INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Offline Causal Reinforcement Learning for Dynamic Marketing Uplift Optimization with Uplift-Decomposed Q-Learning

Gabriel Hayes, Grace Tan*

College of Business, Virginia Tech, USA

Abstract

Marketing teams own large logs of past campaigns but can rarely experiment with new contact policies online. Uplift models estimate the incremental effect of a single contact, yet repeated contacts interact through fatigue, carryover and unsubscription, so one-shot uplift targeting is myopic. We study offline causal reinforcement learning for this setting and propose Uplift-Decomposed Q-learning (UDQ). Under sequential ignorability, the advantage of contacting a customer equals the immediate causal uplift plus a discounted long-term uplift; UDQ estimates the first term with an honest causal forest, learns only the second by temporal-difference updates with shrinkage, and acts on a lower confidence bound built from the forest's standard errors. On the real Hillstrom randomized e-mail experiment (64,000 customers), one-shot targeting policies earn no more than blanket mailing, which motivates the dynamic view. On a 12-period benchmark calibrated on the same experiment, UDQ reaches a net value of 0.180 per customer under uniform logging, 13% above myopic causal-forest targeting, 26% above conservative Q-learning, 75% above a reimplemented causal-forest-plus-DQN framework trained in a learned simulator, and 90% of an oracle agent. Under rule-based logging it ties the best baselines while sending 27% fewer e-mails. Ablations show that the gains come from anchoring the Q-function on the causal-forest uplift and from uncertainty-based pessimism, whereas the learned long-term term must be strongly shrunk and adds little at this data scale.

Keywords *:uplift modeling, offline reinforcement learning, causal forest, heterogeneous treatment effects, customer relationship management, pessimism*

1. Introduction

Direct marketers decide, period after period, whom to contact and with which message. The incremental, or uplift, view of this decision asks how much a contact changes a customer's behavior rather than how likely the customer is to respond, and it is now standard practice in database marketing [1]–[3]. Estimators of conditional average treatment effects (CATEs), notably causal forests [4], [5] and meta-learners [6], rank customers by expected uplift, and policy-evaluation studies on randomized marketing data show that CATE-based targeting earns more than uniform policies [7]. Most uplift pipelines, however, treat each campaign as an isolated decision. In practice the same customers are contacted repeatedly and the effect of today's e-mail depends on what was sent last week: responsiveness decays under contact pressure, part of a campaign's effect persists into later periods, and every message carries a small risk that the customer unsubscribes and

is lost to the channel. Treating these decisions as a Markov decision process and optimizing long-run value with reinforcement learning (RL) has a long history in database marketing [8], [9]. Recently, Wang *et al.* [10] integrated causal forests with deep Q-networks: CATEs estimated by a causal forest bootstrap the Q-function, and a Double/Dueling DQN agent is refined inside a counterfactual simulation environment that models carryover, adaptation and interaction effects of repeated interventions. Their framework makes heterogeneous causal effects a first-class input to sequential targeting, and it is the starting point of this work.

Two obstacles remain when only historical campaign logs are available. First, a simulator calibrated from history is itself a model, and a policy optimized against it exploits its errors, the familiar model-exploitation problem of model-based policy optimization [11]. Second, learning directly from logs, i.e., offline RL [12], avoids the simulator but faces a statistical difficulty that is acute in marketing: the effect of an action is a few percentage points of response probability, whereas the outcome is a noisy binary event. Generic offline algorithms such as conservative Q-learning (CQL) [13] and batch-constrained Q-learning (BCQ) [14] must recover these small contrasts from bootstrapped, noisy targets, and they ignore the causal structure that uplift modeling has exploited for two decades. This paper takes the view that the immediate part of an action's value is exactly an uplift and should be estimated as one. We propose Uplift-Decomposed Q-learning (UDQ), an offline causal RL algorithm that writes the value of contacting a customer as the value of not contacting, plus the immediate uplift estimated by an honest causal forest, plus a learned long-term uplift. Only the last term is learned by temporal-difference (TD) updates. It is shrunk toward zero, so that UDQ falls back to myopic causal targeting wherever the data carry no evidence of long-term effects, and actions are selected by a lower confidence bound built from the forest's standard errors, following the pessimism principle for offline RL [15]. Our contributions are fourfold. (i) We formalize offline dynamic uplift optimization and derive an uplift decomposition of the action-value function under sequential ignorability, which separates an identifiable immediate CATE from a long-term continuation contrast. (ii) We propose UDQ, which turns causal-forest Q-bootstrapping from a one-off initialization into a permanent structural anchor, adds shrinkage and uncertainty-based pessimism, and needs no simulator. (iii) We build a 12-period benchmark calibrated on the real Hillstrom randomized e-mail experiment, implementing a counterfactual response model with fatigue, carryover, interaction and unsubscription, and compare UDQ with eleven baselines, including a full reimplementation of the causal-forest-plus-DQN framework, under two logging regimes. (iv) We report on real randomized data how well one-shot uplift estimators rank customers, and on the benchmark when the causal decomposition helps and when a learned simulator is preferable.

II. Related Work

A. Uplift Modeling and Heterogeneous Treatment Effects

Uplift modeling originated in direct marketing, where tree learners with uplift-specific splitting criteria were developed for single and multiple treatments [1], [16], [17]; surveys and benchmark studies are given in [2], [3]. In parallel, causal inference produced general CATE estimators: meta-learners that combine off-the-shelf regressors [6], orthogonal and residualization-based estimators [18], [19], and causal forests with honest splitting and valid confidence intervals [4], [5]. Hitsch *et al.* [7] compare CATE-based targeting policies by off-policy evaluation on randomized marketing data. Sequential treatment decisions are studied as dynamic treatment regimes in biostatistics [20], and temporal uplift models describe how marketing actions change future purchase intensity [21]. These works estimate effects over time but do not learn a long-horizon contact policy from logs.

B. Reinforcement Learning for Marketing

Pednault *et al.* [8] cast direct-mail targeting as sequential cost-sensitive decision making and solved it with batch RL on campaign histories, and Theocharous *et al.* [9] optimized life-time value in ad recommendation with off-policy evaluation before deployment. Deep Q-learning [22] and its Double and Dueling variants later became the default function approximators for such problems. The causal-forest-plus-DQN framework

discussed in Section I adds heterogeneous causal effects and an explicit model of repeated-contact dynamics, but relies on a simulator for policy learning. Wang et al. develop a dual-process sequential recommendation agent with evolving memory and confidence-gated deliberation, offering a valuable example of adaptive decision-making over user histories [31].

C. Offline and Causal Reinforcement Learning

Fitted Q iteration [23] established batch RL with function approximation. With deep networks, learning from fixed data suffers from extrapolation error on actions that are rare in the data: BCQ restricts the greedy step to actions the behavior policy would take [14], [24], and CQL penalizes the values of out-of-distribution actions [13]; see [12] for a review. Pessimism with respect to an uncertainty quantifier is provably efficient and minimax optimal for offline RL [15], and model-based offline methods penalize uncertain model transitions [11]. Causal RL [25] studies how causal knowledge improves RL. UDQ connects these lines: its uncertainty quantifier is the standard error of an honest causal forest, and its causal knowledge is the identification of the immediate uplift from sequentially randomized logs. Carbon-aware offline reinforcement learning for logistics provides a useful example of evaluating sequential policies from historical data while respecting coupled operational constraints [32].

III. Problem Formulation

A. Setting

A firm interacts with customers $i = 1, \ldots, n$ over T periods. In period t it observes a state s_t that contains customer attributes, recency-frequency-monetary (RFM) quantities and the recent contact history, and chooses an action $a_t \in A = \{0, 1, \ldots, K\}$, where 0 means no contact. It then observes a response $y_t \in \{0, 1\}$ (a website visit) and an unsubscription indicator $u_t \in \{0, 1\}$; an unsubscribed customer can no longer be contacted. The reward is

$$r_t = y_t - c \cdot \mathbf{1}[a_t \neq 0] - v \cdot u_t, \qquad (1)$$

where c is the cost of a contact and v is the value of keeping a subscriber, which prices the long-term loss caused by an unsubscription. The goal is a policy π that maximizes $J(\pi) = E_\pi[\Sigma_t \gamma^{t-1} r_t]$, and only a logged dataset $D = \{(s_t, a_t, y_t, u_t, s_{t+1})\}$ generated by a behavior policy π_b is available.

B. Identification of the Immediate Uplift

Let $Y_t(a)$ denote potential outcomes. *Assumption 1 (sequential ignorability and positivity):* given s_t, the action a_t is independent of the potential rewards and next states, and $\pi_b(a|s) \geq \eta > 0$ for every action admissible in s. The assumption holds by design when contacts are randomized, as in A/B tests, or when a known business rule is mixed with random exploration. Under Assumption 1 the immediate uplift

$$\tau(s, a) = E[r_t \mid s_t = s, a_t = a] - E[r_t \mid s_t = s, a_t = 0] \qquad (2)$$

is a CATE with covariates s, identified from D and estimable by any CATE learner.

C. Uplift Decomposition

Proposition 1: For any policy π with value function V^π and action-value function Q^π,

$$Q^\pi(s, a) - Q^\pi(s, 0) = \tau(s, a) + \gamma \cdot \Delta^\pi(s, a), \qquad (3)$$

where $\Delta^\pi(s, a) = E[V^\pi(s_{t+1}) \mid s, a] - E[V^\pi(s_{t+1}) \mid s, 0]$ is the long-term uplift of a.

Proof: Subtract the Bellman equations $Q^\pi(s, a) = E[r_t + \gamma V^\pi(s_{t+1}) \mid s, a]$ written for a and for 0; by Assumption 1 both conditional expectations are functionals of the logged distribution. ■

The identity has two consequences. The myopic uplift policy $\operatorname{argmax}_a \tau(s, a)$ is optimal exactly when the long-term term does not change the ranking of actions, and that term is the only part of the action contrast that requires bootstrapping. In e-mail marketing τ is small relative to outcome noise but well suited to CATE estimators, whereas Δ captures fatigue and attrition and is typically negative for frequent contacts.

IV. Uplift-Decomposed Q-Learning

A. Causal Uplift Module

We fit a causal forest (CausalForestDML of EconML [26]) with the visit outcome y_t, the multi-valued treatment a_t and covariates s_t on all logged transitions of subscribed customers. Outcome and propensity nuisances are fitted by gradient boosting with cross-fitting, following the orthogonalization of [19], and the

forest uses honest splitting [4]. It returns the visit uplift $\hat{\tau}_y(s, a)$ and a standard error $\hat{\sigma}(s, a)$ obtained with the bootstrap-of-little-bags procedure [5]. Because unsubscription can occur only after a contact, its hazard $\hat{h}(s, a)$ is modeled by a logistic regression on contacted transitions. The estimated immediate uplift of the reward is the anchor

$$\hat{b}(s, a) = \hat{\tau}_y(s, a) - c - v\cdot\hat{h}(s, a), \quad \hat{b}(s, 0) = 0. \quad (4)$$

B. Decomposed Q-Network

Following Proposition 1, UDQ parameterizes the action-value function as

$$Q_\theta(s, a) = V_\theta(s) + \mathbf{1}[a \neq 0]\cdot(\hat{b}(s, a) + G_\theta(s, a)), \quad (5)$$

where a shared multilayer perceptron outputs the value of not contacting $V_\theta(s)$ and K long-term terms $G_\theta(s, a)$, which estimate $\gamma\Delta(s, a)$ together with any residual error of $\hat{b}$. Every logged transition, whatever its action, informs V_θ, so the action-independent part of the value is learned from all data, and the network never has to rediscover the small immediate contrasts from noisy TD targets. Fig. 1 summarizes the pipeline.

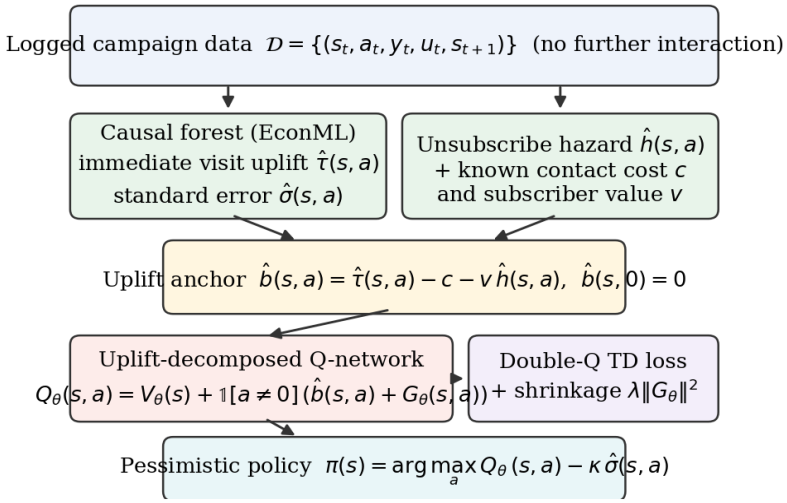


Figure 1. Overview of UDQ. The causal forest and the hazard model are fitted once on the logged data; only the value of not contacting Vϑ and the long-term uplift Gϑ are learned by TD updates, and actions are chosen by a lower confidence bound.

C. Training Objective and Pessimistic Policy

Let $\hat{\sigma}(s, 0) = 0$ and define the pessimistic greedy action $\tilde{a}(s) = \text{argmax}_{a\in A(s)}\,[Q_\theta(s, a) - \kappa\hat{\sigma}(s, a)]$, where $A(s)$ excludes contacts for unsubscribed customers. UDQ minimizes

$$L(\theta) = E_D[(Q_\theta(s,a) - y)^2] + \lambda\cdot E_D[\Sigma_{a\neq 0}\, G_\theta(s,a)^2], \quad (6)$$

with target $y = r + \gamma(1 - d)Q_{\theta^-}(s', \tilde{a}(s'))$, a periodically synchronized target network θ^- [22] and terminal indicator d. Selecting the next action with θ and evaluating it with θ^- reduces overestimation [27]. The shrinkage weight λ pulls G_θ toward zero. Because $\hat{b}$ is fixed, the penalty acts only on the long-term term; as $\lambda \to \infty$ the minimizer has $G_\theta \equiv 0$ and the policy reduces to the pessimistic myopic rule $\text{argmax}_a\,[\hat{b}(s, a) - \kappa\hat{\sigma}(s, a)]$. UDQ thus nests causal-forest targeting with a lower confidence bound, and λ controls how far the policy may depart from it where the data support a long-term effect. The deployed policy $\pi(s) = \tilde{a}(s)$ is a lower-confidence-bound rule in the sense of [15]: an action whose immediate uplift is uncertain must earn its place through a larger estimated value. Algorithm 1 lists the procedure.

Algorithm 1: Uplift-Decomposed Q-Learning (UDQ)

Input: logged data D, cost c, subscriber value v, shrinkage λ, pessimism κ, steps N, sync period M.

1: Fit the causal forest on subscribed transitions $\rightarrow \hat{\tau}_y$, $\hat{\sigma}$; fit the hazard model $\hat{h}$.

2: Cache $\hat{b}(s, \cdot)$, $\hat{b}(s', \cdot)$ and $\hat{\sigma}(s', \cdot)$ for all transitions (Eq. 4).

3: Initialize θ at random; $\theta^- \leftarrow \theta$.

4: **for** $k = 1, \ldots, N$ **do**

5: Sample a minibatch from D.

6: $\tilde{a}(s') \leftarrow \text{argmax}_{a \in A(s')} Q_\theta(s', a) - \kappa\hat{\sigma}(s', a)$.

7: Target $r + \gamma(1 - d)Q_{\theta^-}(s', \tilde{a}(s'))$; Adam step on Eq. (6).

8: Every M steps: $\theta^- \leftarrow \theta$.

9: **end for**

10: **return** $\pi(s) = \text{argmax}_{a \in A(s)} Q_\theta(s, a) - \kappa\hat{\sigma}(s, a)$.

D. Relation to Causal-Forest Bootstrapping

Wang *et al.* [10] initialize the Q-network of a DQN agent at $Q_{\text{init}}(s, a) = \hat{\tau}(s, a) - c(a)$ for $a \neq 0$ and $Q_{\text{init}}(s, 0) = 0$, and then refine it by simulation. This bootstrapping rule is an effective way to inject causal knowledge into a value function, and UDQ builds directly on it: Eq. (5) with $G_\theta \equiv 0$ reproduces their initial Q-function up to the action-independent term V_θ. The difference is that UDQ keeps the causal estimate inside the Q-function throughout learning instead of letting TD updates overwrite it. Section VI compares the two uses, initialization (the CF-init DQN baseline) and permanent anchoring (UDQ), on identical data.

V. Experimental Setup

A. Real Randomized Data

The Hillstrom MineThatData e-mail dataset [28] records a two-week campaign in which 64,000 customers who had purchased in the previous twelve months were randomly assigned, about one third each, to a men's merchandise e-mail, a women's merchandise e-mail or no e-mail. Covariates are recency, purchase history in dollars, men's and women's merchandise indicators, a new-customer indicator, zip-code type and purchase channel; outcomes are visit, conversion and spend. Visit rates are 18.3%, 15.1% and 10.6% in the three arms. We use the data directly, to evaluate one-shot uplift estimators on held-out randomized samples, and indirectly, to calibrate the response model of the dynamic benchmark.

B. Dynamic Benchmark

Customers cannot be re-contacted in hindsight, so long-horizon policies are evaluated in a semi-synthetic environment whose primitives are learned from the real experiment. We implement the counterfactual response model of Wang *et al.* [10], in which the outcome is a baseline plus a base treatment effect modulated by carryover, adaptation and interaction terms, and use it as the ground truth. The visit probability of customer i in period t under action a is

$$p_t(a) = f_0(x_t) + C_t + \mathbf{1}[a \neq 0](\tau_a(x_t)F_{a,t} + \Sigma_b I_{ab} K_{b,t}), \quad (7)$$

where the baseline f_0 and base uplifts τ_a are gradient-boosted classifiers fitted on the control and treatment arms of all 64,000 real customers (a T-learner), $C_t = \Sigma_{j<t}\, \beta_{a(j)} \exp(-\lambda_c(t - j))$ is the carryover stock, $K_{b,t}$ are decayed interaction stocks and $F_{a,t} = \exp(-g_i(S_{a,t} + \eta S_{-a,t}))$ is the adaptation (fatigue) factor. We extend the original model in three ways that matter for long-horizon decisions: fatigue acts on decayed exposure stocks $S_{a,t} = \Sigma_{j<t}\, \mathbf{1}[a_j = a]\rho^{t-j}$, so responsiveness recovers when contacts are spaced; fatigue rates g_i differ across

customers (1.5 and 0.7 times $g_0 = 0.8$ for new and established customers); and every e-mail triggers an unsubscription with probability $\sigma(-5.5 + 0.9\Sigma_a S_{a,t} + 0.7m_{it})$, where m_{it} flags a message that does not match the customer's purchase history. Visits convert to purchases according to a logistic model fitted on real visitors; a purchase resets recency and adds a spend drawn from real purchases to the history, so baseline propensities evolve with the RFM state. The calibrated base uplifts reproduce the segment effects of the experiment, e.g., +7.2 versus an empirical +6.9 percentage points for the men's e-mail among men's-only buyers and +7.6 versus +7.1 among women's-only buyers.

Table I collects all parameters. The remaining ones are $\rho = 0.7$, $\eta = 0.5$, $\beta_a = 0.25\bar{\tau}_a$, $I_{12} = 0.15\bar{\tau}_1$, $I_{21} = 0.15\bar{\tau}_2$, one-period half-lives for carryover and interaction, $c = 0.025$, $v = 0.5$, $T = 12$ bi-weekly periods and $\gamma = 0.95$. They were fixed before any learning method was run; a check with hand-crafted rules confirmed that spacing contacts can beat myopic targeting under these dynamics, so the long-horizon problem is not trivial.

Table 1PARAMETERS OF THE DYNAMIC BENCHMARK

Component	**Setting**
Horizon, period length	$T = 12$, two weeks
Discount factor	$\gamma = 0.95$
Contact cost, subscriber value	$c = 0.025$, $v = 0.5$ (visit units)
Baseline f_0, base uplift τ_a	GBM T-learner on the real arms
Mean base uplift $\bar{\tau}_1$, $\bar{\tau}_2$	0.077, 0.046
Exposure decay ρ, cross-fatigue η	0.7, 0.5
Fatigue rate g_i	0.8×1.5 (new), 0.8×0.7 (established)
Carryover β_a, half-life	$0.25\bar{\tau}_a$, one period
Interaction I_{12}, I_{21}, half-life	$0.15\bar{\tau}_1$, $0.15\bar{\tau}_2$, one period
Unsubscription logit	$-5.5 + 0.9\Sigma_a S_{a,t} + 0.7m_{it}$
Conversion, spend	logistic on real visitors; empirical spend
Logged / evaluated customers	5,000 / 5,000, 12 periods each

C. Logged Data and Evaluation Protocol

Customers are split 70/30 into a training pool and a test pool. Each logged dataset contains 5,000 training-pool customers observed for 12 periods (60,000 transitions) under one of two behavior policies. *Uniform logging* repeats the randomized design of the original experiment, choosing each action with probability 1/3. *Rule-ε logging* sends the message that matches the purchase history every period to recent buyers (recency of at most six months) and every other period to the others, and replaces this choice by a uniformly random action with probability $\varepsilon = 0.3$. Each learned policy is deployed on 5,000 test customers in the true environment with common random numbers. We report the *net value*, i.e., the expected 12-period reward per customer minus that of never contacting, together with e-mails per customer and the share of customers lost to unsubscription; results are averaged over two independent datasets per logging policy.

D. Methods Compared

Simple policies: random, blanket (men's e-mail every period) and the logging policy. *Static uplift:* a T-learner and a causal forest fitted on first-period data, whose assignment is repeated every period. *Myopic CF:* greedy on the anchor $\hat{b}$ of Eq. (4), i.e., UDQ without learned term. *Offline RL:* Offline DQN (Double DQN with dueling head [29] on the fixed data), CQL [13] and discrete BCQ [24]. *CF-init DQN:* Offline

DQN whose network is first fitted to the causal-forest Q-function of Section IV-D. *CF-DQN:* our reimplementation of the complete causal-forest-plus-DQN framework. Its learned simulator uses a gradient-boosted baseline fitted on logged no-contact periods, the causal forest evaluated at an empty contact history as base effect, carryover, adaptation (on cumulative exposure counts, as originally specified) and interaction parameters plus response scales estimated by nonlinear least squares on logged visits, and logistic conversion and unsubscription models; a Double DQN with dueling head and prioritized experience replay [30], initialized by causal-forest bootstrapping, is trained in it for 60 episodes of 512 customers. Because the true environment extends the original response model in three ways (Section V-B), this simulator is misspecified by design, which we take into account when interpreting its results. *Oracle DQN:* the same agent trained in the true environment for 100 episodes; it uses information unavailable offline and serves only as a reference.

E. Implementation

All Q-networks share the 256-128-64 ReLU architecture of the CF-DQN framework and are trained with Adam (learning rate 3×10^{-4}, batch 256, 6,000 steps, target synchronization every 400 steps). Causal forests use 100 honest trees, minimum leaf size 40, subsample ratio 0.45 and two-fold cross-fitting (EconML 0.17.0). Hyper-parameters of UDQ ($\lambda \in \{0, 1, 3\}$, $\kappa \in \{0.5, 1\}$), CQL ($\alpha \in \{0.1, 0.5, 2\}$) and BCQ (threshold $\in \{0.1, 0.3\}$) were selected on a separate development seed and evaluated on training-pool customers outside the dataset; the selected values are $\lambda = 3$, $\kappa = 0.5$, $\alpha = 0.1$ and 0.1. All networks were implemented in NumPy and run on a single CPU core with 3 GB of memory.

VI. Results

A. Uplift Estimation on Real Data

Table II and Fig. 2 report one-shot results on three random 70/30 splits of the real experiment. Heterogeneity is modest. For the men's e-mail, the uplift among the top 30% of customers ranked by any learner is 8.0–8.9 percentage points against an average effect of 7.7; for the women's e-mail the causal forest and the DR-learner raise it from 4.5 to about 6.5. These two learners also rank customers best for the women's e-mail (normalized Qini 4.9 and 4.9), while the S-learner is best for the men's e-mail (2.5); all differences lie within about one standard deviation across splits. With a contact cost of 0.025 visits, every learner recommends contacting 98–100% of customers, and the induced policies earn IPW-estimated net values of 4.7–4.9 visits per 100 customers, indistinguishable from the 4.9 of the blanket men's e-mail. In this campaign the value of personalization therefore does not come from deciding whom to contact once; it comes from deciding how often to contact, which is the dynamic problem studied next. We use the causal forest in all dynamic experiments because it is competitive in ranking and, unlike the meta-learners, provides standard errors.

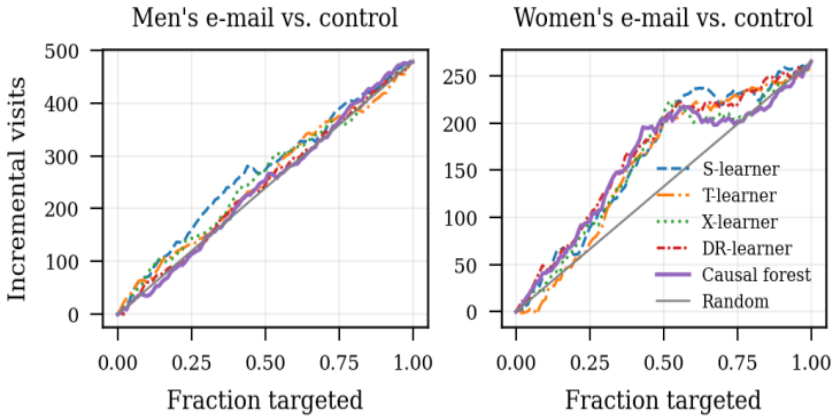


Figure 2. Qini curves of one-shot uplift estimators on a held-out split of the real Hillstrom experiment (visit outcome). The grey line corresponds to random targeting.

Table 2ONE-SHOT UPLIFT ESTIMATION ON THE REAL HILLSTROM EXPERIMENT

Learner	Qini men's	Qini women's	Top 30% men's	Top 30% women's	IPW net value
S-learner	2.5±1.2	4.5±0.5	8.9±1.2	5.5±0.1	4.8±0.4
T-learner	1.6±0.8	3.8±0.5	8.5±1.2	5.4±0.2	4.7±0.3
X-learner	1.5±1.3	4.3±0.8	8.2±1.7	6.0±0.3	4.7±0.6
DR-learner	0.9±1.4	4.9±0.3	8.0±1.7	6.6±0.5	4.8±0.6
Causal forest	1.4±1.8	4.9±1.2	8.0±1.4	6.5±0.2	4.9±0.4
Blanket men's e-mail	–	–	–	–	4.9±0.5

Normalized Qini coefficient (×100), uplift among the top 30% of the ranking (percentage points) and IPW-estimated net value of the induced one-step policy (visits per 100 customers, cost 0.025 per e-mail); mean ± s.d. over three random 70/30 splits stratified by arm.

B. Main Results on the Dynamic Benchmark

Table III and Fig. 3 summarize deployment in the true environment. Static uplift policies, which repeat a first-period assignment, do worse than random contacting: they keep mailing the same customers and lose 23–26% of them to unsubscription. The logging rule is a reasonable heuristic (15.2 against 12.2 for random contacting), and myopic causal-forest targeting improves on both (15.9 and 16.3 under the two regimes).

Under uniform logging UDQ obtains the highest net value, 18.0 per 100 customers. This is 13% above myopic CF, 20% above Offline DQN, 26% above CQL and 90% of the oracle agent (20.1). UDQ sends 5.5 e-mails per customer instead of 8.2 for myopic CF and loses 7.0% of customers instead of 14.2%, close to the oracle's 5.4 e-mails and 6.6%. Because every action is frequent in uniformly logged data, the BCQ filter never binds and BCQ coincides with Offline DQN. The reimplemented CF-DQN framework reaches 10.3, below the logging policy: trained in its learned simulator, the agent contacts about as often as random logging (7.9 e-mails) and loses 13% of customers.

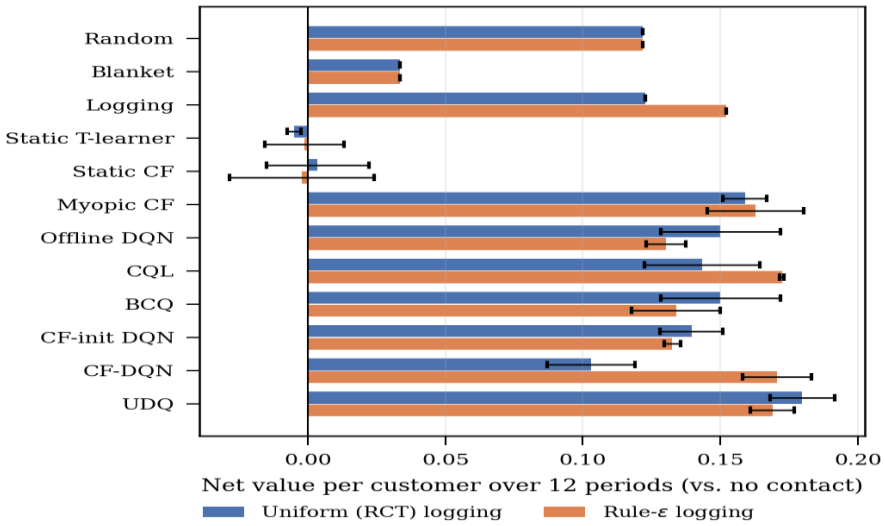


Figure 3. Net value per customer over 12 periods relative to never contacting (mean ± s.d. over two logged datasets) under uniform and rule-ε logging.

Table 3DEPLOYMENT RESULTS ON THE DYNAMIC BENCHMARK

Method	**Net unif.**	**Mails unif.**	**Unsub. unif.**	**Net rule-ε**	**Mails rule-ε**	**Unsub. rule-ε**
Random	12.2±0.0	7.6	13.9	12.2±0.0	7.6	13.9
Blanket	3.4±0.0	10.6	30.4	3.4±0.0	10.6	30.4
Logging policy	12.3±0.0	7.6	13.7	15.2±0.0	7.6	10.9
Static T-learner	-0.5±0.3	8.6	23.0	-0.1±1.4	8.5	22.5
Static CF	0.4±1.9	10.0	26.2	-0.2±2.6	9.1	25.3
Myopic CF	15.9±0.8	8.2	14.2	16.3±1.8	7.6	13.7
Offline DQN	15.0±2.2	6.8	10.5	13.0±0.7	7.0	13.7
CQL	14.3±2.1	6.9	10.5	17.2±0.1	7.1	9.8
BCQ	15.0±2.2	6.8	10.5	13.4±1.6	6.6	11.3
CF-init DQN	14.0±1.2	7.0	12.3	13.3±0.3	7.1	13.7
CF-DQN	10.3±1.6	7.9	13.3	17.1±1.3	7.1	12.4
UDQ (ours)	**18.0±1.2**	**5.5**	**7.0**	**16.9±0.8**	**5.2**	**6.9**
Oracle DQN†	20.1	5.4	6.6	20.1	5.4	6.6

Net value per 100 customers over 12 periods relative to never contacting (mean ± s.d. over two logged datasets), e-mails per customer and percentage of customers lost to unsubscription, on 5,000 test customers. †Trained online in the true environment; it does not depend on the logs and is shown as a reference only.
Under rule-ε logging the picture changes. CQL (17.2), CF-DQN (17.1) and UDQ (16.9) lie within 0.3 points of each other, less than the variation across datasets. CQL benefits from regularizing toward a behavior policy that is itself good, and the learned simulator of CF-DQN is accurate enough in this regime for its agent to profit. UDQ reaches its value with 5.2 e-mails per customer instead of 7.1 and 7.1, i.e., 27% fewer contacts, and loses 6.9% of customers instead of 9.8% and 12.4%. UDQ is thus the only method that is best or within the dataset-to-dataset variation of the best under both logging regimes; CQL and CF-DQN fall below myopic CF under uniform logging, the former because it is pulled toward a random behavior policy and the latter because of simulator mismatch.

C. Ablation Study

Table IV isolates the components of UDQ. Removing shrinkage ($\lambda = 0$) is the most harmful change (4.9 and 4.5 points): an unconstrained long-term term absorbs the noise of the TD targets and overrides good immediate decisions. Using the causal forest only to initialize an Offline DQN brings no benefit over random initialization (14.0 vs. 15.0 under uniform and 13.3 vs. 13.0 under rule-ε logging), because TD learning soon overwrites the initial values, whereas keeping the same estimate as a structural anchor (UDQ with $\kappa = 0$) is 2.1 and 3.3 points better. Pessimism adds 1.9 points under uniform and 0.3 under rule-ε logging.

The most informative comparison is with the shrinkage limit of UDQ, i.e., myopic causal-forest targeting with the same lower confidence bound ($\lambda \to \infty$, $\kappa = 0.5$). It obtains 18.0 and 17.2, within the dataset-to-dataset variation of UDQ. At this data scale, therefore, essentially all of UDQ's advantage over generic offline RL comes from anchoring decisions on the causal-forest uplift and from uncertainty-based pessimism; the learned long-term correction, once properly regularized, neither helps nor hurts, and a practitioner with similar data can deploy the $\lambda \to \infty$ member of the family without any TD training. A larger pessimism coefficient ($\kappa = 1$) contacts too rarely and loses value (15.0 and 14.1).

Table 4ABLATION OF UDQ COMPONENTS (NET VALUE PER 100 CUSTOMERS)

Variant	Net unif.	Mails unif.	Net rule-ε	Mails rule-ε
UDQ ($\lambda = 3$, $\kappa = 0.5$)	**18.0±1.2**	**5.5**	**16.9±0.8**	**5.2**
w/o pessimism ($\kappa = 0$)	16.1±0.4	7.9	16.6±1.6	7.6
w/o shrinkage ($\lambda = 0$)	13.1±3.5	6.9	12.4±2.9	5.7
CF as initialization only	14.0±1.2	7.0	13.3±0.3	7.1
No causal forest (Offline DQN)	15.0±2.2	6.8	13.0±0.7	7.0
Myopic CF ($G \equiv 0$, $\kappa = 0$)	15.9±0.8	8.2	16.3±1.8	7.6
$\lambda \to \infty$, $\kappa = 0.5$ (myopic CF + LCB)	18.0±1.5	5.7	17.2±0.6	5.2
$\lambda \to \infty$, $\kappa = 1.0$	15.0±1.1	3.3	14.1±1.0	3.1

$\lambda \to \infty$ removes the learned long-term term, leaving the causal-forest anchor with the lower confidence bound; these rows re-fit the same datasets and forests (identical seeds).

D. Policy Behavior

Fig. 4 shows contact intensity and subscriber retention over time under uniform logging. Averaged over the horizon, myopic CF contacts 68% of customers per period, CQL 58%, CF-DQN 66%, UDQ 46% and the oracle 45%. In the first period the rates are 96% for myopic CF, 84% for UDQ and 79% for the oracle; in

the last period they are 49%, 31% and 61%. At the end of the horizon UDQ retains 93.0% of subscribers, against 85.8% for myopic CF, 89.5% for CQL, 86.7% for CF-DQN and 93.4% for the oracle. The oracle raises its contact rate toward the end of the horizon, when fatigue and unsubscription no longer have future consequences; UDQ does not reproduce this horizon effect, which illustrates the limited long-term signal that its learned term extracts from the logs. Lower contact pressure is how UDQ converts uncertainty about the uplift into long-horizon value: contacts whose benefit cannot be established from the logs are skipped, which also limits fatigue and unsubscription.

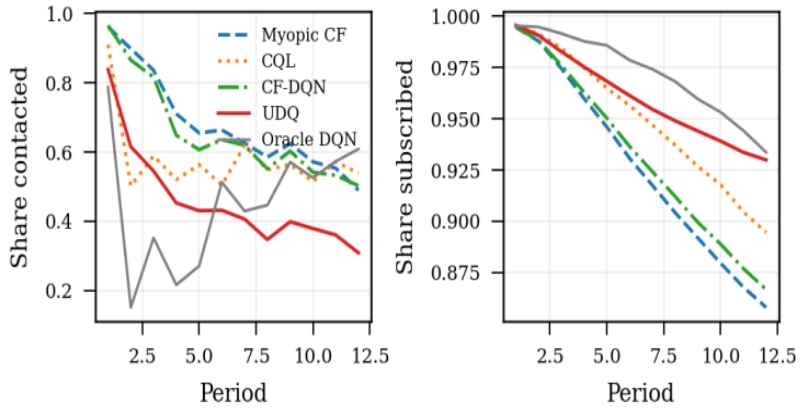


Figure 4Share of customers contacted (left) and share still subscribed (right) in each period under uniform logging (mean over two datasets). The oracle agent is trained in the true environment

E. Sensitivity to Hyper-Parameters

Table V lists the development-seed results used for model selection. UDQ is insensitive to the shrinkage weight once $\lambda \geq 1$ (18.0–18.5 per 100 customers) and to $\kappa \in \{0.5, 1\}$, whereas $\lambda = 0$ falls below myopic CF (16.3), which confirms on independent data that shrinkage is essential. CQL degrades monotonically as α grows, consistent with conservatism toward a uniformly random behavior policy, and the BCQ threshold has no effect because that behavior policy makes every action equally likely.

Table 5DEVELOPMENT-SEED RESULTS USED FOR MODEL SELECTION

Method	Grid	Net value (×100)
Myopic CF	–	16.3
UDQ, κ = 0.5	λ = 0 / 1 / 3	15.9 / 18.4 / 18.5
UDQ, κ = 1.0	λ = 0 / 1 / 3	16.1 / 18.0 / 18.1
CQL	α = 0.1 / 0.5 / 2	12.9 / 11.5 / 10.2
BCQ	τ = 0.1 / 0.3	13.6 / 13.6

Uniform logging, 5,000 customers, development seed; evaluated on 8,000 training-pool customers outside the dataset.

F. Computational Cost

On one CPU core, fitting the causal forest on about 57,000 transitions took 22 s and training UDQ 20 s more, against 18–25 s for the generic offline methods and 69 s for CF-DQN, including forest fit, simulator calibration and about 5,750 updates; the oracle agent needed 105 s. A deployed UDQ decision costs one forest prediction and one network evaluation per customer.

VII. Discussion and Limitations

Three lessons emerge. First, in offline dynamic uplift optimization the action contrasts are small and noisy, and the most effective use of logged data is to estimate them with a dedicated causal estimator and keep that estimate inside the decision rule; generic offline RL learns them only indirectly and depends on the quality of the behavior policy. Second, pessimism about the uplift acts as a data-driven frequency cap: by declining contacts whose benefit is uncertain, UDQ reduces fatigue and unsubscription, which is where most of the long-horizon value lies in our benchmark. Third, learning the long-term correction from 60,000 transitions is hard; the remaining gap of 2.1 points to the oracle agent shows that long-horizon structure is still left unexploited, and closing it will require larger logs, several campaigns or uncertainty estimates for the long-term term itself. The study has limitations. The dynamic benchmark is semi-synthetic: base responses are learned from a real randomized experiment, but fatigue, carryover, interaction and unsubscription follow an assumed parametric model, which also shapes the comparison with simulator-based methods; real repeated-contact data with randomized schedules would provide a stronger test. Results rest on two logged datasets per regime evaluated on 5,000 test customers, so differences below about one point per 100 customers should not be over-interpreted. The analysis assumes that logging decisions depend only on observed states, which holds for randomized or rule-based campaigns but not for discretionary targeting. Finally, hyper-parameters were tuned on a single development seed, and all networks were kept small to fit a single CPU core.

VIII. Conclusion

We studied offline causal reinforcement learning for repeated marketing contacts and proposed UDQ, which decomposes the action-value function into the value of not contacting, a causal-forest uplift and a long-term correction, and acts pessimistically on the forest's uncertainty. On a benchmark calibrated on a real randomized experiment, UDQ was the best method under uniform logging and tied for best under rule-based logging while sending 27% fewer e-mails; it clearly outperformed a reimplementation of the causal-forest-plus-DQN framework trained in a learned simulator under uniform logging and matched it under rule-based logging. The ablations point to a simple recipe: estimate the immediate uplift with an honest causal forest, keep it inside the value function rather than using it only for initialization, regularize long-term corrections strongly, and act on lower confidence bounds. Future work will test the approach on real multi-period campaigns with randomized schedules, attach uncertainty estimates to the long-term term, and measure how its contribution grows with the size of the logs.

References

N. J. Radcliffe and P. D. Surry, “Real-world uplift modelling with significance-based uplift trees,” Stochastic Solutions, Edinburgh, U.K., White Paper TR-2011-1, 2011.

P. Gutierrez and J.-Y. Gérardy, “Causal inference and uplift modelling: A review of the literature,” in *Proc. Int. Conf. Predictive Applications and APIs*, ser. Proc. Mach. Learn. Res., vol. 67, 2017, pp. 1–13.

F. Devriendt, D. Moldovan, and W. Verbeke, “A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics,” *Big Data*, vol. 6, no. 1, pp. 13–41, 2018, doi: 10.1089/big.2017.0104.

S. Wager and S. Athey, “Estimation and inference of heterogeneous treatment effects using random forests,” *J. Amer. Statist. Assoc.*, vol. 113, no. 523, pp. 1228–1242, 2018, doi: 10.1080/01621459.2017.1319839.

S. Athey, J. Tibshirani, and S. Wager, “Generalized random forests,” *Ann. Statist.*, vol. 47, no. 2, pp. 1148–1178, 2019, doi: 10.1214/18-AOS1709.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, “Metalearners for estimating heterogeneous treatment effects using machine learning,” *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 10, pp. 4156–4165, 2019, doi: 10.1073/pnas.1804597116.

J. Wang, J. Liu, Y. Ge, and B. Jiang, “Risk-aware ensemble learning for credit default prediction under economic uncertainty,” World Journal of Economics and Business Research, vol. 1, no. 1, pp. 18–32, 2023.

E. Pednault, N. Abe, and B. Zadrozny, “Sequential cost-sensitive decision making with reinforcement learning,” in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 259–268, doi: 10.1145/775047.775086.

G. Theocharous, P. S. Thomas, and M. Ghavamzadeh, “Personalized ad recommendation systems for life-time value optimization with guarantees,” in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 1806–1812.

J. Wang, Y. Tan, B. Jiang, B. Wu, and W. Liu, “Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies,” *Symmetry*, vol. 17, no. 4, Art. no. 610, 2025, doi: 10.3390/sym17040610.

T. Yu, G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn, and T. Ma, “MOPO: Model-based offline policy optimization,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 14129–14142.

S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline reinforcement learning: Tutorial, review, and perspectives on open problems,” arXiv:2005.01643, 2020.

A. Kumar, A. Zhou, G. Tucker, and S. Levine, “Conservative Q-learning for offline reinforcement learning,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1179–1191.

S. Fujimoto, D. Meger, and D. Precup, “Off-policy deep reinforcement learning without exploration,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, ser. Proc. Mach. Learn. Res., vol. 97, 2019, pp. 2052–2062.

Y. Jin, Z. Yang, and Z. Wang, “Is pessimism provably efficient for offline RL?” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, ser. Proc. Mach. Learn. Res., vol. 139, 2021, pp. 5084–5096.

F. Zhou, “Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders,” Innovation and Technology Studies, vol. 3, no. 1, pp. 27–41, 2026.

Y. Zhao, X. Fang, and D. Simchi-Levi, “Uplift modeling with multiple treatments and general response types,” in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2017, pp. 588–596, doi: 10.1137/1.9781611974973.66.

X. Nie and S. Wager, “Quasi-oracle estimation of heterogeneous treatment effects,” *Biometrika*, vol. 108, no. 2, pp. 299–319, 2021, doi: 10.1093/biomet/asaa076.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, “Double/debiased machine learning for treatment and structural parameters,” *Econometrics J.*, vol. 21, no. 1, pp. C1–C68, 2018, doi: 10.1111/ectj.12097.

B. Chakraborty and S. A. Murphy, “Dynamic treatment regimes,” *Annu. Rev. Statist. Appl.*, vol. 1, pp. 447–464, 2014, doi: 10.1146/annurev-statistics-022513-115553.

X. Zhang, K. Wang, Z. Wang, B. Du, S. Zhao, R. Wu, X. Shen, T. Lv, and C. Fan, “Temporal uplift modeling for online marketing,” in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2024, pp. 6247–6256, doi: 10.1145/3637528.3671560.

J. Chen, X. Han, X. Zhang, and Y. Yang, “Exact concept-level explanations for early detection of financial misstatements,” Eurasia Journal of Science and Technology, vol. 5, no. 2, pp. 34–48, 2023.

D. Ernst, P. Geurts, and L. Wehenkel, “Tree-based batch mode reinforcement learning,” *J. Mach. Learn. Res.*, vol. 6, pp. 503–556, 2005.

S. Fujimoto, E. Conti, M. Ghavamzadeh, and J. Pineau, “Benchmarking batch deep reinforcement learning algorithms,” arXiv:1910.01708, 2019.

Z. Deng, J. Jiang, G. Long, and C. Zhang, “Causal reinforcement learning: A survey,” arXiv:2307.01452, 2023.

K. Battocchi, E. Dillon, M. Hei, G. Lewis, P. Oka, M. Oprescu, and V. Syrgkanis, “EconML: A Python package for ML-based heterogeneous treatment effects estimation,” 2019. [Online]. Available: https://github.com/py-why/EconML

H. van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100, doi: 10.1609/aaai.v30i1.10295.

K. Hillstrom, “The MineThatData e-mail analytics and data mining challenge,” MineThatData blog, Mar. 2008. [Online]. Available: https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

D. Teng, “TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows,” World Journal of Engineering Research, vol. 3, no. 6, pp. 55–66, 2025.

T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

G. Wang, P. Li, E. Feng, A. Shen, and M. Rhee, “A Dual-Process Memory-Augmented Large Language Model Agent for Explainable Sequential Recommendation,” IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3741894.

P. Li, M. Bian, and Y. Lin, “Carbon-Aware Offline Reinforcement Learning for Joint Inventory Replenishment and Transportation Scheduling in Multi-Echelon Supply Chains,” Highlights in Business, Economics and Management, vol. 69, pp. 251–262, 2026.