0.0650 \pm 0.0340$	100	-61.6988 ± 275.7519	0.0436 ± 0.0028
qx.sparse†	$+0.0662 \pm 0.0481$	100	-45.4959 ± 291.4514	0.0438 ± 0.0029
q_s.medium†	$+0.0703 \pm 0.0354$	100	-72.9714 ± 274.3878	0.0435 ± 0.0027
q_s.dense†	$+0.0777 \pm 0.0364$	100	-42.5860 ± 265.4093	0.0439 ± 0.0027

which vanishes if *either* the propensity ($\bar{e} = e$) *or* the outcome models ($\bar{\mu}_t = \mu_t$) are correct—the double-robustness property. In an RCT e is known exactly, so η_{DR} is unbiased for $\tau(x)$ regardless of the quality of $\hat{\mu}_t$; the outcome models affect only variance.

8.5 Variance reduction, IPW vs. AIPW

Both targets share the population minimizer $\hat{\tau}^\star = \tau$ (Appendix 8.2), so the choice is one of variance, not bias. The IPW target passes the full outcome Y through the weight $(T - e)/(e(1 - e))$; the AIPW target passes only the residual $Y - \hat{\mu}_T$. When $\hat{\mu}_T$ captures most of the outcome variation,

Table 17: Per-model results on Megafon (20 iterations, observational sampling, bias_strength 3.0). Sorted by mean $\hat{Q}$ (lower is better; $\hat{Q} \geq 0$ is worse than the zero-effect predictor). † = Q-loss learner.

Model	mean $\hat{Q}\pm$std	degen%	Qini	PV
q_s.aipw†	-0.0295 ± 0.0034	0	3794.2385 ± 79.6700	0.2616 ± 0.0019
t.xgb	-0.0264 ± 0.0012	0	3526.4837 ± 76.4238	0.2620 ± 0.0011
q_r.xgb†	-0.0258 ± 0.0013	0	3519.7269 ± 80.2869	0.2611 ± 0.0011
q_dml.xgb†	-0.0257 ± 0.0014	0	3516.7999 ± 77.7305	0.2610 ± 0.0013
x.xgb	-0.0252 ± 0.0007	0	3541.5270 ± 58.9545	0.2623 ± 0.0012
s.ext.xgb	-0.0239 ± 0.0008	0	3458.2457 ± 70.9271	0.2615 ± 0.0014
qx.aipw†	-0.0213 ± 0.0011	0	3293.5675 ± 79.6716	0.2576 ± 0.0017
s.xgb	-0.0200 ± 0.0008	0	3241.9013 ± 89.4978	0.2578 ± 0.0013
q_r.ridge†	-0.0075 ± 0.0014	0	1601.6481 ± 234.7343	0.2358 ± 0.0031
q_dml.lasso†	$+0.0014 \pm 0.0026$	55	142.0169 ± 139.8567	0.2205 ± 0.0047
q_dml.elastic†	$+0.0024 \pm 0.0032$	70	156.4946 ± 144.6492	0.2195 ± 0.0045
qx.dense†	$+0.0452 \pm 0.0075$	100	1377.4736 ± 126.1930	0.2242 ± 0.0031
qx.medium†	$+0.0461 \pm 0.0084$	100	1368.7158 ± 110.9434	0.2243 ± 0.0031
qx.sparse†	$+0.0470 \pm 0.0090$	100	1294.3468 ± 129.7675	0.2237 ± 0.0034
q_s.sparse†	$+0.0550 \pm 0.0144$	100	1963.1337 ± 150.9267	0.2310 ± 0.0033
q_s.medium†	$+0.0635 \pm 0.0102$	100	2134.4406 ± 145.1145	0.2333 ± 0.0044
q_s.dense†	$+0.0671 \pm 0.0123$	100	2124.8190 ± 137.0104	0.2327 ± 0.0038

Table 18: Benchmark datasets. n, treatment rate, and covariate count are as used in our pipeline; "True τ" marks datasets with known ground-truth CATE (enabling PEHE/regret). Outcome rate is the marginal $\mathbb{E}[Y]$ (event rate for binary outcomes).

Dataset	Domain	Type	n	Treat%	Feat.	Outcome (rate)	True τ
Hillstrom	Email marketing	Real RCT	64,000	67%	5	visit (0.147)	no
SynthUplift	Payment routing (sim.)	Synthetic	50,000	70%	12	decline (0.111)	yes
Sandercock	Stroke trial (IST)	Real RCT	19,435	50%	13	outcome (0.092)	no
Criteo	Online advertising	Real RCT	14,000,000	85%	12	visit (0.047)	no
RetailHero	Retail promotion (X5)	Real RCT	200,000	50%	10	purchase (0.620)	no
Megafon	Telecom uplift (sim.)	Synthetic RCT	600,000	50%	50	conversion (0.204)	yes

$\mathrm{Var}(\eta_{\mathrm{DR}}) \ll \mathrm{Var}(\eta_{\mathrm{IPW}})$, raising the gradient signal-to-noise ratio and curing the degeneracy the raw IPW target exhibits in low-SNR regimes (Section 4.3). The boundary case is finite-sample: when n is too small to estimate $\hat{\mu}_t$ well, the outcome-model error inflates variance and the nuisance-free IPW target is preferable.

8.6 Empirical Test of Proposition 3: Causal Feature Selection

We test Proposition 3 on synthetic RCT data ($n = 8000$, 5 seeds, 150 epochs, Gumbel temperature annealed $2.0 \to 0.05$) with three planted feature roles: an effect *modifier* M ($\tau(x) = 2x_0$), a strongly *prognostic* non-modifier Z (contributing $3x_1$ to the baseline outcome but nothing to τ), and six *noise* features N. Treatment is randomized ($e = 0.5$), so *every* feature is non-confounding; the question is therefore whether the gate selects on *effect-modification* rather than on outcome-predictiveness. We report the learned gate probability of each role for three training targets—outcome MSE, Q-loss with the IPW target, and Q-loss with the AIPW target—across sparsity weights λ.

Table 20 and Figure 2 show a clean separation. **Under the AIPW target**, the gate keeps the modifier M (weight ≈ 0.86, flat in λ) while driving the prognostic Z and noise N toward zero—from 0.69 at $\lambda = 0$ to 0.09 at $\lambda = 1$. The separation is present *even at* $\lambda = 0$ ($Z : 0.69$ vs. $M : 0.86$), so it is not an artifact of the L_1 gate penalty: the doubly-robust Q-loss itself supplies the force. **MSE and the raw IPW target do not select causally**—their gates stay uniformly high and, only under strong λ, shrink *all* features together (the modifier included), i.e. brute-force sparsity rather than effect-modifier retention. This confirms Proposition 3 in its AIPW-scoped form and explains it mechanistically: the AIPW target's outcome-model correction removes the prognostic baseline from

Table 19: Feature-gate impact: mean $\hat{Q}$ and degeneracy % (in parentheses) for Q-S-AIPW as the Gumbel-gate sparsity weight varies ($[64, 32]$, lr 10^{-3}, seed 0, 20 iterations). The sparsity weight is the L_1 penalty on the gate probabilities: larger values prune more features. Lower $\hat{Q}$ is better.

Dataset	sparsity 0.001	0.01	0.05	0.1
Hillstrom	-0.0031 (0%)	-0.0036 (0%)	-0.0037 (0%)	-0.0037 (0%)
Sandercock	$+0.0070$ (100%)	$+0.0067$ (100%)	$+0.0055$ (100%)	$+0.0014$ (85%)
SynthUplift	-0.0037 (5%)	-0.0044 (0%)	-0.0059 (0%)	-0.0027 (5%)
Criteo	$+0.0009$ (70%)	$+0.0001$ (35%)	-0.0000 (40%)	-0.0000 (35%)
RetailHero	-0.0005 (20%)	-0.0009 (0%)	-0.0009 (0%)	-0.0009 (0%)
Megafon	-0.0273 (0%)	-0.0293 (0%)	-0.0348 (0%)	-0.0251 (0%)

Table 20: Gate probability per feature role (mean over 5 seeds) by training target and sparsity weight λ. M: effect modifier (should be kept); Z: prognostic non-modifier (Proposition 3: pruned); N: noise. Only the AIPW target prunes Z while retaining M; per-seed spread is shown in Figure 2.

Target	λ	gate(M)	gate(Z)	gate(N)
MSE	0.00	0.840	0.812	0.806
MSE	0.10	0.838	0.811	0.792
MSE	1.00	0.827	0.717	0.603
Q-loss (IPW)	0.00	0.838	0.815	0.804
Q-loss (IPW)	0.10	0.836	0.808	0.787
Q-loss (IPW)	1.00	0.830	0.740	0.594
Q-loss (AIPW)	0.00	0.861	0.693	0.645
Q-loss (AIPW)	0.10	**0.861**	**0.358**	0.247
Q-loss (AIPW)	0.50	**0.857**	**0.101**	0.095
Q-loss (AIPW)	1.00	**0.853**	**0.085**	0.083

the pseudo-outcome, leaving the gate to key on τ-relevant variation. It also delimits the claim—the effect is a finite-sample, variance-driven phenomenon realized under the low-variance target, not a population-gradient one, consistent with the argument in Proposition 3.

8.7 Q-Learner Training Algorithm

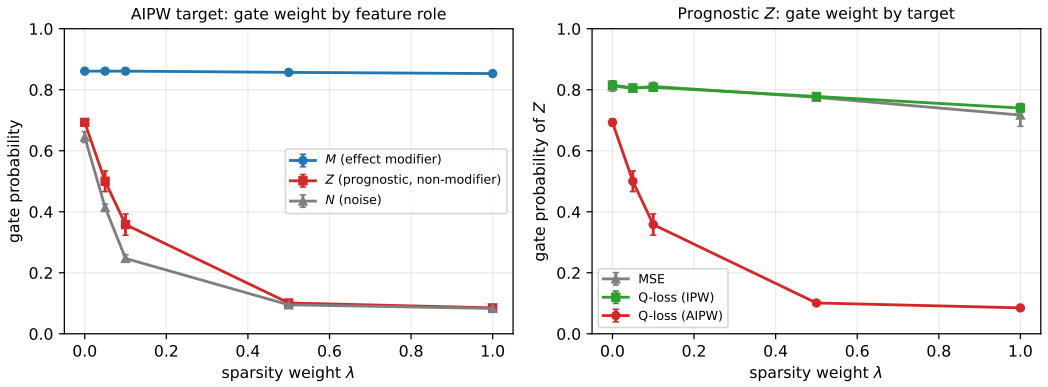

Figure 2: Gate weights in the Proposition 3 synthetic test. **Left:** under the AIPW target the modifier M is retained (weight ≈ 0.86, flat in λ) while the prognostic Z and noise N collapse as λ grows. **Right:** only the AIPW target prunes the prognostic Z; MSE and IPW keep it, their gates shrinking only under brute-force L_1 (which also shrinks the modifier).

Table 21: Q-statistic vs. policy value agreement, real RCT datasets (Q-S/Q-X, 20 iterations).

Dataset	n	Treat%	Q-S/QX wins Q-stat	Q-S/QX wins PV	Q-stat↔PV agree	Within-model ρ
Hillstrom	64k	67%	11/20	**13/20**	**18/20 (90%)**	**0.726*****
Sandercock	19k	50%	12/20	5/20	13/20 (65%)	0.235

Algorithm 1 Q-Learner training (generic across the S/X/DML backbones).

Require: data $\{(x_i, t_i, y_i)\}_{i=1}^{N}$; target $\in \{\text{IPW}, \text{AIPW}\}$; backbone $\hat{\tau}(\cdot\,; \theta)$; epochs E; sparsity weight λ; temperature schedule $\beta : 1.0 \to 0.1$
Nuisances (cross-fitted, out-of-fold): propensity e (known RCT constant, or estimated $\hat{e}(x)$); if AIPW, outcome models $\hat{\mu}_0(x), \hat{\mu}_1(x)$.

for $i = 1, \ldots, N$ **do** ▷ pseudo-outcome, $\mathbb{E}[\eta_i \mid x_i] = \tau(x_i)$
  if target = IPW **then**
    $\eta_i \leftarrow y_i(t_i - e) \,/\, \big(e(1-e)\big)$
  else
    $\eta_i \leftarrow (\hat{\mu}_1 - \hat{\mu}_0)(x_i) + \frac{t_i - e}{e(1-e)}\big(y_i - \hat{\mu}_{t_i}(x_i)\big)$
  end if
end for
initialize gate logits g and backbone parameters θ
for epoch $= 1, \ldots, E$ **do**
  $m \leftarrow \sigma\big((g + \epsilon)/\beta\big)$, $\epsilon \sim \text{Gumbel}(0, 1)$; $\tilde{x}_i \leftarrow m \odot x_i$
  $\hat{\tau}_i \leftarrow \hat{\tau}(\tilde{x}_i; \theta)$ ▷ S / X / DML head
  $\mathcal{L} \leftarrow \frac{1}{N}\sum_i \big[\hat{\tau}_i{}^2 - 2\,\hat{\tau}_i\,\eta_i\big] + \lambda\,\frac{1}{d}\sum_j \sigma(g_j)$ ▷ Q-loss + sparsity
  $(\theta, g) \leftarrow$ optimizer step on $\nabla\mathcal{L}$; anneal β ▷ end-to-end
end for
Ensure: trained $\hat{\tau}(\cdot)$ and learned feature gates g