NEAR: NEIGHBORHOOD EFFECT-AWARE REGULARIZATION FOR TWO-STAGE CATE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the conditional average treatment effect (CATE) from observational data is a central task in causal inference. A common approach is two-stage CATE learning, which first estimates nuisance functions and then uses them to fit a target CATE model. However, the resulting second-stage CATE targets or losses depend on estimated nuisance functions and can be noisy. Existing methods use regularization and information sharing to stabilize second-stage estimation, yet it remains unclear how noisy first-stage CATE estimates should guide the selection of observations to regularize together. We propose Neighborhood Effect-Aware Regularization (NEAR), a modular second-stage regularizer that uses first-stage plug-in CATE estimates to construct a fixed graph and regularizes CATE predictions over pairs with similar plug-in CATEs. We show how plug-in CATE estimation errors control true CATE variation along the selected pairs and derive finite-sample CATE risk bounds for squared-loss second-stage regression using a stabilized version of the doubly robust pseudo-outcome. Across six two-stage learners and three benchmarks, NEAR lowers mean CATE estimation error. Graph support comparisons and perturbation diagnostics further support using plug-in CATE similarity to select pairs for regularization.

1 INTRODUCTION

Estimating how treatment effects vary with observed covariates is central to personalized decision-making. The conditional average treatment effect (CATE) is the expected difference between potential outcomes under treatment and control conditional on covariates (Shalit et al., 2017; Curth & Van der Schaar, 2021). Because each unit reveals only one potential outcome, individual treatment effects are unavailable as direct supervision (Rubin, 1974). Two-stage CATE learners address this problem by first estimating nuisance functions and then fitting a target CATE model using imputed effects (Künzel et al., 2019), pseudo-outcomes (Curth & Van der Schaar, 2021; Kennedy, 2023; Fisher, 2024), or residualized losses (Nie & Wager, 2021). However, the resulting second-stage CATE targets or losses depend on estimated nuisance functions and can be noisy.

Existing work has explored regularization and information sharing to improve CATE estimation. Some methods adapt regularization in second-stage learning or share information across treatment-specific models (Melnychuk et al., 2026; Chauhan et al., 2024). Pairwise, local, and graph structures have also been used for different purposes, including treatment-effect fusion, pairwise training, and representation learning (Padilla et al., 2026; Nagalapatti et al., 2024; Yao et al., 2018; Cohen et al.; Zaccagnino et al., 2026; Cao et al., 2026). For second-stage CATE regularization, these ideas raise a more specific question: which pairs of observations should be regularized together when only noisy first-stage CATE estimates are available? Ideally, such pairs would have similar true CATEs, but the true CATEs are not directly observed.

Choosing these pairs is therefore nontrivial. Regularizing two observations with substantially different true CATEs can shrink genuine treatment effect heterogeneity and introduce bias, whereas failing to regularize two observations with similar true CATEs can forgo a potential reduction in variance (Chen & Liu, 2017). Two-stage learning nevertheless provides a natural source of information for making this choice. Treatment-specific outcome regressions can be fitted before the target CATE model, and their difference provides a plug-in CATE estimate for each training observation. Although these estimates are noisy, they are available before second-stage optimization and can

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.

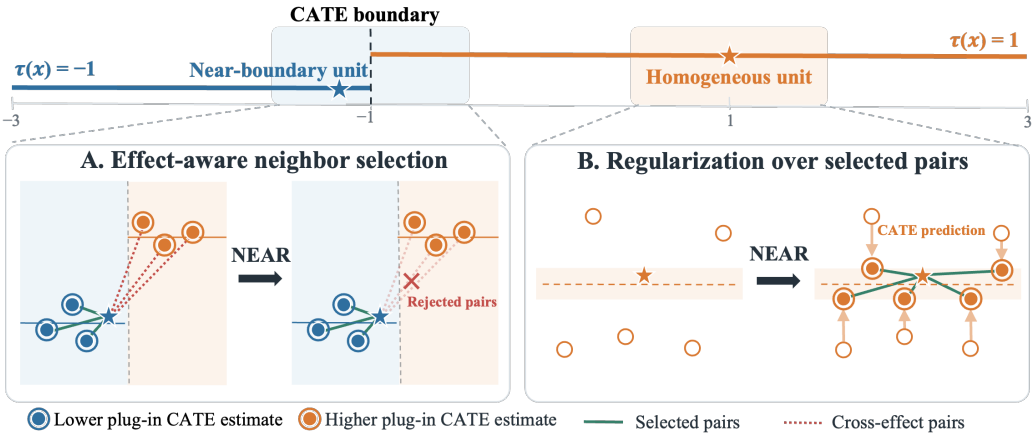


Figure 1: A schematic of NEAR. The upper curve illustrates a CATE function with a sharp change. (A) Pair selection excludes units with large plug-in CATE differences. (B) Pairwise regularization encourages similar second-stage predictions over selected pairs.

therefore be used to identify observations with similar estimated CATEs for regularization. This leads to the following question:

> *Can noisy first-stage plug-in CATE estimates be used to select pairs for second-stage regularization, and how do errors in those estimates affect CATE risk?*

To address this question, we propose ***N**eighborhood **E**ffect-**A**ware **R**egularization* (NEAR), a modular second-stage regularizer. NEAR fits treatment-specific outcome regressions, uses their difference to form plug-in CATE estimates, constructs a fixed graph connecting observations with similar plug-in CATEs, and regularizes their second-stage CATE predictions over the selected pairs. Figure 1 illustrates this procedure. We show how plug-in CATE estimation errors control true CATE variation along selected pairs. For squared-loss second-stage regression with a stabilized doubly robust (DR) pseudo-outcome, we derive finite-sample CATE risk bounds that quantify how this variation affects estimation risk. Experiments across six two-stage learners and three benchmarks show lower mean CATE estimation error with NEAR and support using plug-in CATE similarity to select pairs for regularization.

Contributions. Our contributions are threefold:

1. We introduce a modular second-stage regularizer that forms plug-in CATE estimates from treatment-specific outcome regressions, uses them to construct a fixed graph, and regularizes CATE predictions over selected pairs with similar plug-in CATEs.
2. We establish how plug-in CATE estimation errors control true CATE variation along selected pairs. For squared-loss second-stage regression with a stabilized DR pseudo-outcome over a bounded function class, we derive finite-sample CATE risk bounds that characterize how this variation affects second-stage estimation. A complementary quadratic analysis at the training covariates characterizes when variance reduction offsets the bias introduced by smoothing.
3. Across six two-stage learners and three benchmarks, NEAR lowers mean outcome-standardized root PEHE in all 18 combinations of learner and benchmark. Additional experiments compare alternative pair selections and show increasing CATE error under controlled perturbations that increase true CATE variation along selected edges.

2 Related Work

Two-stage learning and modular regularization. Two-stage CATE learners construct second-stage targets or objectives using imputed effects (Künzel et al., 2019), residualized losses (Nie & Wa-

ger, 2021), or influence-function-based pseudo-outcomes (Kennedy, 2023). Orthogonal statistical learning characterizes loss constructions that reduce the sensitivity of target estimation to nuisance estimation error (Foster & Syrgkanis, 2023). Recent work has also modified regularization and information sharing within CATE learners. Overlap-Adaptive Regularization adapts second-stage regularization strength to treatment overlap (Melnychuk et al., 2026), while HyperITE shares information across treatment-specific models through hypernetwork-based soft weight sharing (Chauhan et al., 2024). These approaches modify second-stage targets, regularization strength, or information sharing across treatment models. They do not address which observations should be regularized together in second-stage CATE estimation.

Pairwise and local structure in CATE learning. Pairwise and local information has been used for several different purposes in treatment-effect estimation. PairNet constructs training losses from pairs of factual outcomes (Nagalapatti et al., 2024), while Rank-Learner uses pairwise comparisons to learn treatment-effect orderings (Arno et al., 2026). Representation-learning methods use local pairings in different ways: SITE preserves local similarity (Yao et al., 2018), PITE aligns local prototypes across treatment groups (Cao et al., 2026), and HERMES dynamically forms pairs from estimated treatment-effect similarity for Siamese representation learning (Zaccagnino et al., 2026). NEAR differs in how these pairs are constructed and where they enter estimation. It forms plug-in CATE estimates from treatment-specific outcome regressions, constructs the graph before second-stage optimization, and then holds the graph fixed while regularizing second-stage CATE predictions.

Graph and fusion regularization. Graph Laplacian regularization encourages smooth predictions over connected observations (Belkin et al., 2006) and exhibits a bias–variance tradeoff that depends on the graph and the values being smoothed (Chen & Liu, 2017). For heterogeneous treatment effects, causal fused lasso orders observations using propensity or prognostic scores and estimates piecewise-constant treatment effects along that ordering (Padilla et al., 2026). LITE uses Laplacian regularization within representation learning for treatment-effect estimation (Cohen et al.). NEAR instead constructs the graph from plug-in CATE similarity and applies the graph penalty to second-stage CATE predictions. Our analysis links errors in the plug-in CATE estimates to true CATE variation along selected pairs and then to CATE risk.

3 Problem Setup

Causal framework and estimand. We work in the potential outcomes framework (Rubin, 1974). Let $\mathcal{D}_n = \{(X_i, A_i, Y_i)\}_{i=1}^n$ denote the observed data from n i.i.d. units, where $X_i \in \mathcal{X}$ is the covariate vector, $A_i \in \{0, 1\}$ is a binary treatment, and $Y_i \in \mathbb{R}$ is the observed outcome. Let $Y_i(a)$ denote the potential outcome under treatment $a \in \{0, 1\}$. Our estimand is the conditional average treatment effect (CATE),

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

Let $e(x) = \Pr(A = 1 \mid X = x)$ denote the propensity score. We use the standard identification conditions (Rubin, 1974; Rosenbaum & Rubin, 1983). Consistency requires $Y_i = Y_i(A_i)$. Conditional unconfoundedness requires $(Y_i(0), Y_i(1)) \perp A_i \mid X_i$. Overlap requires an $\varepsilon > 0$ such that $\varepsilon \le e(x) \le 1-\varepsilon$ for every x in the covariate support. Under these conditions, the CATE is identified by $\tau(x) = \mu_1(x) - \mu_0(x)$, where $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$.

Base second-stage problem. Let $\widehat{\eta}$ collect estimated nuisance functions such as $\widehat{e}, \widehat{\mu}_0, \widehat{\mu}_1$. We consider two-stage learners whose target model has parameters $\theta \in \Theta$ and is fitted by minimizing an empirical objective that decomposes over observations:

$$\widehat{\mathcal{L}}(\theta; \widehat{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta; \widehat{\eta}). \tag{2}$$

Write $f_\theta : \mathcal{X} \to \mathbb{R}$ for the learner's final scalar CATE prediction, regardless of whether θ parameterizes a single effect model or multiple internal branches, as in the X-learner. This output-level notation allows the same pairwise penalty to be added while leaving the learner-specific terms ℓ_i unchanged. The concrete objectives are listed in Appendix B.1. Although the model shares parameters across observations, this objective contains no explicit penalty on differences between selected pairs of CATE predictions.

Second-stage problem with pairwise regularization. We add pairwise regularization to this output interface while retaining the learner-specific loss. Let $\mathcal{P}_n = \{(i,j) : 1 \le i < j \le n\}$ be the unordered pairs in the training sample, and let $w = (w_{ij})_{(i,j)\in\mathcal{P}_n}$ denote nonnegative, data-dependent pair weights. When the support is nonempty, we normalize the weights so that $\sum_{(i,j)\in\mathcal{P}_n} w_{ij} = 1$. If no pair is retained, set all weights to zero. Write

$$\Omega_w(\theta) = \sum_{(i,j)\in\mathcal{P}_n} w_{ij}\{f_\theta(X_i) - f_\theta(X_j)\}^2.$$

For $\lambda \ge 0$, the resulting optimization problem is

$$\widehat{\theta}_{\lambda,w} \in \arg\min_{\theta\in\Theta}\{\widehat{\mathcal{L}}(\theta;\widehat{\eta}) + \lambda\Omega_w(\theta)\}. \tag{3}$$

NEAR uses first-stage plug-in CATE estimates to construct these weights before second-stage fitting. Their nonzero entries select the coupled predictions, and their magnitudes determine each pair's relative contribution.

4 METHOD

4.1 PLUG-IN CATE SIMILARITY

NEAR constructs its graph before target-model fitting and holds it fixed during second-stage optimization. First-stage outcome regressions give

$$\widetilde{\tau}_i = \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i), \qquad k_{ij} = \exp\left\{-\frac{(\widetilde{\tau}_i - \widetilde{\tau}_j)^2}{2\sigma_\tau^2}\right\}, \quad i \ne j. \tag{4}$$

The scale $\sigma_\tau > 0$ is defined from the scaled median absolute deviation of the training plug-in CATE values, with the fallbacks in Equation (57) when that quantity is zero. Larger k_{ij} means greater estimated effect similarity. Neighborhoods are defined by plug-in CATE values, so their members need not be close in covariate space. A dense graph would contain $O(n^2)$ edges and couple units with dissimilar estimated effects. We instead select a sparse edge set from these scores.

4.2 EFFECT-AWARE NEIGHBORHOOD CONSTRUCTION

We first limit candidate pairs by their standardized effect discrepancy,

$$d_{ij} = \frac{|\widetilde{\tau}_i - \widetilde{\tau}_j|}{\sigma_\tau}, \qquad \mathcal{C}_\rho(i) = \{j \ne i : d_{ij} \le \rho\}, \quad \rho \in [0,\infty]. \tag{5}$$

The choice $\rho = \infty$ disables filtering. Each node then selects the indices of at most $K \in \{1,\dots,n-1\}$ eligible neighbors with largest scores, using deterministic tie-breaking,

$$\mathcal{N}_{K,\rho}(i) = \mathrm{TopK}_{j\in\mathcal{C}_\rho(i)}\, k_{ij}. \tag{6}$$

Mutual selection retains only pairs chosen by both endpoints,

$$\mathcal{E}_{K,\rho} = \{\{i,j\} : i < j,\ j \in \mathcal{N}_{K,\rho}(i),\ i \in \mathcal{N}_{K,\rho}(j)\}. \tag{7}$$

Thus the graph is undirected, each node has degree at most K, and $|\mathcal{E}_{K,\rho}| \le nK/2$. Normalize scores over the nonempty retained set,

$$w_{ij} = \frac{k_{ij}}{\sum_{\{a,b\}\in\mathcal{E}_{K,\rho}} k_{ab}}, \qquad \{i,j\} \in \mathcal{E}_{K,\rho}. \tag{8}$$

Nonedges have weight zero. If the graph is empty, all weights are zero.

4.3 NEAR SECOND STAGE

For $\boldsymbol{f}_\theta = (f_\theta(X_i))_{i=1}^n$, the selected graph gives

$$\Omega_{K,\rho}(\boldsymbol{f}_\theta) = \sum_{\{i,j\}\in\mathcal{E}_{K,\rho}} w_{ij}\{f_\theta(X_i) - f_\theta(X_j)\}^2. \tag{9}$$

Using this penalty in Equation (3) defines $\widehat{\theta}_{\lambda,K,\rho}$. The radius ρ limits candidate discrepancies, K caps local degree, and λ controls the penalty strength. Graph construction and training are given in Algorithms 1 and 2 in the appendix.

5 PLUG-IN PAIR SELECTION AND CATE RISK

We first relate plug-in CATE estimation errors to true CATE differences across selected edges, then analyze out-of-sample CATE risk. The function-risk result uses squared loss with the stabilized DR pseudo-outcome. A fixed-node quadratic model separately makes the bias–variance trade-off explicit. Full statements and proofs appear in Appendix A.

5.1 FROM PLUG-IN ERROR TO SELECTED-EDGE CATE VARIATION

Write $\tau_i = \tau(X_i)$, $\delta_i = \widetilde{\tau}_i - \tau_i$, and extend Equation (9) to node vectors by

$$\Omega_{K,\rho}(\boldsymbol{a}) = \sum_{\{i,j\}\in\mathcal{E}_{K,\rho}} w_{ij}(a_i - a_j)^2. \tag{10}$$

This is the weighted graph Dirichlet energy (Belkin et al., 2006). Using $w_{ji} = w_{ij}$, let $d_i^w = \sum_{j\neq i} w_{ij}$, $d_{\max}^w = \max_i d_i^w$, and for any node vector $\boldsymbol{a}$ define $V_n(\boldsymbol{a}) = \min_{c\in\mathbb{R}} \frac{1}{n}\sum_i (a_i - c)^2$.

Lemma 1 (Plug-in error and selected-edge CATE variation)**.** *For the graph in Section 4, every pair satisfies*

$$\big||\widetilde{\tau}_i - \widetilde{\tau}_j| - |\tau_i - \tau_j|\big| \le |\delta_i - \delta_j|. \tag{11}$$

For finite ρ,

$$\sqrt{\Omega_{K,\rho}(\boldsymbol{\tau})} \le \sqrt{\Omega_{K,\rho}(\widetilde{\boldsymbol{\tau}})} + \sqrt{\Omega_{K,\rho}(\boldsymbol{\delta})} \le \rho\sigma_\tau + \sqrt{\Omega_{K,\rho}(\boldsymbol{\delta})}, \tag{12}$$

$$\Omega_{K,\rho}(\boldsymbol{\delta}) \le 2nd_{\max}^w V_n(\boldsymbol{\delta}). \tag{13}$$

If $m = |\mathcal{E}_{K,\rho}| > 0$, then $d_{\max}^w \le K\exp(\rho^2/2)/m$.

The proof is given in Appendix A.2. The radius limits plug-in CATE differences on selected edges, while $|\delta_i - \delta_j|$ controls the discrepancy between plug-in and true CATE differences. A common shift in all plug-in CATE errors cancels from these pairwise differences. The weighted-degree factor captures concentration of the normalized edge weights beyond the degree cap K. The bounds apply to the graph constructed from the observed sample.

5.2 OUT-OF-SAMPLE CATE RISK

Let $\mathcal{R}(f) = \mathbb{E}_{X'}[\{f(X') - \tau(X')\}^2]$, where X' is an independent covariate from the training population. Write $\Omega(f)$ for $\Omega_{K,\rho}$ evaluated at $(f(X_i))_{i=1}^n$. For the fitted stabilized DR pseudo-outcomes $\widehat{Z}_i$, let Z_i^0 be the same score evaluated at the true nuisance functions and define $\zeta_n^2 = \frac{1}{n}\sum_i(\widehat{Z}_i - Z_i^0)^2$. The fitted outcome-regression predictions are finite on the training sample, and fitted propensities lie in $[0,1]$. The fitted predictor approximately minimizes $\widehat{J}(f) = \frac{1}{n}\sum_i(f(X_i) - \widehat{Z}_i)^2 + \lambda\Omega(f)$ over a fixed function class $\mathcal{F}$, up to error $\epsilon_{\text{opt}} \ge 0$.

Assumption 1 (Sampling and function class)**.** *Observations are i.i.d., $|\tau(X)| \le T$, and $Y - \mu_A(X)$ is conditionally mean-zero and ν-sub-Gaussian given (X, A). The fixed class $\mathcal{F}$ consists of measurable functions, is compact in the uniform norm, $\sup_{f\in\mathcal{F}}\|f\|_\infty \le B$, and admits deterministic uniform t-covers with centers in $\mathcal{F}$ and log cardinality at most $H_{\mathcal{F}}(t)$. The stabilized DR score uses a fixed augmentation threshold $q \in (0, 1/2)$, as defined in Appendix A.5.1.*

The reference noise is then conditionally sub-Gaussian with parameter $\sigma = \nu/q$. For $t > 0$, $\alpha \in (0,1)$, put $v = H_{\mathcal{F}}(t) + \log(8/\alpha)$, $u_0 = \log(4/\alpha)$, and

$$\Gamma_n = 4(B+T)t + (B+T)^2\sqrt{\frac{v}{2n}}, \qquad \Pi_n = \sigma\left[B\sqrt{\frac{2v}{n}} + t\sqrt{1 + 2\sqrt{\frac{u_0}{n}} + \frac{2u_0}{n}}\right]. \tag{14}$$

Theorem 1 (CATE function risk)**.** *Under Assumption 1 and the identification conditions, with probability at least $1 - 3\alpha/4$, simultaneously for every fitted nuisance function whose outcome-regression predictions are finite and whose propensity predictions lie in $[0,1]$, every nonnegative graph, $\lambda \ge 0$, and $\widehat{f} \in \mathcal{F}$ with objective gap at most ϵ_{opt},*

$$\mathcal{R}(\widehat{f}) + \lambda\Omega(\widehat{f}) \le \inf_{g\in\mathcal{F}}\{\mathcal{R}(g) + \lambda\Omega(g)\} + 2\Gamma_n + 4\Pi_n + 4B\zeta_n + \epsilon_{\text{opt}}. \tag{15}$$

The proof is given in Appendix A.7. The infimum retains both function approximation error and the cost of smoothing a comparator. The remaining terms separate sampling error, fitted-score error, and approximate optimization. The concentration events are uniform over the fixed function class, while the fitted-score discrepancy is isolated in ζ_n. The bound therefore applies to a graph and nuisance estimates fitted on the same training sample. For the Exponential Linear Unit (ELU) class on a bounded covariate domain, parameters bounded by M give $H_{\mathcal{F}}(t) \leq p \log(1 + 2MC/t)$, with p and C defined in Equation (38). Taking $t = (B+T)/n$ gives sampling terms of order $\sqrt{\log n/n}$ for a fixed class and confidence level. Corollary 2 gives convergence to the best population CATE risk in the fixed function class when the fitted-score error, optimization error, and comparator regularization cost vanish.

Lemma 1 also controls the graph cost in Equation (15). With $Q_n(g) = \frac{1}{n}\sum_i \{g(X_i) - \tau_i\}^2$,

$$\sqrt{\Omega(g)} \leq \sqrt{\Omega(\widetilde{\boldsymbol{\tau}})} + \sqrt{\Omega(\boldsymbol{\delta})} + \sqrt{2nd^w_{\max} Q_n(g)}. \tag{16}$$

Thus selected edges with large true CATE differences can increase regularization bias even when the second-stage function class approximates the CATE well.

Bias–variance interpretation and pointwise use of plug-in CATE estimates. For an unrestricted quadratic fit at fixed training nodes with conditionally unbiased targets, the exact risk decomposition implies improvement whenever the variance reduction exceeds $\lambda\Omega_{K,\rho}(\boldsymbol{\tau})/4$ (Appendix A.3). Shifting all plug-in CATE values by a common constant leaves the graph unchanged, and $\Omega(f + c) = \Omega(f)$. The graph penalty therefore does not anchor predictions to the plug-in estimates' absolute level (Appendix A.4). The experiments examine selected-edge CATE variation, the bias–variance decomposition, and pointwise use of the plug-in CATE alongside final CATE error.

6 Experiments

We evaluate whether NEAR improves CATE estimation across two-stage learners, whether pair selection adds value beyond pointwise use of plug-in CATE estimates, and how plug-in CATE errors affect selected-edge CATE variation and final error.

6.1 Experimental Setup

Learners and benchmarks. We evaluate the doubly robust (DR), R-, inverse-variance-weighted (IVW), regression-adjustment (RA), propensity-weighting (PW), and X-learners on the Infant Health and Development Program (IHDP) (Hill, 2011), the Atlantic Causal Inference Conference (ACIC) 2016 benchmark (Dorie et al., 2019), and Synthetic. The semi-synthetic benchmarks retain observed covariates and supply simulated potential outcomes. Synthetic uses independent covariates $X_{ij} \sim \mathrm{Unif}(-1, 1)$, $j = 1, \ldots, 10$, treatment $A_i \mid X_i \sim \mathrm{Bernoulli}\{e_\gamma(X_i)\}$, and $Y_i(a) = X_{i1} + X_{i2}^2 + \sin(\pi X_{i3}) + 0.5X_{i4}X_{i5} + a\tau_{\mathrm{sharp}}(X_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$.

We set $e_\gamma(x) = \mathrm{sigmoid}\{\gamma(0.8x_1 - 0.6x_2 + 0.4x_3)\}$, clipped to $[0.02, 0.98]$, with $\gamma = 3.0$, $n = 1000$, and

$$\tau_{\mathrm{sharp}}(x) = \begin{cases} -1, & x_1 < 0, \\ 2, & x_1 \geq 0. \end{cases}$$

Evaluation and comparators. We use ten Synthetic repetitions, 100 IHDP splits, and the 77 ACIC simulation settings. Each ACIC setting is one reported evaluation unit, so its mean and standard deviation are computed across these 77 settings. For IHDP and ACIC, NEAR selects from the 48 combinations of $\lambda \in \{0.1, 1, 10, 100\}$, $K \in \{1, 2, 4, 8\}$, and $\rho \in \{2.146, 1.794, 1.354\}$. One configuration per learner and dataset is selected by mean learner-specific validation loss on prespecified calibration units and then fixed for all reported outer evaluations. Each model is refitted on its outer training set. Synthetic uses a $640/160/200$ training-validation-test split in each repetition and refits on the first 800 observations after validation. Each learner is compared without NEAR (None) and with NEAR under the same training settings. Within each evaluation unit, these comparisons share nuisance predictions, target initialization, architecture, and optimization settings. Further comparisons use DR as a common estimator. Constant and debiased OAR (dOAR) (Melnychuk et al., 2026) compare regularization strategies. Direct plug-in prediction and plug-in distillation evaluate

pointwise use of the first-stage estimates. Direct uses $\widetilde{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x)$ as the CATE prediction, and distillation penalizes pointwise deviations from the training values $\widetilde{\tau}_i$. Full objectives, calibration details, and implementation settings are in Appendix B.

Metric. We report outcome-standardized root PEHE (Hill, 2011; Shalit et al., 2017), $\sqrt{\epsilon_{\text{PEHE}}} = \frac{1}{s_Y}\sqrt{\frac{1}{n_{\text{te}}}\sum_{i=1}^{n_{\text{te}}}\{\widehat{\tau}(X_i) - \tau(X_i)\}^2}$, where n_{te} is the test-set size, $s_Y > 0$ is the outcome standard deviation estimated on training data, and the CATE values inside the sum are on the original outcome scale. Lower values indicate more accurate CATE estimation. Tables report mean $\pm$ standard deviation unless stated otherwise. Unless stated otherwise, every root PEHE reported below uses this outcome standardization. Boldface marks the lowest displayed mean in each comparison, including ties.

Table 1: Root PEHE over 10 Synthetic repetitions, 100 IHDP splits, and 77 ACIC simulation settings. Gain is $100(\text{None} - \text{NEAR})/\text{None}$ in percent, computed from unrounded means.

Learner	Synthetic			IHDP			ACIC		
	None	NEAR	Gain ↑	None	NEAR	Gain ↑	None	NEAR	Gain ↑
DR	0.93 ± 0.14	$\mathbf{0.47 \pm 0.07}$	49.54	0.69 ± 0.34	$\mathbf{0.49 \pm 0.22}$	28.75	0.60 ± 0.20	$\mathbf{0.47 \pm 0.22}$	21.76
R	0.99 ± 0.15	$\mathbf{0.46 \pm 0.09}$	53.46	1.04 ± 0.66	$\mathbf{0.50 \pm 0.22}$	51.77	1.42 ± 1.07	$\mathbf{0.48 \pm 0.22}$	66.44
IVW	2.00 ± 0.22	$\mathbf{0.77 \pm 0.33}$	61.34	0.96 ± 0.52	$\mathbf{0.72 \pm 0.18}$	25.12	0.86 ± 0.34	$\mathbf{0.50 \pm 0.22}$	42.11
RA	0.56 ± 0.06	$\mathbf{0.48 \pm 0.06}$	13.75	0.60 ± 0.27	$\mathbf{0.54 \pm 0.23}$	9.44	0.50 ± 0.21	$\mathbf{0.47 \pm 0.22}$	6.92
PW	2.03 ± 0.41	$\mathbf{0.56 \pm 0.37}$	72.19	2.07 ± 0.93	$\mathbf{0.64 \pm 0.23}$	69.20	1.42 ± 0.21	$\mathbf{0.62 \pm 0.23}$	56.34
X	0.55 ± 0.08	$\mathbf{0.43 \pm 0.07}$	21.95	0.68 ± 0.28	$\mathbf{0.53 \pm 0.21}$	22.02	0.60 ± 0.41	$\mathbf{0.49 \pm 0.23}$	17.89

6.2 Results Analysis

NEAR lowers mean root PEHE in all 18 learner and benchmark combinations (Table 1). The reductions occur across pseudo-outcome, residualized-loss, weighting, and multi-branch learners.

Table 2: Regularization strategies with DR as the common estimator. Constant is the fixed, nonadaptive regularization baseline from OAR.

Regularizer	IHDP	ACIC
None	0.69 ± 0.34	0.60 ± 0.20
Constant	0.65 ± 0.33	0.58 ± 0.21
dOAR	0.61 ± 0.30	0.55 ± 0.21
NEAR	$\mathbf{0.49 \pm 0.22}$	$\mathbf{0.47 \pm 0.22}$

With DR held fixed, Table 2 compares regularization strategies. Constant and dOAR have lower mean error than None, while NEAR has the lowest mean error on both benchmarks. The dOAR comparator combines overlap-adaptive regularization with its debiasing correction. NEAR selects pairs of CATE predictions to couple using plug-in CATE similarity.

First-stage estimates already provide substantial predictive information. Direct plug-in prediction and pointwise distillation both have lower mean error than None (Table 3), and NEAR has the lowest mean error on all three benchmarks. The paired comparison in Table 7 gives lower error for NEAR than Direct on 70 of 100 IHDP splits and 65 of 77 ACIC settings, and lower error than Distill on 66 of 100 and 53 of 77, respectively.

Table 3: First-stage plug-in CATE use with DR as the common estimator. Direct predicts the plug-in estimate, Distill penalizes pointwise deviation from it, and NEAR regularizes selected pairs.

Dataset	None	Direct	Distill	NEAR
Synthetic	0.93 ± 0.14	0.52 ± 0.05	0.56 ± 0.12	$\mathbf{0.47 \pm 0.07}$
IHDP	0.69 ± 0.34	0.53 ± 0.27	0.54 ± 0.27	$\mathbf{0.49 \pm 0.22}$
ACIC	0.60 ± 0.20	0.50 ± 0.22	0.49 ± 0.21	$\mathbf{0.47 \pm 0.22}$

Which graph supports help? Table 4 compares graph supports with uniform weights and separately validated λ. We denote the uniform-weight NEAR variant by NEAR-U. It improves over None on both benchmarks, while the covariate, random, and shuffled plug-in graphs in this comparison increase error. Covariate edges follow feature proximity, and shuffling breaks the link between estimated effects and units. Among these graph constructions, selecting edges by plug-in CATE similarity gives the lowest mean error. Gaussian and uniform weights on the same NEAR graph have the same displayed mean on IHDP, while Gaussian weighting is lower on ACIC, 0.47 versus 0.49 (Table 10). At $K = 16$, root PEHE stays at 0.51 as ρ decreases from 0.30 to 0.01, even as edge retention falls from 0.99 to 0.27. At $\rho = 0.0001$, fewer than one percent of edges remain, the isolated-node ratio reaches 0.96, and root PEHE rises to 0.64 (Table 9).

Table 4: Graph support comparison with DR. All graphs use uniform weights and separately validated λ. Random matches the edge count of NEAR-U.

Dataset	None	Covariate	Random	Shuffled	NEAR-U
IHDP	0.69 ± 0.34	0.74 ± 0.13	0.84 ± 0.14	0.85 ± 0.14	$\mathbf{0.51 \pm 0.26}$
ACIC	0.60 ± 0.20	0.62 ± 0.23	0.66 ± 0.28	0.66 ± 0.29	$\mathbf{0.49 \pm 0.21}$

Selected-edge CATE variation and risk. For a nonempty retained graph $G = (\{1, \ldots, n\}, \mathcal{E}_{K,\rho})$, we use the normalized selected-edge CATE variation $\widetilde{I}_G(\tau) = \Omega_{K,\rho}(\boldsymbol{\tau})/\{2V_n(\boldsymbol{\tau})\}$, where $V_n(\boldsymbol{\tau})$ is the empirical variance of the true CATE across training nodes, as defined in Section 5. The metric is defined when this variance is positive. Twice this variance is the mean squared CATE difference for two independent draws from the empirical node distribution. Thus $\widetilde{I}_G(\tau) = 1$ is the empirical independent-pair reference, while values below one indicate smaller true CATE differences along the selected edges. Perturbing only the plug-in CATE used for graph construction raises $\Omega_{K,\rho}(\boldsymbol{\delta})$ from 1.02 to 3.00, the normalized selected-edge CATE variation from 0.23 to 0.67, and root PEHE from 0.47 to 0.64, while the retained edge count and isolated-node ratio remain nearly unchanged (Table 12 in the appendix). Outcomes, treatments, and the base loss remain fixed throughout this perturbation. Because only the plug-in CATE values used to construct the graph are perturbed, the experiment isolates how degrading this first-stage information changes selected-edge CATE variation and final CATE error.

Table 5: Synthetic selected-edge diagnostic over ten repetitions. Oracle uses the true CATE to form a graph with the same active nodes and edge count as NEAR.

Graph	$\sqrt{\epsilon_{\text{PEHE}}}$	$\widetilde{I}_G(\tau)$
Oracle	$\mathbf{0.27 \pm 0.15}$	$\mathbf{0.00 \pm 0.00}$
NEAR	0.47 ± 0.07	0.21 ± 0.03

Table 6: Synthetic CATE bias–variance decomposition over 50 fits. Values are in original outcome units, and MSE is CATE prediction mean squared error.

Method	Bias^2	Variance	MSE
None	**0.40**	3.83	4.23
NEAR	0.43	**0.46**	**0.89**

The Oracle graph reduces normalized selected-edge CATE variation from 0.21 to 0.00 and mean root PEHE from 0.47 to 0.27 (Table 5). It provides a reference for the gain available from more accurate CATE information in pair selection. Table 6 shows squared CATE bias increasing from 0.40 to 0.43, while prediction variance falls from 3.83 to 0.46 and MSE from 4.23 to 0.89. We fit each method on 50 independently generated Synthetic training samples and evaluate every fit on the same 2,048 independently drawn covariates. Definitions and the standard decomposition (Geman et al., 1992) are in Appendix C.7.

6.3 Robustness and Sensitivity

NEAR also lowers mean root PEHE for smooth linear, grouped piecewise-constant, and nonlinear interacting effects (1.15 to 0.37, 1.35 to 0.69, and 1.49 to 0.64, Table 11). Their CATE functions are $1 + x_1 + 0.5x_2$, two groups defined by the sign of x_1x_2, and $\sin(3\pi x_1) + x_2x_3$, respectively. These cases test sharing along continuously varying effects, within separate effect groups, and across nonlinear interactions. Table 8 further compares full-sample nuisance fitting with two- and five-fold

cross-fitting. For $L = 1, 2, 5$ nuisance folds, None has mean root PEHE $0.93, 1.49, 1.33$, while NEAR has $0.47, 0.62, 0.56$. Cross-fitting raises mean error relative to $L = 1$ for both methods in this diagnostic, while NEAR remains lower at each fold count. The ordering is therefore unchanged when nuisance predictions are cross-fitted. The full sample-size and noise sweep extends Figure 2(a) to $n \in \{500, 1000, 2000, 5000\}$ and $\sigma_\varepsilon \in \{0.5, 1, 2\}$. It fixes $(\lambda, K, \rho) = (10, 2, 1.354)$ over all 12 matched conditions, and NEAR has lower mean root PEHE in every condition (Figure 3).

Figure 2(b) shows U-shaped error curves whose minima shift toward smaller λ as K grows. Total edge weight stays at one, so this shift reflects changes in the selected edges and their relative weights. Figure 2(c) shows that moderate radius filtering maintains low error, while very small radii isolate units and increase error. Together, the panels show how K and ρ determine the selected edges and λ controls smoothing over them. The current implementation evaluates all candidate pairs before sparsification, giving $O(n^2)$ pair evaluations in the worst case. Because the graph depends on scalar plug-in CATE differences, sorting those values can reduce graph-construction cost. Table 13 reports graph construction times from 0.03 to 3.09 seconds and total training times from 1.13 to 1.48 times None as n increases from 1,000 to 10,000.

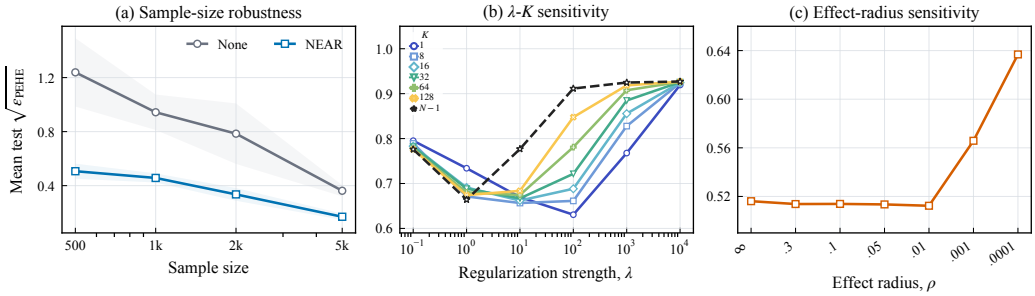


Figure 2: DR diagnostics. Panel (a) shows Synthetic root PEHE versus sample size, with mean and standard deviation over ten repetitions. Panel (b) shows mean IHDP root PEHE versus λ for different K, at $\rho = 1.794$, over ten splits. The legend label $N-1$ denotes the cap $K = n-1 = 671$. Panel (c) shows mean IHDP root PEHE versus ρ, at $\lambda = 10, K = 16$, over ten separate splits. At $\rho = \infty$, the radius filter is disabled.

7 Conclusion

We develop NEAR, a modular regularizer that uses first-stage plug-in CATE estimates to select pairs for second-stage CATE regularization without changing the base learner's objective. We link plug-in CATE error to true CATE variation on selected edges and show how this variation enters finite-sample CATE risk, with a complementary fixed-node bias–variance analysis. The analysis connects first-stage CATE accuracy, true CATE variation on selected edges, and the smoothing bias in the second stage. Across six learners and three benchmarks, NEAR lowers mean CATE error in all 18 combinations of learner and benchmark. Graph support, Oracle, and perturbation diagnostics further show lower CATE error when selected pairs have smaller true CATE differences.

Limitations and future work. The finite-sample function-risk analysis focuses on squared-loss second-stage regression with a stabilized DR pseudo-outcome over bounded function classes. Extending these guarantees to the other second-stage objectives evaluated here is an interesting next step. The current graph is fixed before target-model fitting, and analyzing graph updates driven by the evolving target model would require additional control of data-dependent pair selection. Approximate neighbor construction that avoids all-pairs scoring while preserving comparable control of selected-edge CATE variation could improve scalability.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing and language polishing. Authors are responsible for checking all claims, proofs, analyses, and reported results.

REFERENCES

Henri Arno, Dennis Frauen, Emil Javurek, Thomas Demeester, and Stefan Feuerriegel. Ranklearner: Orthogonal ranking of treatment effects. *arXiv preprint arXiv:2602.03517*, 2026.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7 (11), 2006.

Fuyuan Cao, Jiaxuan Zhang, and Xiaoli Li. Pite: Multi-prototype alignment for individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19871–19879, 2026.

Vinod Kumar Chauhan, Jiandong Zhou, Ghadeer Ghosheh, Soheila Molaei, and David A Clifton. Dynamic inter-treatment information sharing for individualized treatment effects estimation. In *International conference on artificial intelligence and statistics*, pp. 3529–3537. PMLR, 2024.

Pin-Yu Chen and Sijia Liu. Bias-variance tradeoff of graph laplacian regularizer. *IEEE Signal Processing Letters*, 24(8):1118–1122, 2017.

David Cohen, Ronen Talmon, and Uri Shaham. Revisiting covariate and hypothesis roles in ite estimation: A new approach using laplacian regularization.

Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International conference on artificial intelligence and statistics*, pp. 1810–1818. PMLR, 2021.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019.

Aaron Fisher. Inverse-variance weighting for estimation of heterogeneous treatment effects. In *Forty-first International Conference on Machine Learning*, 2024.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51 (3):879–908, 2023.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

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

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

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.

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Overlap-adaptive regularization for conditional average treatment effect estimation. In *International Conference on Learning Representations*, volume 2026, pp. 53748–53788, 2026.

Lokesh Nagalapatti, Pranava Singhal, Avishek Ghosh, and Sunita Sarawagi. Pairnet: Training with observed pairs to estimate individual treatment effect. *arXiv preprint arXiv:2406.03864*, 2024.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Oscar Hernan Madrid Padilla, Yanzhen Chen, Carlos Misael Madrid Padilla, and Gabriel Ruiz. A causal fused lasso for interpretable heterogeneous treatment effects estimation. *Journal of Machine Learning Research*, 27(40):1–56, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Rocco Zaccagnino, Gerardo Benevento, Delfina Malandrino, Donatello Telesca, and Alessia Ture. Hermes: Heterogeneous effects representation with matched embeddings using siamese networks. *Transactions on Machine Learning Research*, 2026.

A Proofs and Risk Analysis

We first establish selected-edge bounds and conditional node results. The function-risk proofs then control fitted-score error, with a stable lower subgraph handling the fitted graph's dependence on the training data.

A.1 Graph and conditional node model

Fix covariates $X_1, \dots, X_n$ and write $\tau_i = \tau(X_i)$ for their true conditional average treatment effects. Let $\widetilde{\tau}_i$ denote the plug-in CATE used to construct the graph, and define $\delta_i = \widetilde{\tau}_i - \tau_i$.

Assumption 2 (Finite node values)**.** *$n \geq 2$ is finite and $\tau_i, \widetilde{\tau}_i \in \mathbb{R}$ for every i.*

Assumption 3 (Graph definition)**.** *Let $K \in \{1, \dots, n-1\}$, $\sigma_\tau > 0$, and $\rho \in [0, \infty]$. For $i \neq j$, define*

$$k_{ij} = \exp\left(-\frac{(\widetilde{\tau}_i - \widetilde{\tau}_j)^2}{2\sigma_\tau^2}\right).$$

If $\rho < \infty$, the eligible candidates for node i are those $j \neq i$ with $|\widetilde{\tau}_i - \widetilde{\tau}_j| \leq \rho\sigma_\tau$. If $\rho = \infty$, all other nodes are eligible. The set $\mathcal{N}_{K,\rho}(i)$ consists of at most K eligible candidates with the largest k_{ij}. Ties are resolved by a fixed rule unchanged by a common additive shift of all plug-in values. A neighborhood with fewer than K eligible candidates is not padded. The unordered retained edge set is

$$\mathcal{E}_{K,\rho} = \big\{\{i,j\} : i < j,\ j \in \mathcal{N}_{K,\rho}(i),\ i \in \mathcal{N}_{K,\rho}(j)\big\}.$$

Write $m = |\mathcal{E}_{K,\rho}|$. If $m > 0$, assign to every retained edge

$$w_{ij} = \frac{k_{ij}}{\displaystyle\sum_{\{a,b\} \in \mathcal{E}_{K,\rho}} k_{ab}}.$$

Weights of nonedges and self-pairs are zero. For $m = 0$, all weights are zero. Unordered edges are counted once throughout.

For a node vector $a = (a_1, \dots, a_n)^\mathsf{T} \in \mathbb{R}^n$, define

$$\Omega_{K,\rho}(a) = \sum_{\{i,j\} \in \mathcal{E}_{K,\rho}} w_{ij}(a_i - a_j)^2, \qquad d_i^w = \sum_{j \neq i} w_{ij}, \qquad d_{\max}^w = \max_i d_i^w. \tag{17}$$

The graph energy and all weighted degrees equal zero for an empty graph. For any node vector a, its centered empirical mean square is

$$V_n(a) = \min_{c \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^n (a_i - c)^2 = \frac{1}{n}\sum_{i=1}^n (a_i - \bar{a})^2, \qquad \bar{a} = \frac{1}{n}\sum_{i=1}^n a_i. \tag{18}$$

The last equality follows by expanding the sum as $\sum_i (a_i - \bar{a})^2 + n(c - \bar{a})^2$.

Assumption 4 (Conditional node model)**.** *For the risk statements only, work under a conditional distribution in which the covariates, τ, the graph, its weights, and $\lambda \geq 0$ are fixed. Whenever plug-in-dependent bounds are used, the plug-in values are also included in $\mathcal{H}$. A random node target $Z \in \mathbb{R}^n$ has finite second moments and*

$$\mathbb{E}[Z \mid \mathcal{H}] = \tau + b, \qquad \mathrm{Cov}(Z \mid \mathcal{H}) = \Sigma,$$

where $\mathcal{H}$ denotes the conditioned information, $b \in \mathbb{R}^n$, and Σ is a symmetric positive semidefinite matrix.

Assumption 5 (Quadratic node fit)**.** *For the risk statements only, the fitted node vector is the exact minimizer over all $f \in \mathbb{R}^n$ of*

$$\frac{1}{n}\sum_{i=1}^n (f_i - Z_i)^2 + \lambda\Omega_{K,\rho}(f). \tag{19}$$

Assumption 6 (Conditionally unbiased node target)**.** *For the improvement corollaries only, impose the additional condition $b = 0$ in Assumption 4.*

The node analysis conditions on the covariates and graph. The improvement results additionally assume conditionally unbiased targets. The covariance matrix may be nondiagonal or singular.

A.2 SELECTED-EDGE DISCREPANCY AND ERROR CONCENTRATION

Lemma 2 (Retained edges and normalization)**.** *Under Assumptions 2 and 3, every node has unweighted retained degree at most K, and $m \le nK/2$. If $m > 0$, all retained weights are positive, $\sum_{\{i,j\}\in\mathcal{E}_{K,\rho}} w_{ij} = 1$, and $\sum_i d_i^w = 2$. If also $\rho < \infty$, then*

$$d_{\max}^w \le \frac{K\exp(\rho^2/2)}{m}. \tag{20}$$

Proof. If $\{i,j\}$ is retained, $j \in \mathcal{N}_{K,\rho}(i)$. Thus the retained neighbors of i form a subset of a set of size at most K. Summing degrees counts each unordered edge twice, so $2m \le nK$. Each score is a positive real number because its exponent is finite. For $m > 0$ the normalization denominator is therefore positive, and summing the normalized scores gives one. Summing the weighted degrees counts each retained weight twice and gives two. For finite ρ, every retained edge is eligible and hence

$$\exp(-\rho^2/2) \le k_{ij} \le 1.$$

Consequently the denominator is at least $m\exp(-\rho^2/2)$. Each retained weight is at most $\frac{\exp(\rho^2/2)}{m}$. There are at most K such terms in each d_i^w, proving (20). □

Proposition 1 (Selected-edge discrepancy and graph energy)**.** *Under Assumptions 2 and 3, every retained edge obeys*

$$\big||\tau_i - \tau_j| - |\widetilde{\tau}_i - \widetilde{\tau}_j|\big| \le |\delta_i - \delta_j|. \tag{21}$$

At graph level,

$$\big|\sqrt{\Omega_{K,\rho}(\tau)} - \sqrt{\Omega_{K,\rho}(\widetilde{\tau})}\big| \le \sqrt{\Omega_{K,\rho}(\delta)}. \tag{22}$$

For finite ρ this yields

$$\sqrt{\Omega_{K,\rho}(\tau)} \le \sqrt{\Omega_{K,\rho}(\widetilde{\tau})} + \sqrt{\Omega_{K,\rho}(\delta)} \le \rho\sigma_\tau + \sqrt{\Omega_{K,\rho}(\delta)}. \tag{23}$$

These claims are valid even when the graph is constructed from the same data used to estimate $\widetilde{\tau}$.

Proof. The identity $\tau_i - \tau_j = (\widetilde{\tau}_i - \widetilde{\tau}_j) - (\delta_i - \delta_j)$ and the scalar triangle inequality show $|\tau_i - \tau_j| \le |\widetilde{\tau}_i - \widetilde{\tau}_j| + |\delta_i - \delta_j|$. Interchanging the first two differences gives the reverse bound and proves (21).

When $m > 0$, choose the orientation $i < j$ for each edge and define an edge vector $v(a) \in \mathbb{R}^m$ with coordinates $\sqrt{w_{ij}}(a_i - a_j)$. Then $\|v(a)\|_2^2 = \Omega_{K,\rho}(a)$ and $v(\tau) = v(\widetilde{\tau}) - v(\delta)$. For finite vectors u, v, Cauchy–Schwarz gives $\|u+v\|_2^2 \le \|u\|_2^2 + 2\|u\|_2\|v\|_2 + \|v\|_2^2$, which proves the Euclidean triangle inequality. Applying it in both directions to these edge vectors proves (22). For finite ρ, eligibility and Lemma 2 give

$$\Omega_{K,\rho}(\widetilde{\tau}) \le \sum_{\{i,j\}\in\mathcal{E}_{K,\rho}} w_{ij}\rho^2\sigma_\tau^2 = \rho^2\sigma_\tau^2.$$

This proves (23). If $m = 0$, all three energies are zero, so all displayed inequalities hold directly. Every argument is pointwise in the realized values and graph. □

Proposition 2 (Centered node error and weight concentration)**.** *Under Assumptions 2 and 3,*

$$\Omega_{K,\rho}(\delta) \le 2nd_{\max}^w V_n(\delta). \tag{24}$$

If $m > 0$ and $\rho < \infty$, then

$$\Omega_{K,\rho}(\delta) \le \frac{2nK\exp(\rho^2/2)}{m} V_n(\delta), \tag{25}$$

$$\sqrt{\Omega_{K,\rho}(\tau)} \le \rho\sigma_\tau + \sqrt{\frac{2nK\exp(\rho^2/2)}{m} V_n(\delta)}. \tag{26}$$

Proof. For every $c \in \mathbb{R}$, the nonnegativity of $((\delta_i - c) + (\delta_j - c))^2$ implies

$$(\delta_i - \delta_j)^2 \le 2(\delta_i - c)^2 + 2(\delta_j - c)^2.$$

Multiplying by w_{ij} and summing each unordered edge once gives

$$\Omega_{K,\rho}(\delta) \le 2\sum_i d_i^w (\delta_i - c)^2 \le 2d_{\max}^w \sum_i (\delta_i - c)^2.$$

Set $c = \bar{\delta}$ and use (18) to obtain (24). This also holds for $m = 0$ because its two sides are zero. Substitute Lemma 2 into (24) for (25), and substitute that result into Proposition 1 for (26). □

Proof of Lemma 1. Proposition 1 gives the pairwise and graph-energy inequalities, and Proposition 2 gives the centered-error bound. Lemma 2 supplies the weighted-degree bound for a nonempty finite-radius graph. These arguments are pointwise in the fitted graph, including graphs constructed from plug-in CATE estimates fitted on the same sample. □

Boundary cases. For a common plug-in error $\delta_i = c$, $V_n(\delta) = \Omega_{K,\rho}(\delta) = 0$, so the plug-in and true graph energies agree exactly. For $\rho = \infty$, the triangle and centered-error bounds remain valid. The finite-radius certificate is then omitted. For $\rho = 0$, retained plug-in values are equal and $\Omega_{K,\rho}(\widetilde{\tau}) = 0$. Changing K or ρ can alter both the retained edges and their weights.

A.3 Conditional fixed-node risk

Lemma 3 (From the scalar penalty to a Laplacian)**.** *Under Assumptions 2 and 3, let W be the symmetric $n \times n$ matrix with off-diagonal entries w_{ij} and zero diagonal. Let $D = \mathrm{diag}(d_1^w, \ldots, d_n^w)$ and $L = D - W$. For every $f \in \mathbb{R}^n$,*

$$f^\mathsf{T} L f = \Omega_{K,\rho}(f). \tag{27}$$

L is real, symmetric, positive semidefinite, and $L\mathbf{1} = 0$.

Proof. Symmetry follows from $w_{ij} = w_{ji}$. Expanding the quadratic form,

$$f^\mathsf{T} L f = \sum_i \sum_{j \ne i} w_{ij} f_i^2 - \sum_i \sum_{j \ne i} w_{ij} f_i f_j = \sum_{\{i,j\} \in \mathcal{E}_{K,\rho}} w_{ij}(f_i^2 + f_j^2 - 2 f_i f_j),$$

which is (27). Its terms are nonnegative, establishing positive semidefiniteness. Each row sum is zero by the definition of d_i^w, proving $L\mathbf{1} = 0$. □

Proposition 3 (Exact conditional risk identity)**.** *Under Assumptions 2, 3, 4, and 5, define*

$$S_\lambda = (I + n\lambda L)^{-1}.$$

The unique solution of (19) *is $\widehat{f}_\lambda = S_\lambda Z$. Its conditional mean squared node error is*

$$\frac{1}{n}\mathbb{E}[\|\widehat{f}_\lambda - \tau\|_2^2 \mid \mathcal{H}] = \frac{1}{n}\|(S_\lambda - I)\tau + S_\lambda b\|_2^2 + \frac{1}{n}\,\mathrm{tr}(S_\lambda \Sigma S_\lambda). \tag{28}$$

The unregularized solution is Z and has conditional risk $(\|b\|_2^2 + \mathrm{tr}\,\Sigma)/n$.

Proof. By Lemma 3, the objective is $\|f - Z\|_2^2/n + \lambda f^\mathsf{T} L f$. Its gradient is $2(f - Z)/n + 2\lambda L f$ and its Hessian is $2I/n + 2\lambda L$. For any nonzero v the Hessian quadratic form is $2\|v\|_2^2/n + 2\lambda v^\mathsf{T} L v > 0$. Thus there is a unique minimizer, determined by $(I + n\lambda L)f = Z$. The matrix is positive definite and invertible, giving the stated solution. It is symmetric, so its inverse is symmetric.

Set the centered target $\varepsilon = Z - \tau - b$. Assumption 4 gives $\mathbb{E}[\varepsilon \mid \mathcal{H}] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^\mathsf{T} \mid \mathcal{H}] = \Sigma$. Then

$$\widehat{f}_\lambda - \tau = (S_\lambda - I)\tau + S_\lambda b + S_\lambda \varepsilon.$$

The first two terms are fixed conditional on $\mathcal{H}$, so their inner product with $S_\lambda \varepsilon$ has conditional expectation zero. For the remaining squared norm, finite second moments permit a finite sum expansion:

$$\mathbb{E}[\|S_\lambda \varepsilon\|_2^2 \mid \mathcal{H}] = \sum_{i,j,k} (S_\lambda)_{ij}(S_\lambda)_{ik}\Sigma_{jk} = \mathrm{tr}(S_\lambda \Sigma S_\lambda).$$

Dividing the expanded error by n proves (28). At $\lambda = 0$, $S_0 = I$ and the final baseline formula follows. □

Proposition 4 (Smoothing bias and variance)**.** *Under the assumptions of Proposition 3,*

$$\frac{1}{n}\|(S_\lambda - I)\tau\|_2^2 \le \frac{\lambda}{4}\Omega_{K,\rho}(\tau), \tag{29}$$

and

$$0 \le \frac{1}{n}\operatorname{tr}(S_\lambda \Sigma S_\lambda) \le \frac{1}{n}\operatorname{tr}\Sigma. \tag{30}$$

Under Assumption 6, the quantity in (29) *is the entire squared-bias term of* (28)*. Otherwise, the additional $S_\lambda b$ term and its cross term must be retained.*

Proof. Lemma 3 permits the spectral decomposition $L = U\,\mathrm{diag}(\ell_1, \dots, \ell_n)U^\mathsf{T}$, where U is orthogonal and each $\ell_j \ge 0$. Thus

$$S_\lambda = U\,\mathrm{diag}\left(\frac{1}{1+n\lambda\ell_1}, \dots, \frac{1}{1+n\lambda\ell_n}\right)U^\mathsf{T}.$$

Write $t_j = (U^\mathsf{T}\tau)_j$. For every $u \ge 0$,

$$\frac{u}{4} - \frac{u^2}{(1+u)^2} = \frac{u(u-1)^2}{4(1+u)^2} \ge 0.$$

It follows that

$$\frac{1}{n}\|(S_\lambda - I)\tau\|_2^2 = \frac{1}{n}\sum_j \frac{(n\lambda\ell_j)^2}{(1+n\lambda\ell_j)^2}t_j^2 \le \frac{\lambda}{4}\sum_j \ell_j t_j^2 = \frac{\lambda}{4}\Omega_{K,\rho}(\tau).$$

The diagonal entries of $U^\mathsf{T}\Sigma U$ are nonnegative because Σ is positive semidefinite. Finite-sum rearrangement of the trace gives

$$\operatorname{tr}(S_\lambda \Sigma S_\lambda) = \sum_j \frac{(U^\mathsf{T}\Sigma U)_{jj}}{(1+n\lambda\ell_j)^2}, \qquad \operatorname{tr}\Sigma = \sum_j (U^\mathsf{T}\Sigma U)_{jj}.$$

Each multiplier lies in $[0, 1]$, proving (30). Finally, (28) contains $S_\lambda b$ as well as $(S_\lambda - I)\tau$. Setting $b = 0$ removes the former term, as claimed. □

Corollary 1 (A sufficient condition for improvement)**.** *Under all assumptions of Proposition 3 and the additional Assumption 6, define the conditional variance reduction*

$$\Delta_{\mathrm{var}}(\lambda) = \frac{1}{n}\{\operatorname{tr}\Sigma - \operatorname{tr}(S_\lambda \Sigma S_\lambda)\}.$$

If

$$\frac{\lambda}{4}\Omega_{K,\rho}(\tau) < \Delta_{\mathrm{var}}(\lambda), \tag{31}$$

then the conditional mean squared node error of $\widehat{f}_\lambda$ is strictly smaller than that of Z. A sufficient, possibly more conservative condition is

$$\frac{\lambda}{4}\left\{\sqrt{\Omega_{K,\rho}(\widetilde{\tau})} + \sqrt{2nd_{\max}^w V_n(\delta)}\right\}^2 < \Delta_{\mathrm{var}}(\lambda). \tag{32}$$

For a nonempty graph and finite ρ, the expression in braces can be replaced by the larger quantity $\rho\sigma_\tau + \sqrt{2nK\exp(\rho^2/2)V_n(\delta)/m}$.

Proof. Subtract the $b = 0$ baseline from (28). The exact difference is

$$\frac{1}{n}\|(S_\lambda - I)\tau\|_2^2 - \Delta_{\mathrm{var}}(\lambda).$$

Proposition 4 upper-bounds the first term by $\lambda\Omega_{K,\rho}(\tau)/4$, proving the first implication. Propositions 1 and 2 give

$$\sqrt{\Omega_{K,\rho}(\tau)} \le \sqrt{\Omega_{K,\rho}(\widetilde{\tau})} + \sqrt{2nd_{\max}^w V_n(\delta)}.$$

Both sides are nonnegative, so squaring and multiplying by $\lambda/4$ preserves the inequality. This proves (32) is sufficient. Equation (26) proves the last replacement under its stated additional conditions. □

Remark 1 (Boundary cases)**.** *For an empty graph or $\lambda = 0$, $S_\lambda = I$ and the node risk is unchanged. If $\Omega_{K,\rho}(\tau) = 0$, then $S_\lambda\tau = \tau$, so smoothing introduces no bias under $b = 0$. Strict variance reduction requires positive noise variance along at least one Laplacian eigenvector with positive eigenvalue.*

A.4 COMMON SHIFTS AND POINTWISE PENALTIES

Proposition 5 (Translation invariance of the graph and penalty)**.** *Under Assumptions 2 and 3, a common shift $\widetilde{\tau}_i \mapsto \widetilde{\tau}_i + c$ leaves the graph and weights unchanged whenever the scale σ_τ is fixed or itself unchanged by translation. The MAD-based scale, with a standard-deviation fallback and then a fixed positive constant fallback, has this property. For any fixed graph, any $f \in \mathbb{R}^n$, and any $c \in \mathbb{R}$,*

$$\Omega_{K,\rho}(f + c\mathbf{1}) = \Omega_{K,\rho}(f). \tag{33}$$

In contrast, the pointwise penalty

$$P(f, \widetilde{\tau}) = \frac{1}{n}\sum_i (f_i - \widetilde{\tau}_i)^2$$

is not generally invariant when only the plug-in CATE values are shifted.

Proof. Each plug-in difference is unchanged by a common shift. If the scale is unchanged, all eligibility tests and scores are unchanged. The fixed tie rule then gives the same neighborhoods, mutual edges, and normalized weights. To verify the asserted scale property, let the median be any fixed order-statistic convention (including the average of the two central values for even n). Adding c shifts every order statistic and its median by c. Hence each absolute deviation from the median is unchanged, as is their median and any fixed multiplicative correction. The mean shifts by c, so every deviation used in a standard deviation is unchanged. Therefore the choice between MAD, standard deviation, and the constant fallback is also unchanged. Equation (33) follows term by term from $(f_i + c) - (f_j + c) = f_i - f_j$. For the pointwise penalty, direct expansion yields

$$P(f, \widetilde{\tau} + c\mathbf{1}) - P(f, \widetilde{\tau}) = -\frac{2c}{n}\sum_i (f_i - \widetilde{\tau}_i) + c^2.$$

For example, if $f = \widetilde{\tau}$ and $c \neq 0$, the difference is $c^2 > 0$. □

The graph penalty acts on pairwise differences of final CATE predictions, while the base objective retains its learner-specific dependence on nuisance estimates.

A.5 FUNCTION RISK, SAMPLING ASSUMPTIONS, AND DR SCORES

Assumption 7 (Sampling and tails)**.** *The observations $O_i = (X_i, A_i, Y_i)$, $i = 1, \ldots, n$, are independent and identically distributed, $n \geq 2$, and $A_i \in \{0, 1\}$. Write $e(x) = \mathbb{P}(A = 1 \mid X = x)$, $\mu_a(x) = \mathbb{E}[Y \mid X = x, A = a]$, and $\tau(x) = \mu_1(x) - \mu_0(x)$. Almost surely, $\|X\|_\infty \leq R_x < \infty$, $|\tau(X)| \leq T < \infty$, and $0 < e(X) < 1$. For every real t,*

$$\mathbb{E}[\exp\{t(Y - \mu_A(X))\} \mid X, A] \leq \exp(t^2\nu^2/2).$$

The constants R_x, T, ν are deterministic. The usual identification assumptions give τ its causal interpretation. They are separate from these estimation conditions.

The sub-Gaussian residual condition allows unbounded outcomes. All risks below use one fixed outcome scale. Put

$$\mathcal{R}(f) = \mathbb{E}_{X'}\{f(X') - \tau(X')\}^2, \qquad Q_n(f) = \frac{1}{n}\sum_i \{f(X_i) - \tau(X_i)\}^2,$$

where X' is independent of training and has the same covariate law. Thus $\mathcal{R}(\widehat{f})$ is random through the training data.

Assumption 8 (Fixed architecture)**.** *Let $\phi(t) = t$ for $t \geq 0$ and $\phi(t) = e^t - 1$ for $t < 0$. For a deterministic $M \geq 0$, let $\mathcal{F}_M$ contain*

$$f_\theta(x) = W_4\phi(W_3\{W_2\phi(W_1 x + b_1) + b_2\} + b_3) + b_4 \tag{34}$$

with affine dimensions $d \to 25 \to 25 \to 37 \to 1$, and every weight and bias coordinate in $[-M, M]$.

Affine input standardization and a positive outcome scale can be absorbed into the first and last affine maps. The bounds are stated for a fixed bound on these effective parameters.

A.5.1 FITTED AND REFERENCE DR SCORES

Write $e_i = e(X_i)$ and $\widehat{e}_i = \widehat{e}(X_i)$. Fix an augmentation threshold $q \in (0, 1/2)$ and denominator protection $d_0 \geq 0$, including the implementation values $q = 0.05$, $d_0 = 10^{-9}$. Define

$$h(p) = \begin{cases} \dfrac{1}{p + d_0}, & p \geq q, \\ 0, & p < q. \end{cases}$$

The fitted score and a reference score used only for analysis are

$$\widehat{Z}_i = \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) + A_i h(\widehat{e}_i)\{Y_i - \widehat{\mu}_1(X_i)\} - (1 - A_i)h(1 - \widehat{e}_i)\{Y_i - \widehat{\mu}_0(X_i)\}, \tag{35}$$

$$Z_i^0 = \tau(X_i) + A_i h(e_i)\{Y_i - \mu_1(X_i)\} - (1 - A_i)h(1 - e_i)\{Y_i - \mu_0(X_i)\}. \tag{36}$$

Fitted outcome regressions are finite and fitted propensities lie in $[0, 1]$. They may use all training observations. The reference uses the *same* d_0 as the actual score. It is neither a new learner nor an oracle input to training. Let

$$\varepsilon_i = Z_i^0 - \tau(X_i), \quad \zeta_n = \left\{ \frac{1}{n} \sum_i (\widehat{Z}_i - Z_i^0)^2 \right\}^{1/2}, \qquad \sigma = \frac{\nu}{q}.$$

The actual graph has nonnegative normalized weights w_{ij} on unordered edges, with each edge counted once. Write

$$\Omega_{K,\rho}(f) = \sum_{i<j} w_{ij}\{f(X_i) - f(X_j)\}^2, \qquad \widehat{J}_D(f) = \frac{1}{n} \sum_i \{f(X_i) - \widehat{Z}_i\}^2 + \lambda \Omega_{K,\rho}(f).$$

An empty graph has zero weights and energy. For a fitted predictor $\widehat{f} \in \mathcal{F}_M$, define its full-objective gap

$$\epsilon_{\text{opt}} = \widehat{J}_D(\widehat{f}) - \min_{g \in \mathcal{F}_M} \widehat{J}_D(g) \geq 0. \tag{37}$$

The minimum exists because the parameter box is compact and the fixed realized objective is continuous in target parameters. The objective-gap formulation covers any approximate minimizer whose objective value is within ϵ_{opt} of the minimum.

A.6 EXPLICIT NEURAL FUNCTION-RISK BOUNDS

For fixed M, define the following constants directly from the architecture:

$$\begin{aligned} p &= 25d + 1675, \\ B &= 23125M^4(dR_x + 1) + 925M^3 + 37M^2 + M, \\ C &= 92500M^3(dR_x + 1) + 2775M^2 + 74M + 1. \end{aligned} \tag{38}$$

For a deterministic covering accuracy $t > 0$ and failure level $\alpha \in (0, 1)$, put

$$H_F = p \log(1 + 2MC/t), \qquad v = H_F + \log(8/\alpha), \qquad u_0 = \log(4/\alpha),$$

$$\Gamma = 4(B + T)t + (B + T)^2 \sqrt{\frac{v}{2n}}, \tag{39}$$

$$\Pi = \sigma \left[B\sqrt{\frac{2v}{n}} + t\sqrt{1 + 2\sqrt{\frac{u_0}{n}} + \frac{2u_0}{n}} \right]. \tag{40}$$

Theorem 2 (Neural function-risk comparison)**.** *Under Assumptions 7 and 8, with probability at least $1 - 3\alpha/4$, the following holds simultaneously for all finite fitted outcome regressions and propensities in $[0, 1]$, nonnegative realized graphs, choices $\lambda \geq 0$, and fitted predictors $\widehat{f} \in \mathcal{F}_M$:*

$$\mathcal{R}(\widehat{f}) + \lambda \Omega_{K,\rho}(\widehat{f}) \leq \inf_{g \in \mathcal{F}_M} \{\mathcal{R}(g) + \lambda \Omega_{K,\rho}(g)\} + 2\Gamma + 4\Pi + 4B\zeta_n + \epsilon_{\text{opt}}. \tag{41}$$

The simultaneous statement includes graphs, nuisance estimates, target fits, and hyperparameters selected from the training sample. The fitted-score and optimization terms refer to the selected fit.

The proof is given in Appendix A.7. The comparison retains function approximation and comparator regularization cost in the infimum, including the case $\tau \notin \mathcal{F}_M$.

A.7 UNIFORM ESTIMATION BOUNDS AND RISK PROOFS

Lemma 4 (Network envelope and cover)**.** *Under Assumptions 7 and 8, every member of $\mathcal{F}_M$ satisfies $\|f\|_\infty \le B$, $\|f_\theta - f_{\theta'}\|_\infty \le C\,\|\theta - \theta'\|_\infty$, and $\log N(t, \mathcal{F}_M, \|\cdot\|_\infty) \le H_F$.*

Proof. Both ELU and the identity fix zero and are 1-Lipschitz. With widths $(m_0, \dots, m_4) = (d, 25, 25, 37, 1)$, the output envelope at layer l obeys

$$r_0 = R_x, \qquad r_l = M(m_{l-1}r_{l-1} + 1).$$

Thus

$$r_1 = M(dR_x + 1), \quad r_2 = 25M^2(dR_x + 1) + M, \quad r_3 = 625M^3(dR_x + 1) + 25M^2 + M,$$

and $r_4 = B$. If the maximum parameter difference is v, subtracting the two affine maps and using the previous envelope gives a difference at layer l of at most $c_l v$, where

$$c_0 = 0, \qquad c_l = m_{l-1}Mc_{l-1} + m_{l-1}r_{l-1} + 1.$$

Here $c_1 = dR_x + 1$, $c_2 = 50M(dR_x + 1) + 1$, $c_3 = 1875M^2(dR_x + 1) + 50M + 1$, and $c_4 = C$. The parameter count is $(25d+25)+(625+25)+(925+37)+(37+1) = p$. For $M > 0$, subdivide each parameter interval into $\lceil 2MC/t \rceil$ intervals of length at most t/C. The midpoint grid gives a function cover of accuracy at most t, with at most $(1 + 2MC/t)^p$ centers, all in the same box. For $M = 0$, the zero function is a one-element cover. □

Lemma 5 (Uniform design error)**.** *With probability at least $1-\alpha/4$, $\sup_{f\in\mathcal{F}_M} |Q_n(f) - \mathcal{R}(f)| \le \Gamma$.*

Proof. For completeness, if $U \in [0, H_0]$, the second derivative of $\log \mathbb{E}e^{sU}$ is the variance under exponential tilting, at most $H_0^2/4$. Integrating twice shows $\log \mathbb{E}e^{s(U-\mathbb{E}U)} \le s^2H_0^2/8$ for $s \ge 0$. Independence, exponential Markov, and $s = 4z/H_0^2$ give an upper-tail probability $\exp(-2nz^2/H_0^2)$ for the sample average. Applying the same argument to $H_0 - U$ gives the lower tail. For each network-cover center use $U = (f(X) - \tau(X))^2 \in [0, (B + T)^2]$, and take a union bound. The resulting center error is at most $(B + T)^2\sqrt{v/(2n)}$ except with probability $\alpha/4$. For two functions at sup distance t, their squared losses differ pointwise by at most $2(B+T)t$. There are one empirical and one population replacement, giving the remaining $4(B + T)t$. The case $B + T = 0$ is immediate. Compactness and continuity of the parametrization imply that the supremum is measurable (a countable dense parameter subset suffices). □

Lemma 6 (Reference-score noise)**.** *Under Assumption 7, $\mathbb{E}[Z_i^0 \mid X_i] = \tau(X_i)$. Conditional on $X_1, \dots, X_n$, the noise vector ε satisfies*

$$\mathbb{E}[\exp(t^\top \varepsilon) \mid X_1, \dots, X_n] \le \exp(\sigma^2 \|t\|_2^2/2), \qquad t \in \mathbb{R}^n.$$

Proof. Each observed-group residual has conditional mean zero given X_i, A_i. Its signed multiplier in (36) has magnitude at most $1/q$, including the zero-multiplier case. Its conditional moment generating function is therefore bounded by $\exp(t_i^2\nu^2/(2q^2))$. Averaging over $A_i \mid X_i$ preserves the bound and the zero mean. Conditional independence across observations then multiplies these bounds. This argument concerns true-nuisance reference scores, not fitted scores. □

Lemma 7 (Quadratic noise bound)**.** *If a mean-zero vector z obeys $\mathbb{E}e^{t^\top z} \le e^{\sigma^2\|t\|_2^2/2}$ for all t, and S is a fixed positive semidefinite matrix, then for every $u > 0$,*

$$\mathbb{P}\Big(z^\top Sz > \sigma^2\{\operatorname{tr} S + 2\sqrt{\operatorname{tr}(S^2)u} + 2\,\|S\|_{\mathrm{op}}\,u\}\Big) \le e^{-u}.$$

Proof. This is the centered case of Theorem 2.1 of Hsu et al. (2012), instantiated with that theorem's matrix $A = S^{1/2}$. We also give the needed derivation. For an independent standard Gaussian vector g and $0 < t < 1/(2\sigma^2\,\|S\|_{\mathrm{op}})$, Gaussian integration and the vector exponential-moment assumption give

$$\mathbb{E}e^{tz^\top Sz} = \mathbb{E}_g\mathbb{E}_z e^{\sqrt{2t}\,g^\top S^{1/2}z} \le \mathbb{E}_g e^{t\sigma^2 g^\top Sg} = \prod_r (1 - 2t\sigma^2 s_r)^{-1/2},$$

where s_r are the eigenvalues of S. All integrands are nonnegative, so interchanging expectations is valid. Using the power series of $-\log(1-x)$, the centered log moment generating function is at most

$$\log \mathbb{E} e^{t(z^\top Sz - \sigma^2 \operatorname{tr} S)} \le \frac{t^2 \sigma^4 \operatorname{tr}(S^2)}{1 - 2t\sigma^2 \|S\|_{\mathrm{op}}}.$$

Let $A_0 = \sigma^4 \operatorname{tr}(S^2)$, $b_0 = 2\sigma^2 \|S\|_{\mathrm{op}}$. For $A_0 > 0$, the choice $t = \sqrt{u}/(\sqrt{A_0} + b_0\sqrt{u})$ in the exponential Markov bound at $2\sqrt{A_0 u} + b_0 u$ makes the exponent $-u$. If $A_0 = 0$, either the matrix or the noise is zero almost surely. This establishes all cases. □

Lemma 8 (Uniform multiplier bound)**.** *Except on an event of probability at most $\alpha/2$,*

$$\sup_{f \in \mathcal{F}_M} \left| \frac{1}{n} \sum_i \varepsilon_i f(X_i) \right| \le \Pi.$$

Proof. Given $X_1, \dots, X_n$, each deterministic covering function has sample squared norm at most nB^2. Lemma 6 and the one-dimensional exponential Markov bound give a two-sided probability at most $2e^{-v}$ above $\sigma B\sqrt{2v/n}$. The cover union costs at most $\alpha/4$. For the covering replacement, Cauchy–Schwarz gives

$$\left| \frac{1}{n} \sum_i \varepsilon_i (f - f_c)(X_i) \right| \le t \sqrt{\frac{1}{n} \sum_i \varepsilon_i^2}.$$

Lemma 7 at $S = I, u = u_0$ bounds the last square root by $\sigma\sqrt{1 + 2\sqrt{u_0/n} + 2u_0/n}$ with failure at most $\alpha/4$. Conditional bounds integrate to the same unconditional probabilities. □

Proof of Theorem 2. For every comparator $g \in \mathcal{F}_M$, the definition of the gap yields $\widehat{J}_D(\widehat{f}) \le \widehat{J}_D(g) + \epsilon_{\mathrm{opt}}$. Writing $\widehat{Z}_i = \tau(X_i) + \varepsilon_i + e_i^Z$ and expanding the two squares gives

$$\begin{aligned} Q_n(\widehat{f}) + \lambda \Omega_{K,\rho}(\widehat{f}) \le Q_n(g) + \lambda \Omega_{K,\rho}(g) + \frac{2}{n} \sum_i \varepsilon_i \{\widehat{f}(X_i) - g(X_i)\} \\ + \frac{2}{n} \sum_i e_i^Z \{\widehat{f}(X_i) - g(X_i)\} + \epsilon_{\mathrm{opt}}. \end{aligned} \tag{42}$$

The noise term is bounded by 4Π on the event of Lemma 8. The score term is at most $4B\zeta_n$, because $|\widehat{f} - g| \le 2B$. Two applications of Lemma 5 cost 2Γ. These events are uniform over the entire fixed class, hence hold for data-dependent fitted and comparator functions. Taking the infimum over g proves the theorem, with total failure $3\alpha/4$. The concentration event is defined from the reference noise and fixed function class and therefore remains valid after selecting the fitted graph and nuisance estimates. □

Proof of Theorem 1. The same argument applies to the compact class $\mathcal{F}$ in Assumption 1. We give the substitutions explicitly. Use its deterministic uniform t-cover in place of the parameter-box cover, replace H_F by $H_{\mathcal{F}}(t)$, and use its envelope B. The reference-score proof uses the conditional residual bound and the coefficient bound $1/q$. The design argument in Lemma 5 then gives

$$\sup_{f \in \mathcal{F}} |Q_n(f) - \mathcal{R}(f)| \le \Gamma_n$$

with failure at most $\alpha/4$: each center loss lies in $[0, (B+T)^2]$, the covering union uses $v = H_{\mathcal{F}}(t) + \log(8/\alpha)$, and the two cover replacements contribute $4(B+T)t$. Compactness in the uniform norm supplies a countable dense subset, which makes the supremum measurable. The multiplier argument in Lemma 8 gives

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i f(X_i) \right| \le \Pi_n$$

with failure at most $\alpha/2$. In that argument the center vectors have squared norm at most nB^2, the cover union costs $\alpha/4$, and the replacement costs at most $t\{\frac{1}{n}\sum_i \varepsilon_i^2\}^{1/2}$, controlled at failure $\alpha/4$

by Lemma 7 with $S = I$. For every $g \in \mathcal{F}$, expanding approximate empirical optimality gives Equation (42). The noise difference is at most $4\Pi_n$, and Cauchy–Schwarz with $|\widehat{f} - g| \le 2B$ bounds the score difference by $4B\zeta_n$. Replacing the two empirical risks costs $2\Gamma_n$. Taking the infimum over g proves the result on an event with total failure at most $3\alpha/4$. This event depends on the reference noise and the fixed function class, so it applies simultaneously to the data-dependent graph, nuisance estimates, and fitted predictor. □

A.8 Fitted-score error and regularization bias

Proposition 6 (Same-sample fitted-score error)**.** *Let $u_{a,i} = \widehat{\mu}_a(X_i) - \mu_a(X_i)$, $r_i = Y_i - \mu_{A_i}(X_i)$, and $p_i = A_i e_i + (1 - A_i)(1 - e_i)$. Define $\widehat{p}_i$ analogously and $F_i = \mathbf{1}\{\mathbf{1}(\widehat{p}_i \ge q) \ne \mathbf{1}(p_i \ge q)\}$. With empirical norm $\|v\|_n = (\frac{1}{n}\sum_i v_i^2)^{1/2}$, the following holds for every finite realization:*

$$\zeta_n \le \|u_1 - u_0\|_n + \frac{\|u_A\|_n}{q} + \frac{\|r\,|\widehat{e} - e|\|_n}{q^2} + \frac{\|rF\|_n}{q}. \tag{43}$$

Under Assumption 7, suppose deterministic bounds $a_0, a_1, a_e \ge 0$ satisfy, jointly with probability at least $1 - \beta_n$,

$$\max_{1\le i\le n} |u_{0,i}| \le a_0, \qquad \max_{1\le i\le n} |u_{1,i}| \le a_1, \qquad \max_{1\le i\le n} |\widehat{e}_i - e_i| \le a_e,$$

and

$$\mathbb{P}\{\mathrm{dist}(e(X), \{q, 1-q\}) \le t\} \le C_q t^\kappa$$

for constants $C_q < \infty$, $\kappa > 0$, $t_0 > 0$ and $0 < t \le t_0$, with $0 < a_e \le t_0$, then, for any $\gamma \in (0, 1)$, with probability at least $1 - \beta_n - 2\gamma$,

$$\zeta_n \le a_0 + a_1 + \frac{\max(a_0, a_1)}{q} + \frac{\nu a_e}{q^2\sqrt{\gamma}} + \frac{\nu\sqrt{C_q}\, a_e^{\kappa/2}}{q\sqrt{\gamma}}. \tag{44}$$

If $a_e = 0$, all gate flips are absent and the last two terms are zero.

Proof. When both arguments of h are at least q, $|h(\widehat{p}) - h(p)| \le |\widehat{p} - p|/q^2$. When both are below q, the difference is zero. When only one is at least q, its magnitude is at most $1/q$. Thus the coefficient difference is at most $|\widehat{e} - e|/q^2 + F/q$. Direct subtraction of the two scores gives

$$\widehat{Z} - Z^0 = (u_1 - u_0) - (2A - 1)h(\widehat{p})u_A + (2A - 1)\{h(\widehat{p}) - h(p)\}r.$$

The empirical norm triangle inequality proves (43).

On the stated nuisance-error event a gate flip implies $\mathrm{dist}(e_i, \{q, 1 - q\}) \le a_e$. Conditional sub-Gaussianity implies $\mathbb{E}[r_i^2 \mid X_i, A_i] \le \nu^2$. Consequently $\mathbb{E}P_n r^2 \le \nu^2$ and

$$\mathbb{E}P_n[r^2\mathbf{1}\{\mathrm{dist}(e, \{q, 1 - q\}) \le a_e\}] \le \nu^2 C_q a_e^\kappa.$$

Two applications of Markov at failure γ, followed by intersection with the nuisance event, give (44). □

Proposition 6 connects fitted-score error to nuisance accuracy and to the propensity margin around the augmentation threshold.

Proposition 7 (Selected-edge bound for comparator regularization cost)**.** *For the actual normalized graph let $\widetilde{\tau}_i = \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)$ and $\delta_i = \widetilde{\tau}_i - \tau(X_i)$. For any comparator g, put $a_i = g(X_i) - \tau(X_i)$ and $d^w_{\max} = \max_i \sum_j w_{ij}$. Then*

$$\begin{aligned}
\sqrt{\Omega_{K,\rho}(\tau)} &\le \sqrt{\Omega_{K,\rho}(\widetilde{\tau})} + \sqrt{\Omega_{K,\rho}(\delta)}, \\
\sqrt{\Omega_{K,\rho}(g)} &\le \sqrt{\Omega_{K,\rho}(\widetilde{\tau})} + \sqrt{\Omega_{K,\rho}(\delta)} + \sqrt{2nd^w_{\max} Q_n(g)}.
\end{aligned} \tag{45}$$

For finite radius, $\sqrt{\Omega_{K,\rho}(\widetilde{\tau})} \le \rho\sigma_\tau$. If $\|g - \tau\|_\infty \le a_\infty$, the last term in (45) *may instead be $2a_\infty$. Also $\Omega_{K,\rho}(\delta) \le 2nd^w_{\max} \min_c \frac{1}{n}\sum_i (\delta_i - c)^2$.*

Proof. Apply the Euclidean norm triangle inequality to vectors indexed by unordered edges with coordinates $\sqrt{w_{ij}}(v_i - v_j)$. Next,

$$\Omega_{K,\rho}(a) \le 2\sum_{i<j} w_{ij}(a_i^2 + a_j^2) = 2\sum_i d_i^w a_i^2 \le 2n d_{\max}^w Q_n(g).$$

The radius bounds every retained plug-in difference by $\rho\sigma_\tau$, and the weights sum to one when nonempty. Uniform approximation bounds every difference $a_i - a_j$ by $2a_\infty$. Replacing a_i by $\delta_i - c$, whose common shift cancels, and minimizing over c proves the last assertion. All empty-graph assertions hold with energy zero. □

The comparator bound retains approximation error when the true CATE lies outside the function class.

A.9 Rates and conditional consistency

Corollary 2 (Fixed-class rate and conditional consistency)**.** *Fix d, R_x, M, T, ν, q with $B + T > 0$, and use $t = (B + T)/n$. For each fixed confidence level, Γ, Π are $O(\sqrt{\log n/n})$, with constants depending only on these fixed quantities and confidence. If, for a sequence of selected fits, $\zeta_n = o_p(1)$, $\epsilon_{\text{opt}} = o_p(1)$, and there exist deterministic comparators $g_n \in \mathcal{F}_M$ such that*

$$\mathcal{R}(g_n) \longrightarrow \inf_{g\in\mathcal{F}_M} \mathcal{R}(g), \qquad \lambda_n \Omega_{K,\rho}(g_n) = o_p(1),$$

then $\mathcal{R}(\widehat{f}_n) \to \inf_{g\in\mathcal{F}_M} \mathcal{R}(g)$ in probability. This gives consistency for τ only when that approximation infimum is zero. The same conclusion with limit zero holds for a sequence of classes if the displayed sampling terms, fitted-score term $B_n\zeta_n$, optimization error, and comparator risk and graph cost all vanish.

Proof. With the specified t,

$$H_F = p\log\left(1 + \frac{2MCn}{B+T}\right), \quad \Gamma = \frac{4(B+T)^2}{n} + (B+T)^2\sqrt{\frac{H_F + \log(8/\alpha)}{2n}}.$$

For $c = 2MC/(B + T)$ and $n \ge 1$, $\log(1 + cn) \le \log n + \log(1 + c)$. Substitution in (40) bounds its two terms by

$$\sigma B\sqrt{\frac{2\{p\log n + p\log(1+c) + \log(8/\alpha)\}}{n}} \quad \text{and} \quad \frac{\sigma(B+T)}{n}\sqrt{1 + 2\sqrt{\frac{u_0}{n}} + \frac{2u_0}{n}},$$

respectively. These explicit inequalities establish the stated rates. Apply Theorem 2 with g_n and drop its nonnegative graph term on the left. For any fixed tolerance, take an arbitrarily small fixed failure probability and then let n grow. The assumed remainders vanish in probability. The deterministic lower bound $\mathcal{R}(\widehat{f}_n) \ge \inf_{\mathcal{F}_M} \mathcal{R}$ gives convergence to that infimum. For varying classes, the same argument applies when the stated quantities vanish. When $B + T = 0$, the target class and CATE are zero and the comparison is immediate. □

B Experimental and Implementation Details

All experiments were conducted on a workstation equipped with an AMD Ryzen 9 9950X CPU (16 cores and 32 threads), 128 GB of system memory, and a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. The implementation used Python 3.9.23 and PyTorch 2.8.0 (Paszke et al., 2019) with CUDA 12.9. All model training and inference used this GPU. Unless a table explicitly states original outcome units, root PEHE in the appendix uses the outcome standardization defined in the main experimental setup.

B.1 Learner-Specific Objectives

Write $\widehat{e}_i = \widehat{e}(X_i)$ and $\widehat{\mu}_{a,i} = \widehat{\mu}_a(X_i)$. Let $\widehat{m}_i = \widehat{e}_i\widehat{\mu}_{1,i} + (1 - \widehat{e}_i)\widehat{\mu}_{0,i}$. The same loss definitions apply to in-sample and cross-fitted nuisance predictions.

DR-learner. Following the OAR implementation (Melnychuk et al., 2026), the DR-learner uses a stabilized version of the DR pseudo-outcome (Kennedy, 2023) with augmentation threshold $q = 0.05$ and denominator protection $d_0 = 10^{-9}$:

$$Z_i^{\mathrm{DR}} = \widehat{\mu}_{1,i} - \widehat{\mu}_{0,i} + \frac{\mathbf{1}\{A_i = 1, \widehat{e}_i \geq q\}}{\widehat{e}_i + d_0}(Y_i - \widehat{\mu}_{1,i}) - \frac{\mathbf{1}\{A_i = 0, 1 - \widehat{e}_i \geq q\}}{1 - \widehat{e}_i + d_0}(Y_i - \widehat{\mu}_{0,i}). \tag{46}$$

For a single-output model f_θ, its base objective is

$$\widehat{\mathcal{L}}_{\mathrm{DR}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{f_\theta(X_i) - Z_i^{\mathrm{DR}}\}^2. \tag{47}$$

The indicators implement the low-propensity stabilization used by OAR. The corresponding augmentation term is omitted when the estimated probability of the observed treatment is below q. The denominator is protected by d_0. All DR regularization comparisons use this same stabilized objective and vary the regularization strategy applied to the second-stage model.

R-learner. The R-learner uses the Robinson-residual objective (Nie & Wager, 2021):

$$\widehat{\mathcal{L}}_{\mathrm{R}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{m}_i - (A_i - \widehat{e}_i)f_\theta(X_i)\}^2. \tag{48}$$

IVW-learner. We use the IVW implementation in OAR (Melnychuk et al., 2026), which weights the stabilized DR squared loss by the squared treatment residual:

$$\widehat{\mathcal{L}}_{\mathrm{IVW}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(A_i - \widehat{e}_i)^2\{f_\theta(X_i) - Z_i^{\mathrm{DR}}\}^2. \tag{49}$$

The weight depends on the observed treatment as well as the estimated propensity. Both None and NEAR use this same empirical objective.

RA-learner. The RA-learner (Curth & Van der Schaar, 2021) constructs the imputed effect

$$Z_i^{\mathrm{RA}} = A_i(Y_i - \widehat{\mu}_{0,i}) + (1 - A_i)(\widehat{\mu}_{1,i} - Y_i), \tag{50}$$

and minimizes

$$\widehat{\mathcal{L}}_{\mathrm{RA}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{f_\theta(X_i) - Z_i^{\mathrm{RA}}\}^2. \tag{51}$$

PW-learner. The PW-learner (Curth & Van der Schaar, 2021) uses $\bar{e}_i^{\mathrm{PW}} = \mathrm{clip}(\widehat{e}_i, \epsilon_{\mathrm{PW}}, 1-\epsilon_{\mathrm{PW}})$, with $\epsilon_{\mathrm{PW}} = 0.05$ in the reported experiments. The propensity-weighted transformed outcome and its regression loss are

$$Z_i^{\mathrm{PW}} = \frac{A_i Y_i}{\bar{e}_i^{\mathrm{PW}}} - \frac{(1 - A_i)Y_i}{1 - \bar{e}_i^{\mathrm{PW}}}, \tag{52}$$

$$\widehat{\mathcal{L}}_{\mathrm{PW}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{f_\theta(X_i) - Z_i^{\mathrm{PW}}\}^2. \tag{53}$$

X-learner. The X-learner (Künzel et al., 2019) does not use a single pooled pseudo-outcome. Let $n_a = \sum_{i=1}^{n}\mathbf{1}\{A_i = a\}$. A control-sample branch $\tau_{\mathrm{c}}(x;\theta_{\mathrm{c}})$ is trained on $D_i^{\mathrm{c}} = \widehat{\mu}_{1,i} - Y_i$ for $A_i = 0$, and a treated-sample branch $\tau_{\mathrm{t}}(x;\theta_{\mathrm{t}})$ is trained on $D_i^{\mathrm{t}} = Y_i - \widehat{\mu}_{0,i}$ for $A_i = 1$. With $\theta = (\theta_{\mathrm{c}}, \theta_{\mathrm{t}})$, the implemented base loss is

$$\widehat{\mathcal{L}}_{\mathrm{X}}(\theta) = \frac{1}{2n_0}\sum_{i:A_i=0}\{\tau_{\mathrm{c}}(X_i;\theta_{\mathrm{c}}) - D_i^{\mathrm{c}}\}^2 + \frac{1}{2n_1}\sum_{i:A_i=1}\{\tau_{\mathrm{t}}(X_i;\theta_{\mathrm{t}}) - D_i^{\mathrm{t}}\}^2. \tag{54}$$

The final scalar CATE prediction is the propensity mixture

$$f_\theta(x) = \widehat{e}(x)\tau_{\mathrm{c}}(x;\theta_{\mathrm{c}}) + \{1-\widehat{e}(x)\}\tau_{\mathrm{t}}(x;\theta_{\mathrm{t}}), \tag{55}$$

where $\widehat{e}(x)$ is the fitted propensity. The unified interface in the main text therefore accommodates a multi-branch parameterization by applying NEAR to the learner's final scalar output f_θ.

For DR-, R-, IVW-, RA-, and PW-learners, f_θ is the single effect model appearing directly in the corresponding base loss. For the X-learner, it is the mixture in Equation (55). In every case, NEAR leaves the learner-specific base loss unchanged and applies its auxiliary pairwise penalty to this final scalar output:

$$\widehat{\mathcal{L}}(\theta;\widehat{\eta}) + \lambda \sum_{\{i,j\}\in\mathcal{E}_{K,\rho}} w_{ij}\{f_\theta(X_i) - f_\theta(X_j)\}^2. \tag{56}$$

B.2 Comparators and Evaluation Details

Following OAR (Melnychuk et al., 2026), Constant applies a fixed, nonadaptive regularization level: representation dropout with probability 0.1 on IHDP and additive Gaussian input noise with standard deviation 0.05 on ACIC. The dOAR comparator uses the corresponding debiased variant, with overlap-adaptive dropout on IHDP and overlap-adaptive input noise on ACIC. Plug-in distillation penalizes pointwise deviation from the plug-in CATE and selects its strength from $\{0.01, 0.1, 1, 10, 100\}$ separately within each outer evaluation unit using the DR-learner validation loss. The Direct plug-in comparator uses $\widehat{\mu}_1(x) - \widehat{\mu}_0(x)$ as its CATE prediction without fitting a second-stage model.

For the Synthetic Oracle diagnostic, the reference is each repetition's realized NEAR graph. The Oracle retains exactly the same active nodes and number of edges. It first pairs adjacent units after sorting active nodes by their true CATE values, connecting an unmatched last unit to the nearest other active unit if needed. It then fills the edge budget in order of increasing true CATE difference. Gaussian scores use the robust scale of true CATE values on active nodes and are normalized across the final edges. The true CATE enters only this Oracle diagnostic. NEAR uses the plug-in CATE estimates defined in Section 4.

The graph-support diagnostic uses uniform edge weights for every graph. NEAR-U retains the same plug-in-CATE mutual top-K edge set as NEAR. Covariate Uniform uses mutual K-NN on standardized covariates. Random Uniform samples the same number of undirected edges as the realized NEAR graph. Shuffled Uniform permutes the plug-in CATE values before applying the same radius and mutual top-K construction. Random and Shuffled use two prespecified graph seeds per evaluation unit. For Random and Shuffled, results are averaged over these seeds within each evaluation unit before aggregation. For each dataset and graph combination, one λ is selected from $\{10, 100, 1000\}$ on IHDP and $\{1, 10, 100\}$ on ACIC by minimizing the mean validation loss over five prespecified calibration units. The selected value is then frozen for all 100 IHDP splits or 77 ACIC settings. The diagnostic that isolates edge weighting compares Gaussian and uniform weights on the same NEAR graph and applies the same validation procedure. All compared variants share nuisance predictions, target initialization, architecture, and optimization settings within each evaluation unit.

B.3 Graph Construction and Model Selection

For the IHDP and ACIC benchmarks in Table 1, the propensity network has hidden width 25. Outcome and target networks use representation width 25 and head width 37 with ELU activations. Nuisance models use learning rate 0.001, batch size 128, and weight decay 0.001. Target models use learning rate 0.005, batch size 64, and weight decay 0.01. Training uses 200 epochs without early stopping, and target predictions use exponential moving-average parameters with decay 0.995.

The canonical graph uses plug-in CATE similarity and normalizes weights over the retained edges. When the bandwidth is not fixed externally, let $m_\tau = \mathrm{median}\{\widetilde{\tau}_j\}_{j=1}^n$, where $\widetilde{\tau}_j = \widehat{\mu}_1(X_j) - \widehat{\mu}_0(X_j)$, and set

$$\sigma_\tau = 1.4826\,\mathrm{median}\{|\widetilde{\tau}_i - m_\tau|\}_{i=1}^n. \tag{57}$$

The factor 1.4826 gives the Gaussian-consistent scaling of the median absolute deviation (Rousseeuw & Croux, 1993). If this scaled median absolute deviation is zero, the standard devi-

Algorithm 1 NEAR graph construction

Input: training data $\mathcal{D}_n$, predictions on the training sample $\{\widehat{\mu}_0(X_i), \widehat{\mu}_1(X_i)\}_{i=1}^n$, bandwidth σ_τ, radius ρ, and neighbor cap K
Output: mutual edge set $\mathcal{E}_{K,\rho}$ and normalized weights $\{w_{ij}\}$

for $i = 1, \ldots, n$ **do**
  Compute $\widehat{\mu}_{a,i} = \widehat{\mu}_a(X_i)$, $a \in \{0, 1\}$, and set $\widetilde{\tau}_i = \widehat{\mu}_{1,i} - \widehat{\mu}_{0,i}$.
end for
Initialize $\mathcal{C}_\rho(i) \leftarrow \varnothing$ for every i.
for each block of unordered pairs $i < j$ **do**
  Compute $k_{ij} = \exp\left\{-\frac{(\widetilde{\tau}_i - \widetilde{\tau}_j)^2}{2\sigma_\tau^2}\right\}$.
  if $\frac{|\widetilde{\tau}_i - \widetilde{\tau}_j|}{\sigma_\tau} \leq \rho$ **then**
    Insert j into $\mathcal{C}_\rho(i)$ and i into $\mathcal{C}_\rho(j)$, retaining k_{ij}.
  end if
end for
for each node i **do**
  Let $\mathcal{N}_{K,\rho}(i)$ contain the indices corresponding to the largest $\min\{K, |\mathcal{C}_\rho(i)|\}$ scores, using deterministic tie-breaking.
end for
Form $\mathcal{E}_{K,\rho}$ from unordered pairs $\{i, j\}$ satisfying $j \in \mathcal{N}_{K,\rho}(i)$ and $i \in \mathcal{N}_{K,\rho}(j)$.
if $\mathcal{E}_{K,\rho} = \varnothing$ **then**
  return $\varnothing, \varnothing$.
end if
Compute $Z_{K,\rho} = \sum_{\{a,b\} \in \mathcal{E}_{K,\rho}} k_{ab}$.
Set $w_{ij} = \frac{k_{ij}}{Z_{K,\rho}}$ for every $\{i, j\} \in \mathcal{E}_{K,\rho}$.
return $\mathcal{E}_{K,\rho}$ and $\{w_{ij}\}$.

ation of the plug-in CATE estimates in the training sample is used. If both quantities are zero, σ_τ is set to one, which makes all similarity scores equal to one.

NEAR jointly selects three hyperparameters because they control different components of its regularizer: ρ sets the standardized plug-in CATE radius, K caps the local neighborhood, and λ controls the penalty strength. Mutual top-K controls graph density. A finite ρ prevents forced neighbor selection when a node has fewer than K sufficiently close plug-in CATE estimates. The benchmark hyperparameter grid is

$$\lambda \in \{0.1, 1, 10, 100\}, \qquad K \in \{1, 2, 4, 8\},$$
$$\rho \in \{2.146, 1.794, 1.354\},$$

giving 48 configurations. These radii correspond to minimum Gaussian similarities of approximately 0.1, 0.2, and 0.4 at the radius boundary through $\exp(-\rho^2/2)$. For Table 1, each learner and dataset configuration is selected by mean validation loss over ten calibration units: ten IHDP splits, or five splits from each of two ACIC settings. The selected configuration is then fixed across all 100 IHDP splits or all 77 ACIC simulation settings. Each ACIC setting contributes one reported outer evaluation result. Synthetic and the diagnostic experiments use their stated internal validation procedures. Nuisance estimation, graph construction, and target training use the outer training sample, and the selected model is evaluated once on the test data. The reported experiments use full-sample nuisance fitting ($L = 1$). NEAR also accepts cross-fitted nuisance estimates, as examined in Table 8.

Mutual top-K caps graph degree, while the radius determines which neighbors are eligible. Figure 2(c) examines this interaction at $K = 16$, including $\rho = \infty$ as the unfiltered case.

Algorithms 1 and 2 summarize the complete graph construction and second-stage training procedures.

Algorithm 2 NEAR training with a fixed graph

Input: training data $\mathcal{D}_n$, base two-stage learner, nuisance model classes and fitting procedure, $\sigma_\tau, K, \rho, \lambda$, observation and edge batch sizes $b_{\text{obs}}, b_{\text{edge}}$, number of updates N_{step}, and step sizes $\{\alpha_t\}_{t=1}^{N_{\text{step}}}$

Output: fitted CATE predictor $f_{\widehat{\theta}_{\lambda,K,\rho}}$

1: Fit $\widehat{\mu}_0, \widehat{\mu}_1$ and any additional nuisance functions required by the base learner using the specified procedure, and obtain their predictions on the training sample.
2: Call Algorithm 1 to obtain $\mathcal{E}_{K,\rho}$ and w_{ij}.
3: Construct the learner-specific second-stage loss and initialize f_θ.
4: **for** $t = 1, \ldots, N_{\text{step}}$ **do**
5: Sample the observation minibatches required by the learner, each of size b_{obs} (arm-specific for the X-learner), and compute the mean base loss over the sampled observations $\widehat{\mathcal{L}}_t(\theta; \widehat{\eta})$.
6: **if** $\lambda > 0$ and $\mathcal{E}_{K,\rho} \neq \varnothing$ **then**
7: Draw $E_{t,1}, \ldots, E_{t,b_{\text{edge}}}$ i.i.d. with replacement from the categorical distribution with probabilities $\{w_{ij}\}$.
8: Write $E_{t,r} = (I_{t,r}, J_{t,r})$ and compute $\widehat{\Omega}_t = \frac{1}{b_{\text{edge}}} \sum_{r=1}^{b_{\text{edge}}} \{f_\theta(X_{I_{t,r}}) - f_\theta(X_{J_{t,r}})\}^2$.
9: **else**
10: Set $\widehat{\Omega}_t = 0$.
11: **end if**
12: Set $\widehat{\mathcal{J}}_t = \widehat{\mathcal{L}}_t + \lambda \widehat{\Omega}_t$.
13: Update $\theta \leftarrow \theta - \alpha_t \nabla_\theta \widehat{\mathcal{J}}_t$.
14: **end for**
15: **return** $f_{\widehat{\theta}_{\lambda,K,\rho}}$.

B.4 Synthetic Design

The Synthetic benchmark uses the following sharp data-generating process (DGP). Each unit $i = 1, \ldots, n$ has $X_i = (X_{i1}, \ldots, X_{i,10})^\top$. The variables are generated in the following order:

$$\begin{aligned} X_{ij} &\overset{\text{iid}}{\sim} \text{Unif}(-1, 1), && j = 1, \ldots, 10, \\ A_i \mid X_i &\sim \text{Bernoulli}\{e_\gamma(X_i)\}, \\ Y_i(a) &= \mu_0(X_i) + a\tau_{\text{sharp}}(X_i) + \varepsilon_i, && a \in \{0, 1\}, \end{aligned}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ independently of X_i, and the observed outcome is $Y_i = Y_i(A_i)$. The same unit-level noise ε_i enters both potential outcomes. The response surface and propensity are

$$\begin{aligned} \mu_0(x) &= x_1 + x_2^2 + \sin(\pi x_3) + 0.5 x_4 x_5, \\ \ell_\gamma(x) &= \gamma(0.8x_1 - 0.6x_2 + 0.4x_3), \\ e_\gamma(x) &= \text{clip}[\text{sigmoid}\{\ell_\gamma(x)\}, 0.02, 0.98], \end{aligned}$$

and the CATE is

$$\tau_{\text{sharp}}(x) = \begin{cases} -1, & x_1 < 0, \\ 2, & x_1 \geq 0. \end{cases}$$

The main performance experiment fixes $\gamma = 3.0$ and $\sigma_\varepsilon = 1$.

Each repetition contains $n = 1000$ observations with an inner train/validation/test split of $640/160/200$. After validation selection, the first two subsets are combined for final fitting and the test subset is evaluated once. Across ten repetitions and six learners, None and NEAR share nuisance predictions, target model initialization, minibatch order, network architecture, and optimization settings. Root PEHE is reported after standardizing outcomes using the same preprocessing as OAR.

Synthetic robustness diagnostics. The sample-size and noise experiment uses the DR-learner at $\gamma = 3.0$, $n \in \{500, 1000, 2000, 5000\}$, and $\sigma_\varepsilon \in \{0.5, 1, 2\}$, with ten matched repetitions per condition. It uses the MAD-based σ_τ from Equation (57) and $(\lambda, K, \rho) = (10, 2, 1.354)$, with the

graph hyperparameters fixed across data conditions. The cross-fitting diagnostic uses the same Synthetic design and compares full-sample nuisance fitting with two- and five-fold cross-fitting over ten matched repetitions. For the bias–variance diagnostic, one independent evaluation sample of 2,048 covariate vectors is drawn from $\mathrm{Unif}(-1,1)^{10}$ and held fixed. Each method is fitted on 50 independently generated training samples, and every fitted CATE function is evaluated on this common sample. Squared CATE bias and CATE prediction variance are computed pointwise across the 50 fits and then averaged over the evaluation sample. Their sum is the CATE prediction MSE (Geman et al., 1992). All quantities are reported on the original outcome scale. Runtime scaling uses $n \in \{1000, 2000, 5000, 10000\}$, three repetitions per size, $\sigma_\varepsilon = 1$, and the same fixed graph hyperparameters as the sample-size diagnostic.

C Additional Results and Diagnostics

C.1 Direct Plug-in and Pointwise Distillation

The Direct comparator predicts $\widetilde{\tau}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x)$ without a second-stage model. Plug-in distillation uses the same plug-in CATE values as a pointwise target. With sg denoting the stop-gradient operator, its DR-learner objective is

$$\widehat{\mathcal{L}}_{\mathrm{DR}}(\theta) + \frac{\lambda_{\mathrm{distill}}}{n} \sum_{i=1}^{n} \{f_\theta(X_i) - \mathrm{sg}(\widetilde{\tau}_i)\}^2.$$

The two comparators therefore reuse the first-stage plug-in CATE without selecting pairs for regularization. Table 3 reports their root PEHE alongside None and NEAR.

Table 7: Paired root PEHE differences (Comparator minus NEAR) for DR. Positive differences and wins indicate lower NEAR error. Standard deviations are computed across the paired evaluation units.

Dataset	Comparator	Mean difference ± SD	NEAR wins
IHDP	Direct	0.037 ± 0.074	$70/100$
IHDP	Distill	0.041 ± 0.079	$66/100$
ACIC	Direct	0.028 ± 0.063	$65/77$
ACIC	Distill	0.017 ± 0.049	$53/77$

C.2 Robustness to Nuisance Estimation

Table 8 compares full-sample nuisance fitting with two- and five-fold cross-fitting on Synthetic. NEAR has lower root PEHE than None for every fold count.

Table 8: Robustness to nuisance estimation on Synthetic over ten matched repetitions. Values are $\sqrt{\epsilon_{\mathrm{PEHE}}}$, reported as mean ± standard deviation. $L = 1$ denotes full-sample nuisance fitting. $L > 1$ denotes L-fold cross-fitting.

Nuisance folds L	None	NEAR
1	0.93 ± 0.14	$\mathbf{0.47 \pm 0.07}$
2	1.49 ± 0.38	$\mathbf{0.62 \pm 0.09}$
5	1.33 ± 0.32	$\mathbf{0.56 \pm 0.12}$

C.3 Robustness to Sample Size and Outcome Noise

Figure 3 reports the complete DR-learner sweep over

$$n \in \{500, 1000, 2000, 5000\}, \qquad \sigma_\varepsilon \in \{0.5, 1, 2\},$$

at $\gamma = 3.0$. The NEAR graph hyperparameters are fixed across the matched data conditions. NEAR has lower mean root PEHE in all 12 conditions.

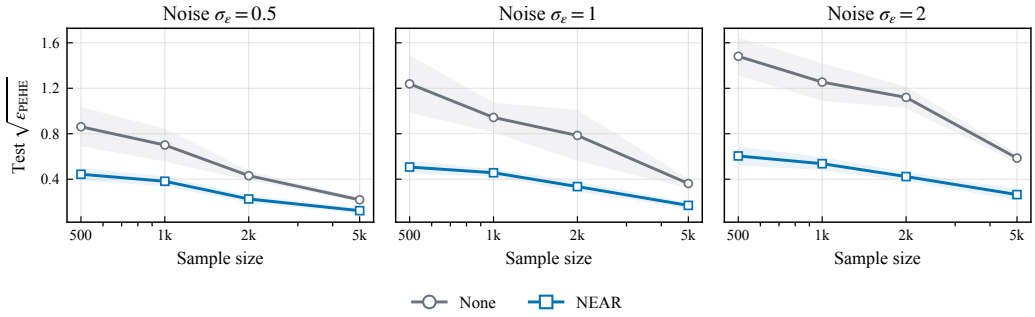


Figure 3: Robustness to sample size and outcome noise for the DR-learner on Synthetic at $\gamma = 3.0$. Panels correspond to $\sigma_\varepsilon \in \{0.5, 1, 2\}$. Lines show means over ten matched repetitions, and shaded bands denote one standard deviation.

C.4 Plug-in CATE Radius Sensitivity

Figure 2(b) uses IHDP splits 0 through 9 and Figure 2(c) uses splits 10 through 19. Table 9 fixes $\lambda = 10$ and varies ρ for $K = 16$ on IHDP. Moderate radii retain almost all edges, while very small radii remove most edges, isolate nodes, and worsen root PEHE. Extreme filtering also changes the true CATE variation among the remaining edges.

Table 9: IHDP plug-in CATE radius sensitivity for DR-learner at $K = 16$ over ten repeated evaluations (mean $\pm$ standard deviation). Retention is relative to $\rho = \infty$. The final column reports normalized selected-edge CATE variation and the isolated-node ratio.

ρ	$\sqrt{\epsilon_{\text{PEHE}}}$	Edge retention	$\widetilde{I}_G(\tau)$ / isolation
∞	0.52 ± 0.24	1.00	0.34 / 0.00
0.30	0.51 ± 0.24	0.99	0.34 / 0.00
0.10	0.51 ± 0.24	0.94	0.33 / 0.01
0.05	0.51 ± 0.24	0.86	0.33 / 0.02
0.01	0.51 ± 0.24	0.27	0.34 / 0.11
0.001	0.57 ± 0.29	0.03	0.35 / 0.70
0.0001	0.64 ± 0.30	< 0.01	0.45 / 0.96

C.5 Graph Support and Edge Weight Diagnostics

The graph support diagnostic in the main text compares graph constructions under uniform edge weights. The plug-in-CATE mutual top-K graph uses $(K, \rho) = (2, 1.354)$ on IHDP and $(4, 1.794)$ on ACIC. One λ is selected for each graph and dataset using five prespecified calibration units and the DR-learner validation loss. The selected value is 10 for all IHDP graphs. On ACIC, Gaussian NEAR selects 10, while uniform NEAR, Covariate, Random, and Shuffled select 1. These values are frozen for the full evaluations.

Table 10 compares Gaussian and uniform weighting on the same NEAR graph, with each variant calibrated by the same validation procedure. The two variants have the same displayed mean on IHDP. Gaussian weights yield lower mean root PEHE on ACIC.

Table 10: DR-learner diagnostic of edge weighting on a fixed NEAR graph. Values are $\sqrt{\epsilon_{\text{PEHE}}}$, reported as mean $\pm$ standard deviation (lower is better).

Dataset	Gaussian	Uniform
IHDP	$\mathbf{0.51 \pm 0.26}$	$\mathbf{0.51 \pm 0.26}$
ACIC	$\mathbf{0.47 \pm 0.21}$	0.49 ± 0.21

C.6 Additional Treatment Effect Structures

The smooth, grouped, and complex effect structures retain the covariates, propensity, response surface, sample size, and noise distribution of the main Synthetic benchmark. Their CATE functions are

$$\tau_{\text{smooth}}(x) = 1 + x_1 + 0.5x_2,$$
$$\tau_{\text{grouped}}(x) = \begin{cases} -1, & x_1 x_2 < 0, \\ 1, & x_1 x_2 \geq 0, \end{cases}$$
$$\tau_{\text{complex}}(x) = \sin(3\pi x_1) + x_2 x_3.$$

For each structure, we use ten repetitions at $n = 1000$ and $\gamma = 3.0$, with the graph hyperparameters and target-model settings fixed from the corresponding main Synthetic experiment. Table 11 shows that NEAR has lower mean root PEHE than None under all three structures.

Table 11: DR-learner results under additional Synthetic treatment effect structures. Values are $\sqrt{\epsilon_{\text{PEHE}}}$, reported as mean $\pm$ standard deviation over ten repetitions (lower is better).

Effect structure	None	NEAR
Smooth	1.15 ± 0.24	$\mathbf{0.37 \pm 0.11}$
Grouped	1.35 ± 0.12	$\mathbf{0.69 \pm 0.10}$
Complex	1.49 ± 0.23	$\mathbf{0.64 \pm 0.10}$

C.7 Bias–Variance Diagnostic

We draw a fixed, independent evaluation sample $\{X'_j\}_{j=1}^{N_{\text{eval}}}$ of size $N_{\text{eval}} = 2{,}048$ from the Synthetic covariate distribution, fit each method on $R = 50$ independently generated training samples, and evaluate every fitted CATE function on that same sample. Writing $\widehat{\tau}_m^{(r)}(X'_j)$ for method m's prediction from training repetition r and $\overline{\tau}_{mj} = \frac{1}{R}\sum_r \widehat{\tau}_m^{(r)}(X'_j)$, we compute

$$B_m^2 = \frac{1}{N_{\text{eval}}} \sum_j \{\overline{\tau}_{mj} - \tau(X'_j)\}^2,$$
$$V_m = \frac{1}{N_{\text{eval}}} \sum_j \frac{1}{R} \sum_r \{\widehat{\tau}_m^{(r)}(X'_j) - \overline{\tau}_{mj}\}^2,$$
$$\text{MSE}_m = \frac{1}{N_{\text{eval}} R} \sum_{j,r} \{\widehat{\tau}_m^{(r)}(X'_j) - \tau(X'_j)\}^2 = B_m^2 + V_m.$$

This is the standard bias–variance decomposition (Geman et al., 1992), averaged over the fixed evaluation sample and reported on the original outcome scale. Table 6 shows that NEAR changes squared CATE bias only slightly, from 0.40 to 0.43, while reducing CATE prediction variance from 3.83 to 0.46 and MSE from 4.23 to 0.89.

C.8 Plug-in Perturbation

This experiment uses ten independently generated repetitions of the same Synthetic design, with the benchmark hyperparameters held fixed.

Let $\widetilde{\tau}_i$ be the first-stage plug-in CATE. We perturb the values used for graph construction as

$$\widetilde{\tau}_i^{(\sigma_\delta)} = \widetilde{\tau}_i + \sigma_\delta \,\text{SD}(\widetilde{\boldsymbol{\tau}}) z_i, \qquad z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),$$

for $\sigma_\delta \in \{0, 0.25, 0.5, 1\}$, where $\text{SD}(\widetilde{\boldsymbol{\tau}})$ is the training-node standard deviation. Outcomes, treatments, true effects, the base loss, and target-training settings remain fixed. With $\delta_i^{(\sigma_\delta)} = \widetilde{\tau}_i^{(\sigma_\delta)} - \tau_i$, we compute

$$\Omega_{K,\rho}(\boldsymbol{\delta}^{(\sigma_\delta)}) = \sum_{\{i,j\} \in \mathcal{E}_{K,\rho}} w_{ij} \{\delta_i^{(\sigma_\delta)} - \delta_j^{(\sigma_\delta)}\}^2$$

on the graph constructed at each perturbation level using the edge weights entering the NEAR loss. Table 12 reports diagnostics for graphs constructed after adding increasing Gaussian perturbations to the plug-in CATE values used for graph construction. The error energy is evaluated on the final retained edges and uses the same weights as the NEAR penalty. As σ_δ increases, $\Omega_{K,\rho}(\boldsymbol{\delta}^{(\sigma_\delta)})$, normalized selected-edge CATE variation, and root PEHE increase together. Edge count and the isolated-node ratio remain stable across the four perturbation levels.

Table 12: Plug-in perturbation diagnostic on Synthetic. Values are means over ten repetitions. Increasing σ_δ perturbs only the plug-in CATE values used for graph construction. Here $\boldsymbol{\delta}$ is the error at each perturbation level. The final two columns report mean retained edge count and isolated-node ratio.

σ_δ	$\Omega_{K,\rho}(\boldsymbol{\delta})$	$\widetilde{I}_G(\tau)$	$\sqrt{\epsilon_{\text{PEHE}}}$	Edges	Isolated
0.00	1.02	0.23	0.47	1713.60	0.06
0.25	1.28	0.28	0.51	1706.70	0.07
0.50	1.90	0.42	0.58	1702.20	0.06
1.00	3.00	0.67	0.64	1711.90	0.06

C.9 Runtime Scaling

The implementation scores all candidate pairs before sparsification, requiring $O(n^2)$ pair evaluations in the worst case. Sorting the scalar plug-in CATE estimates before radius and mutual top-K selection can reduce graph-construction cost. Table 13 shows that graph construction time increases from 0.03 seconds at $n = 1{,}000$ to 3.09 seconds at $n = 10{,}000$, while total NEAR training takes 1.13 to 1.48 times as long as training the corresponding None model over this range.

Table 13: Runtime scaling on Synthetic over three repetitions, reported in seconds as mean $\pm$ standard deviation.

n	Graph	None	NEAR
1,000	0.03 ± 0.00	23.49 ± 1.45	34.48 ± 1.95
2,000	0.10 ± 0.02	47.52 ± 0.98	70.50 ± 0.30
5,000	0.72 ± 0.06	135.38 ± 1.38	195.23 ± 4.34
10,000	3.09 ± 0.08	258.67 ± 6.64	292.76 ± 1.76