Decision-Making Under Selective Labels: Optimal Finite-Domain Policies and Beyond

Dennis Wei [1]

Abstract

Selective labels are a common feature of high-stakes decision-making applications, referring to the lack of observed outcomes under one of the possible decisions. This paper studies the learning of decision policies in the face of selective labels, in an online setting that balances learning costs against future utility. In the homogeneous case in which individuals' features are disregarded, the optimal decision policy is shown to be a threshold policy. The threshold becomes more stringent as more labels are collected; the rate at which this occurs is characterized. In the case of features drawn from a finite domain, the optimal policy consists of multiple homogeneous policies in parallel. For the general infinite-domain case, the homogeneous policy is extended by using a probabilistic classifier and bootstrapping to provide its inputs. In experiments on synthetic and real data, the proposed policies achieve consistently superior utility with no parameter tuning in the finite-domain case and lower parameter sensitivity in the general case.

1. Introduction

The problem of *selective labels* is common to many high-stakes decision-making scenarios affecting human subjects. In these scenarios, individuals receive binary decisions, which will be referred to generically as acceptance or rejection. If the decision is to accept, then an outcome label is observed, which determines the utility of the decision. However if the decision is to reject, no outcome is observed. In lending for example, the decision is whether to offer or deny the loan, and the outcome of repayment or default is observed only if the loan is made. In pre-trial bail decisions, the outcome is whether a defendant returns to court without committing another offense, but there is no opportunity to observe it if bail is denied. In hiring, a candidate's job performance is observed only if they are hired.

The prevalence and challenges of selective labels were recently emphasized by Lakkaraju et al. (2017), who studied the evaluation of machine learning models in comparison to human decision-makers using data labelled selectively by the human decisions themselves. The subject of the present paper is the *learning* of decision policies in the face of selective labels. This problem was addressed indirectly by De-Arteaga et al. (2018), who proposed label imputation in regions of high human confidence, and more deeply by Kilbertus et al. (2020). In the latter paper, the goal is to maximize expected utility (possibly including a fairness penalty) over a held-out population, given data and labels collected selectively by a suboptimal existing policy. Kilbertus et al. (2020) showed that an existing policy that is deterministic, commonly achieved by thresholding the output of a predictive model, may condemn future policies to suboptimality. However, if the existing policy is stochastic and "exploring", then the optimal policy can be learned and a stochastic gradient ascent algorithm is proposed to do so.

This paper studies an online formulation of the selective labels problem, presented in Section 2, that accounts for the costs of decisions taken during learning and seeks to maximize discounted total reward. This contrasts with Kilbertus et al. (2020) where learning costs do not enter into the objective of held-out utility. Also unlike Kilbertus et al. (2020), there is no need for labelled data from an existing exploring policy. The online formulation brings the problem closer to one of contextual bandits, with which comparisons are made throughout the paper.

The approach taken herein is to first solve a simpler special case and then explore the extent to which this solution can generalize. Specifically, in Section 3, it is assumed that individuals are drawn from a homogeneous population, without features to distinguish them. By formulating the problem as a partially observable Markov decision process (POMDP) and applying dynamic programming, the optimal acceptance policy is shown to be a threshold policy on the estimated probability of success. Properties of the optimal policy are derived. These show that the policy becomes more stringent (i.e., the rejection set grows) as more observa-

[1] IBM Research, Yorktown Heights, NY, USA. Correspondence to: <dwei@us.ibm.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

tions are collected, which is reminiscent of upper confidence bound (UCB) policies (Auer et al., 2002; Chu et al., 2011; Abbasi-Yadkori et al., 2011). The rate of convergence of the decision threshold is characterized.

Generalizing from the homogeneous case to one with features X, Section 4 shows that the optimal decision policy for any finite feature domain $\mathcal{X}$ consists of multiple optimal homogeneous policies in parallel, one for each $x \in \mathcal{X}$, and each with an effective discount factor that depends on the probability distribution of X. For infinite and continuous domains, Section 5 proposes to leverage the optimal homogeneous policy, using a probabilistic classifier (e.g. logistic regression) and bootstrap estimates of uncertainty to supply the inputs required by the homogeneous policy.

The proposed policies are evaluated in experiments (reported in Section 6) on synthetic data and two real-world datasets featuring high-stakes decisions. Several conventional, selective labels, and contextual bandit baselines are used for comparison; efforts are made to re-implement or adapt some of these. In the finite-domain case, while one of the baselines can achieve optimal utility, the advantage of the optimal policy of Section 4 is that it does so without parameter tuning. In the general case, the extended homogeneous policy of Section 5 exhibits the highest utility and lower parameter sensitivity than the next best alternatives.

Other related work In addition to contextual bandits (Bietti et al., 2020; Foster et al., 2018; Agarwal et al., 2014; Joseph et al., 2016), the selective labels problem is related to policy learning (Dudík et al., 2011; Swaminathan & Joachims, 2015; Athey & Wager, 2017; Kallus, 2018) and causal inference (Hernán & Robins, 2020) in that only the outcome resulting from the selected action is observed. It is distinguished by there being no observation at all in the case of rejection. Notwithstanding this difference, it is possible to view the online formulation considered herein as a simpler special kind of bandit problem, as noted in Section 2. This simplicity makes it amenable to an optimal dynamic programming approach as opted for in this paper.

Limited feedback phenomena similar to selective labels have been considered in the literature on social aspects of ML, specifically as they relate to fairness and performance evaluation (Kallus & Zhou, 2018; Coston et al., 2020) and applications such as predictive policing (Lum & Isaac, 2016; Ensign et al., 2018). Notably, Bechavod et al. (2019) study a similar selective labels problem and the effect of group fairness constraints on regret. These problems, in which algorithm-driven decisions affect subsequent data observation, fit into a larger and growing literature on dynamics induced by the deployment of ML models (Liu et al., 2018; Hashimoto et al., 2018; Hu & Chen, 2018; Mouzannar et al., 2019; Heidari et al., 2019; Zhang et al., 2019; Perdomo et al., 2020; Creager et al., 2020; Rosenfeld et al., 2020; Tsirtsis & Gomez-Rodriguez, 2020; Zhang et al., 2020). One distinction is that limited feedback problems such as selective labels are present independent of whether and how humans respond to ML decisions.

2. Problem Formulation

The selective labels problem studied in this paper is as follows: Individuals $i = 0, 1, \dots$ arrive sequentially with features $x_i \in \mathcal{X}$. A decision of *accept* ($a_i = 1$) or *reject* ($a_i = 0$) is made based on each individual's x_i according to a *decision policy* $\Pi : \mathcal{X} \mapsto [0, 1]$, where $\Pi(x) = \Pr(A = 1 \mid x)$ is the probability of acceptance. The policy is thus permitted to be stochastic, although it will be seen that this is not needed in some cases. If the decision is to accept, then a binary outcome y_i is observed, with $y_i = 1$ representing *success* and $y_i = 0$ *failure*. If the decision is to reject, then no outcome is observed, hence the term *selective labels*. Individuals' features and outcomes are independently and identically distributed according to a joint distribution $p(x, y) = p(y \mid x)p(x)$.

Decisions and outcomes incur rewards according to $a_i(y_i - c)$ for $c \in (0, 1)$, following the formulation of Kilbertus et al. (2020); Corbett-Davies et al. (2017), i.e., a reward of $1 - c$ if acceptance leads to success, $-c$ if acceptance leads to failure, and 0 if the individual is rejected. The assumptions underlying this formulation deserve further comment. As noted by Kilbertus et al. (2020), the cost of rejection, whose general form is $(1 - a_i)g(y_i)$, is unobservable due to the lack of labels (although rejection is presumably negative for the individual). It is assumed therefore that g is constant, the reward from success is greater than g, and the reward (i.e. cost) from failure is less than g. Domain knowledge can inform the reward/cost of success/failure relative to rejection. For example in lending, the decision-maker's (lender's) rewards are fairly clear: interest earned in the case of success (repayment), loss of principal (or some expected fraction thereof) in the case of failure (default), and little to no cost for rejection. The individual's rewards may also be taken into account although harder to quantify, for example accomplishing the objective of the loan (e.g. owning a home) or damage to creditworthiness from a default (Liu et al., 2018). It might even be possible to learn the cost of rejection through an alternative feedback mechanism. For example, a lender could follow up with a subset of its rejected applicants to understand the impact on their lives. In any case, once the three reward values are determined, they can then be linearly transformed to $1 - c$, $-c$, and 0 without loss of generality.

The objective of *utility* is quantified by the expectation of the discounted infinite sum of rewards,

$$\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i a_i(y_i-c)\right]=\mathbb{E}\left[\sum_{i=0}^{\infty}\gamma^i\Pi(x_i)(\rho(x_i)-c)\right] \tag{1}$$

for some discount factor $\gamma < 1$, where we have defined the conditional *success probability* $\rho(x) := p(Y = 1 \mid x)$. The right-hand side of (1) results from taking the conditional expectation given x_i, leaving an expectation over $x_i \sim p(x)$. The right-hand expectation indicates that the problem of determining policy $\Pi(x)$ can be decomposed (at least conceptually) over values of X. This is clearest in the case of a discrete domain $\mathcal{X}$, for which the expectation is a sum, weighted by $p(x)$. The decomposition motivates the study of a simpler problem in which x is fixed or dropped, resulting in a homogeneous population. This "homogeneous" problem is the subject of Section 3. We then consider how to leverage the solution to the homogeneous problem in later sections.

It is also possible to treat the selective labels problem as a special contextual bandit problem with two possible actions (accept/reject), where the reward from rejection is furthermore taken to be zero as discussed above. The following sections show that the approach of starting with the homogeneous setting allows the optimal policy to be determined in the case of finite $\mathcal{X}$. An empirical comparison with contextual bandit algorithms is reported in Section 6. It should also be noted that while the cost of rejection is assumed to be a constant, the *relative* utility of rejection is $c - \rho(x_i)$ from (1), which is not constant and requires estimation of $\rho(x)$.

Kilbertus et al. (2020) formulate a fairness objective in addition to utility but this will not be considered herein.

3. The Homogeneous Case

In the homogeneous case with no features X, the success probability reduces to a single parameter $\rho := p(y_i = 1)$. If ρ is known, then the policy that maximizes (1) is immediate: $\Pi^*(\rho) = \mathbb{1}(\rho > c)$, where $\mathbb{1}(\cdot)$ is the indicator function that yields 1 when its argument is true. The optimal utility is

$$V^*(\rho,\infty)=\sum_{i=0}^{\infty}\gamma^i\max\{\rho-c,0\}=\frac{\max\{\rho-c,0\}}{1-\gamma}. \tag{2}$$

As will be explained more fully below, the ∞ in $V^*(\rho,\infty)$ denotes exact knowledge of ρ, i.e. from an infinite sample.

The challenge of course is that ρ is not known but must be learned as decisions are made. The approach taken herein is to regard the case of known ρ as a Markov decision process (MDP) with state ρ and no dynamics (i.e. $\rho_{i+1} = \rho_i$). The case of unknown ρ is then treated as the corresponding partially observable MDP (POMDP) using a *belief state* for ρ (Bertsekas, 2005, Sec. 5.4).

To define the belief state, a beta distribution prior is placed on ρ: $\rho_0 \sim B(\sigma_0, \nu_0 - \sigma_0)$, where the shape parameters $\alpha = \sigma_0$, $\beta = \nu_0 - \sigma_0$ are expressed in terms of a number σ_0 of "pseudo-successes" in ν_0 "pseudo-observations". Since ρ is the parameter of a Bernoulli random variable, the beta distribution is a conjugate prior. It follows that the posterior distribution of ρ before individual i arrives, given $\nu'_i = \sum_{j=0}^{i-1} a_j$ outcomes and $\sigma'_i = \sum_{j=0}^{i-1} a_j y_j$ successes observed thus far, is also beta, $\rho_i \sim B(\sigma_i, \nu_i - \sigma_i)$, with $\sigma_i = \sigma_0 + \sigma'_i$ and $\nu_i = \nu_0 + \nu'_i$. Thus we define the pair $\mu_i := \sigma_i/\nu_i = \mathbb{E}[\rho_i]$ and ν_i as the belief state for ρ, equivalently using the mean μ_i in place of σ_i. The acceptance policy is also made a function of the belief state, $\Pi(\mu_i, \nu_i)$.

The initial state (μ_0, ν_0), i.e. the parameters of the prior, can be chosen based on an initial belief about ρ. This choice is clearer when outcome data has already been collected by an existing policy, in which case ν_0 can be the number of outcomes observed and μ_0 the empirical mean.

Define $V^\Pi(\mu, \nu)$ to be the value function at state (μ, ν) under policy Π, i.e., the expected discounted sum of rewards from following Π starting from state (μ, ν). The index i is dropped henceforth because the dependence is on (μ, ν), irrespective of the number of rounds needed to attain this state. In Appendix A.1, the dynamic programming recursion that governs $V^\Pi(\mu, \nu)$ is derived. By optimizing this recursion with respect to the acceptance probabilities $\Pi(\mu, \nu)$, we obtain the following result.

Theorem 1. *For the homogeneous selective labels problem, the optimal acceptance policy that maximizes discounted total reward* (1) *is a threshold policy:* $\Pi^*(\mu, \nu) = \mathbb{1}(V^*(\mu, \nu) > 0)$*, where the optimal value function* $V^*(\mu, \nu)$ *satisfies the recursion*

$$V^*(\mu,\nu)=\max\left\{\mu-c+\gamma\left[\mu V^*\left(\frac{\mu\nu+1}{\nu+1},\nu+1\right)\right.\right. \\ \left.\left.+(1-\mu)V^*\left(\frac{\mu\nu}{\nu+1},\nu+1\right)\right],0\right\}. \tag{3}$$

Theorem 1 shows that the optimal homogeneous policy does not require stochasticity. It also shows that the problem is one of *optimal stopping* (Bertsekas, 2005, Sec. 4.4): in each state (μ, ν), there is the option $(\Pi(\mu, \nu) = 0)$ to stop accepting and thus stop observing, which freezes the state at (μ, ν) thereafter with zero reward. The optimal policy is thus characterized by the *stopping* or *rejection set*, the set of (μ, ν) at which it is optimal to stop because the expected reward from continuing is negative.

In the limiting case as $\nu \to \infty$, $V^*(\mu, \nu)$ and $\Pi^*(\mu, \nu)$ are known explicitly. This is because the mean μ converges to the true success probability ρ, by the law of large numbers. We therefore have $\Pi^*(\mu, \infty) = \mathbb{1}(\mu > c)$ and $V^*(\mu, \infty)$ as given in (2), explaining the previous notation. The corresponding stopping set is the interval $[0, c]$.

Connection to one-armed bandit The above formulation and dynamic programming solution are related to the "one-armed bandit" construction of Weber (1992) and its corresponding Gittins index. Specifically, upon defining belief state (μ, ν), the homogeneous problem conforms to the formulation of Weber (1992): (μ, ν) is the state ($x_j(t)$ in Weber's notation), rewards are a function of this state, and (μ, ν) evolve in a Markov fashion upon each acceptance. One might expect therefore that the optimal homogeneous policy of Theorem 1 is equivalent to the Gittins index policy, and indeed this is the case. For a "one-armed bandit" where the cost of the "reject" arm is taken to be zero, it suffices to determine whether the expected discounted total reward that appears in the Gittins index, $\sup_\tau \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t R_j(x_j(t)) \mid x_j(0) = x\right]$, is positive. Here the supremum is taken over stopping times τ. The proposed dynamic programming approach summarized by Theorem 1 can be seen as an explicit way of computing the supremum (which Weber does not discuss): we either stop at $\tau = 0$, or continue so that τ is at least 1 and consider the same stopping question for the possible next states $x_j(1)$, weighted appropriately.

Approximation of optimal policy For finite ν, a natural way of approximating $V^*(\mu, \nu)$ is as follows: Choose a large integer N, which will also index the approximation, $V^N(\mu, \nu)$, and set $V^N(\mu, N+1) = V^*(\mu, \infty)$, the infinite-sample value function (2). Then use (3) with V^N in place of V^* to recursively compute $V^N(\mu, \nu)$ for $\nu = N, N-1, \ldots$. The corresponding policy is $\Pi^N(\mu, \nu) = \mathbb{1}(V^N(\mu, \nu) > 0)$. Note that (3) is valid for all $\mu \in [0, 1]$, not just integer multiples of $1/\nu$; this can be seen by allowing the initial parameter σ_0 to range over real values.

Figure 1 plots the result of the above computation for $N = 1000$, $c = 0.8$, and $\gamma = 0.99$ (a second example is in Appendix B). The plot suggests that $V^N(\mu, \nu) \geq V^N(\mu, \nu+1)$ and that $V^N(\mu, \nu)$ is a non-decreasing convex function of μ for all ν. It also shows that $V^N(\mu, \nu)$ is quite close to $V^N(\mu, 1001) = V^*(\mu, \infty)$ for large $\nu > 100$.

Properties of optimal policy The properties suggested by Figure 1 do in fact hold generally (all proofs in Appendix A).

Proposition 2. *The optimal value function $V^*(\mu, \nu)$ is non-decreasing and convex in μ for all ν.*

Proposition 3. *The optimal value function $V^*(\mu, \nu)$ is non-increasing in ν, i.e. $V^*(\mu, \nu) \geq V^*(\mu, \nu+1)\ \forall\ \mu \in [0, 1]$.*

Other optimal stopping problems are known to have similar monotonicity and convexity properties (Bertsekas, 2005).

Monotonicity in both μ and ν implies that the stopping set at sample size ν, $\{\mu : V^*(\mu, \nu) \leq 0\}$, is an interval $[0, c_\nu]$ that grows as ν increases, $c_\nu \leq c_{\nu+1} \leq \cdots \leq c$. In other words, the acceptance policy is more lenient in early stages and gradually approaches the policy for known ρ. The following result bounds the difference $c - c_\nu$.

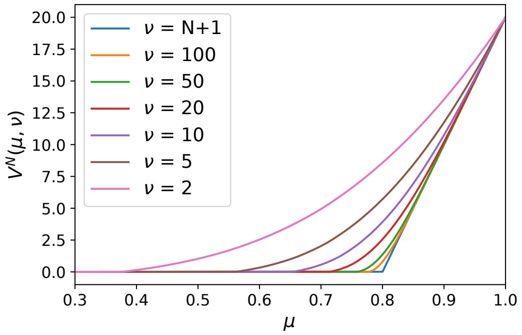


Figure 1. Optimal value function approximations $V^N(\mu, \nu)$ for $N = 1000$, $c = 0.8$, and $\gamma = 0.99$.

Proposition 4. *The difference between the acceptance threshold c_ν for sample size ν and the infinite-sample threshold c is bounded as follows:*

$$\begin{aligned} c - c_\nu &\leq \frac{\gamma \cdot {}_2F_1(1, \nu; \nu+2; \gamma)}{\nu + 1 - \gamma \cdot {}_2F_1(1, \nu; \nu+2; \gamma)}(1 - c) \\ &\leq \frac{\gamma \min\{1/(1-\gamma), \nu+1\}}{\nu + 1 - \gamma \min\{1/(1-\gamma), \nu+1\}}(1 - c), \end{aligned}$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function.

From the second, looser upper bound above, it can be seen that for $\nu > 1/(1-\gamma)$, $c - c_\nu$ decays as $O(1/\nu)$. It is interesting to compare this behaviour to UCB policies (Auer et al., 2002; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Filippi et al., 2010; Li et al., 2017). An acceptance threshold c_ν is equivalent to adding a margin $c - c_\nu$ to the mean μ (i.e., yielding a UCB) and comparing with c. Typically however, confidence intervals are proportional to the standard deviation and scale as $1/\sqrt{\nu}$, as is the case for a beta or binomial distribution. The $1/\nu$ rate implied by Proposition 4 for large ν is thus faster.

The analysis that leads to Proposition 4 can be extended to also provide bounds on the approximation $V^N(\mu, \nu)$.

Proposition 5. *For $\nu = N+1, N, \ldots$ and all $\mu \in [0, 1]$,*

$$\begin{aligned} 0 &\leq V^*(\mu, \nu) - V^N(\mu, \nu) \\ &\leq \frac{\gamma^{N+2-\nu} {}_2F_1(1, N+1; N+3; \gamma)}{N+2} V^*(1, \nu). \end{aligned}$$

Similar to Proposition 4, for $N > 1/(1-\gamma)$, the approximation error decays as $1/N$ (γ^N/N for fixed ν).

In Appendix C, the case of *undiscounted average* reward (in contrast to (1)) over an infinite horizon is also analyzed. There the optimal policy is found to have positive acceptance probability regardless of the belief state. This is reminiscent of the exploring policies of Kilbertus et al. (2020) and contrasts with the case of discounted total reward (1).

4. The Finite-Domain Case

In this section, we move from the homogeneous case to one where features X are present and take a finite number of values. As discussed in Section 2, the decomposability of (1) into a sum over $x \in \mathcal{X}$ implies that the optimal decision policy consists of multiple optimal homogeneous policies in parallel, one for each $x \in \mathcal{X}$. Accordingly, a beta distribution is now posited for each conditional success probability $\rho(x)$, parametrized by state variables $\mu(x)$ and $\nu(x)$: $\nu(x)$ is the number of acceptances with feature value x, plus pseudo-counts from the prior on $\rho(x)$, while $\mu(x)$ is the fraction of successes among acceptances at x, again accounting for prior pseudo-counts.

The difference with respect to the homogeneous case is that the *effective discount factor* seen at each value x is not equal to the γ in (1) but depends on x as follows:

$$\bar{\gamma}(x) = \frac{\gamma p(x)}{1 - \gamma(1 - p(x))}. \tag{4}$$

Intuitively, the effective discount factor $\bar{\gamma}(x)$ arises because successive arrivals of individuals with value x are separated not by one time unit but by a random, geometrically distributed time that depends on $p(x)$.

Denote by $\Pi^*(\mu, \nu; \gamma)$ the optimal homogeneous policy that uses discount factor γ in (3) and (2), and $V^*(\mu, \nu; \gamma)$ the corresponding optimal value function. Then the optimal finite-domain policy can be stated as follows.

Theorem 6. *Assume that the features X have finite cardinality, $|\mathcal{X}| < \infty$. Then the optimal acceptance policy is to use optimal homogeneous policies $\Pi^*\big(\mu(x), \nu(x); \bar{\gamma}(x)\big)$ independently for each $x \in \mathcal{X}$, where $\bar{\gamma}(x)$ is the effective discount factor in* (4)*.*

Appendix A.5 provides a derivation of (4) to prove Theorem 6.

Computing the effective discount factors (4) requires knowledge of the distribution $p(x)$. In the usual case where $p(x)$ is not known, $\bar{\gamma}(x)$ may be estimated empirically. Denoting by $I_1, I_2, \ldots, I_m$ the inter-arrival times observed at x thus far, the estimated effective discount factor is

$$\hat{\bar{\gamma}}(x) = \frac{1}{m} \sum_{j=1}^{m} \gamma^{I_j}. \tag{5}$$

5. The General Case

We now consider the general case in which the features X are continuous or X is still discrete but the cardinality of $\mathcal{X}$ is large. In these cases, it is no longer possible or statistically reliable to represent the state of knowledge by counts of acceptances and successes.

In this paper, we investigate the extent to which the optimal homogeneous policy can be successfully carried over to the general setting. The development of more involved policies is left to future work. The continued use of the homogeneous policy is motivated by two reasons: first, its optimality for finite domains, which might be used to approximate an infinite or continuous domain, and second, the ease of computing the approximation $V^N(\mu, \nu)$ (taking milliseconds on a MacBook Pro for $N = 1000$ in Figure 1).

The application of the homogeneous policy, i.e. $\Pi^*(\mu(x), \nu(x); \bar{\gamma})$, requires three inputs: (1) The mean parameter $\mu(x)$ of the beta distribution assumed for the conditional success probability $\rho(x)$; (2) The sample size parameter $\nu(x)$ of $\rho(x)$; (3) The discount factor $\bar{\gamma}$ that determines the trade-off between exploration and exploitation. These are discussed in turn below.

Conditional mean $\mu(x)$ With $\rho(x)$ assumed to be random, we have $\Pr(Y = 1 \mid x) = \mathbb{E}[\rho(x)] = \mu(x)$. Estimation of $\mu(x)$ is equivalent therefore to the standard probabilistic classification problem of approximating $\Pr(Y = 1 \mid x)$. This may be accomplished by training a model $\hat{\mu}(x)$ to minimize log loss (e.g. logistic regression) on accepted individuals, i.e., those for which Y labels are available.

Conditional sample size $\nu(x)$ In the finite case, $\nu(x)$ is the sample size parameter of the beta posterior for $\rho(x)$. It is thus equal (possibly with a constant offset) to the number of labels observed for x and may be seen as a measure of confidence in the conditional mean $\mu(x)$. This suggests measuring confidence in the predictions of the model $\hat{\mu}(x)$ used to approximate $\mu(x)$ in the more general case.

The above idea is realized herein via bootstrap sampling. For a given x, let $\hat{\mu}_1(x), \ldots, \hat{\mu}_K(x)$ be K estimates of the conditional mean from K models trained on bootstrap resamples of the labelled population. (The "master" model $\hat{\mu}(x)$ from above, trained on the full labelled population, is separately maintained.) This set of K estimates is regarded as an approximation to the posterior distribution of $\rho(x)$, in a similar spirit as in Eckles & Kaptein (2014); Osband & Roy (2015). Fitting a beta distribution to $\hat{\mu}_1(x), \ldots, \hat{\mu}_K(x)$, the parameter $\nu(x)$ is estimated by the method of moments as

$$\hat{\nu}(x) = \frac{\bar{\hat{\mu}}(x)(1 - \bar{\hat{\mu}}(x))}{\operatorname{var}(\hat{\mu}(x))} - 1, \tag{6}$$

where $\bar{\hat{\mu}}(x)$ and $\operatorname{var}(\hat{\mu}(x))$ are the sample mean and sample variance of $\hat{\mu}_1(x), \ldots, \hat{\mu}_K(x)$.

Note that the above methods of estimating $\mu(x)$ and $\nu(x)$ do not require specification of prior parameters $\mu_0(x)$, $\nu_0(x)$, unlike in the homogeneous and finite-domain cases.

Discount factor $\bar{\gamma}$ In the homogeneous and finite-domain cases, the effective discount factor $\bar{\gamma}(x)$ used by the policy is either equal to the given discount factor γ in (1) or can be determined from the probability distribution of X. For the general case, $\bar{\gamma}(x) \equiv \bar{\gamma}$ is left as a single free parameter of the policy (independent of x), to be tuned to balance exploration and exploitation. The results in Section 6 indicate that performance is relatively insensitive to $\bar{\gamma}$.

Intuitively, one expects good values for $\bar{\gamma}$ to lie below γ, as they do in the finite-domain case (4). The reason is that when the conditional mean $\mu(x)$ is estimated by a model, accepting an individual and observing a label at one value x also decreases uncertainty in $\mu(x)$ at other values x through an update to the model. In contrast, in the finite-domain case, $\rho(x)$ for $x \in \mathcal{X}$ are modelled independently. This indirect reduction of uncertainty (related to the intrinsic exploration of greedy policies analyzed by Bastani et al. (2020); Kannan et al. (2018)) reduces the need for exploration and favours smaller $\bar{\gamma}$. Further analysis and the possibility of setting $\bar{\gamma}$ automatically are left for future work.

Practical aspects The quantities $\hat{\mu}(x)$ and $\hat{\nu}(x)$ are ideally updated after each new observation. For computational efficiency, online learning and the online bootstrap (Eckles & Kaptein, 2014; Osband & Roy, 2015) are used to perform these updates. Specifically, this work makes use of the Vowpal Wabbit (VW) library[1] for online learning and the "double-or-nothing" bootstrap: for each of the K bootstrap samples, a new observation is added twice with probability $1/2$ or is not added. Other unit-mean distributions over non-negative integers (e.g. $\text{Poisson}(1)$ in Bietti et al. (2020)) could also be used.

The estimate $\hat{\nu}(x)$ (6) is generally not an integer. In this work, $\hat{\nu}(x)$ is simply rounded to the nearest integer and truncated if needed to $N+1$, the largest value in the approximation $V^N(\mu, N+1) = V^*(\mu, \infty)$. To handle real-valued $\hat{\mu}(x)$, recursion (3), which is valid for all $\mu \in [0, 1]$ as discussed in Section 3, is pre-computed on a dense grid of μ values and then linearly interpolated as needed.

6. Experiments

Experiments are conducted on synthetic data (Section 6.2) to evaluate the optimal finite-domain policy of Section 4, as well as on two real-world datasets with high-stakes decisions (Section 6.3) to evaluate the extended homogeneous policy of Section 5. In all cases, a selective labels problem is simulated from a labelled dataset of (x_i, y_i) pairs by presenting features x_i of individuals one by one and only revealing the outcome y_i to the algorithm if the decision is to accept. In addition, to provide an initial training (i.e. exploration) set, the first B_0 individuals are always accepted and their outcomes are observed. Notably, the rewards/costs incurred from collecting this training data are counted toward the total utility. The effects of varying B_0 are studied.

The proposed policies are compared to a conventional baseline, the selective-labels-specific method of Kilbertus et al. (2020), and contextual bandit algorithms, described in Section 6.1. For both computational efficiency and fair comparison, the supervised learning models upon which all of these policies rely are trained online using VW. For the finite-domain experiments in which modelling is not necessary, the Bayesian approach of Section 4 is used to update $\mu(x)$, $\nu(x)$ for all policies. Appendix D.2 provides more details.

6.1. Baselines

Greedy (G) This baseline represents the conventional approach of training a success probability model $\hat{\mu}(x)$ on the initial training set of size B_0, and then accepting and collecting labels from only individuals for whom the prediction $\hat{\mu}(x_i)$ exceeds the threshold c. The labels of accepted individuals are used to update the model. Since the policy $\mathbb{1}(\hat{\mu}(x_i) > c)$ maximizes the immediate expected reward, this baseline will be referred to as the greedy policy.

Consequential Learning (CL, CLVW) The CL algorithm (Kilbertus et al., 2020, Alg. 1) is re-implemented for the case of no fairness penalty ($\lambda = 0$) and policy updates after every acceptance/observation ($N = 1$). These settings bring it in line with other methods compared. Update equations are given in Appendix D.1.1. While the paper of Kilbertus et al. (2020) does link to a code repository, no code was available as of this writing. Furthermore, CL uses "plain" stochastic gradient updates, whereas VW uses a more sophisticated algorithm. For this reason, a VW version of CL (CL-VW) was also implemented, also described in Appendix D.1.1.

Contextual bandit algorithms As noted in Section 2, the selective labels problem can be treated as a contextual bandit problem. Accordingly, four representative contextual bandit algorithms are compared: ϵ-greedy (ϵG) (Langford & Zhang, 2008), bootstrap Thompson sampling/bagging (Eckles & Kaptein, 2014; Osband & Roy, 2015), Online Cover (Agarwal et al., 2014), and RegCB (Foster et al., 2018), which is a generalization of LinUCB (Chu et al., 2011; Abbasi-Yadkori et al., 2011). These are chosen because they are practical algorithms extensively evaluated in a recent contextual bandit "bake-off" (Bietti et al., 2020) and are implemented in VW. More specifically, based on the recommendations of Bietti et al. (2020), the chosen variants are greedy bagging (B-g), Online Cover with no uniform

[1] https://vowpalwabbit.org

exploration (C-nu), and optimistic RegCB (R-o). Parameter settings and tuning are discussed in Appendix D.3.

The selective labels problem herein differs from a two-arm contextual bandit in that the cost of rejection is assumed to be zero, an assumption that is not used by the four algorithms above. In an attempt to mitigate this possible disadvantage, each algorithm was given the option of observing one pass through the entire dataset in which all individuals are rejected with a cost of zero. This did not appear to improve performance appreciably, possibly because the reward estimators in VW are already initialized at zero.

RegCB-Optimistic (R-o, R-osl) The R-o algorithm is of particular interest for two reasons. First, it is a UCB policy with similar structure to the optimal homogeneous policy, as discussed in Section 3. Second, it performed best overall in the bake-off of Bietti et al. (2020). In the experiments herein however, the VW implementation of R-o performed less well (see Appendix D.5). The likely reason is that it does not take advantage of the zero-cost assumption for rejection, despite the rejection pass through the data mentioned above.

To improve the performance of R-o, a specialized version that does exploit the zero-cost assumption was implemented, referred to as R-osl. R-o makes the decision with the highest UCB on its expected reward. Since rejection is assumed to have zero cost while the expected reward of acceptance is $\hat{\mu}(x) - c$, this reduces to determining whether the UCB on $\hat{\mu}(x)$ exceeds c. More details are in Appendix D.1.2.

6.2. Finite-Domain Experiments

The experiments on synthetic data address the finite-domain setting and focus on two questions: (1) the effect of having to estimate the effective discount factors $\bar{\gamma}(x)$ on the performance of the optimal policy, and (2) comparison of the optimal policy to various baselines.

Synthetic data generation Given a cardinality $|\mathcal{X}|$, the probability distribution $p(x)$ is sampled from the flat Dirichlet (i.e. uniform) distribution over the $|\mathcal{X}| - 1$-dimensional simplex. Success probabilities $\rho(x) = \Pr(Y = 1 \mid x)$ are sampled from the uniform distribution over $[0, 1]$, independently for $x = 0, \dots, |\mathcal{X}| - 1$. Then T pairs (x_i, y_i), $i = 0, \dots, T - 1$ are drawn from the joint distribution of (X, Y). This generation procedure is repeated 1000 times for each cardinality $|\mathcal{X}|$ and threshold c. Means and standard errors in the means are computed from these repetitions.

For evaluation, rewards are summed using the discount factor $\gamma = 0.999$. The number of rounds T is set to $5/(1-\gamma)$ so that the sum of truncated discount weights, $\sum_{t=T}^{\infty} \gamma^t$, is less than 1% of the total sum $\sum_{t=0}^{\infty} \gamma^t$.

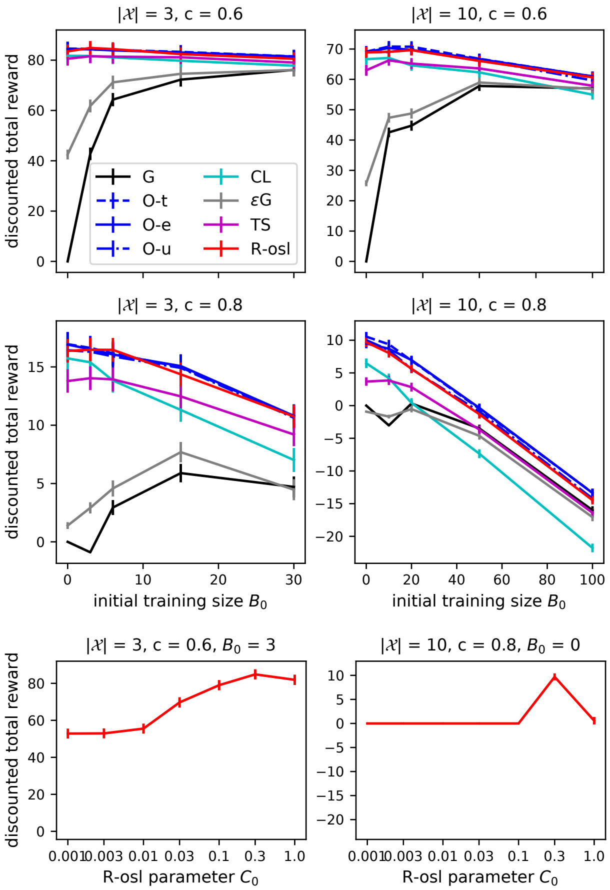


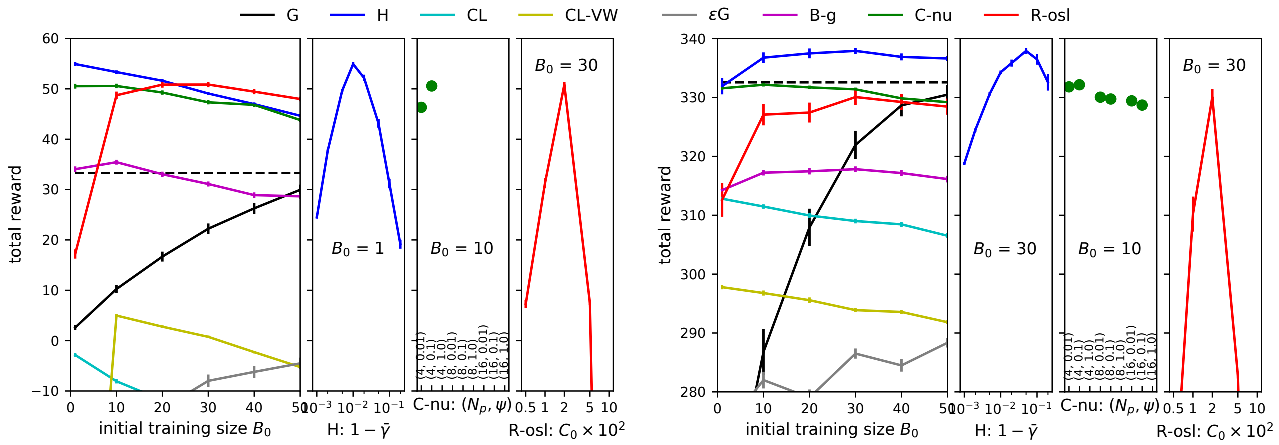
Figure 2. Discounted total rewards (discount factor $\gamma = 0.999$) on finite domains $\mathcal{X}$.

Homogeneous policy variants Three variants of the optimal/homogeneous policy are compared. The first (abbreviated O-t, 't' for "true") is given access to $p(x)$ and computes the effective discount factors $\bar{\gamma}(x)$ using (4). The second, more realistic variant (O-e, "estimate") is not given $p(x)$ and instead estimates $\bar{\gamma}(x)$ using (5). The third (O-u, "uniform") does not estimate $\bar{\gamma}(x)$, instead assuming a uniform distribution $p(x) = 1/|\mathcal{X}|$ and using that in (4).

Modifications to baselines The most noteworthy change is to bagging, which is an approximation of Thompson sampling (TS). The latter can be implemented directly in the finite-domain case. TS chooses acceptance with probability $\Pr(\rho_i(x) > c)$, i.e., the probability that the reward from acceptance is greater than zero. Other modifications are described in Appendix D.1.

Results Figure 2 shows the discounted total rewards achieved for $|\mathcal{X}| \in \{3, 10\}$ and $c \in \{0.6, 0.8\}$. Plots for additional $(|\mathcal{X}|, c)$ pairs are in Appendix D.5. In general, greater differences are seen as c varies compared to $|\mathcal{X}|$.

In the first two rows of Figure 2, the total reward is plotted as a function of the size B_0 of the initial training batch. For



(a) FICO, $c = 0.8$

(b) COMPAS, $c = 0.6$

Figure 3. Total rewards (discount factor $\gamma = 1$) on real-world datasets. The dashed black line indicates the maximum total reward achieved by the greedy policy (G) at $B_0 = 80$ for FICO and $B_0 = 90$ for COMPAS.

the greedy policy (G), the curves tend to be increasing. The initial increase from zero occurs because some training data is needed for the predicted success probabilities to reach the threshold c for accepting any individuals. Further increases are due to continued improvement in these estimates. ϵG behaves similarly. For the other policies, the curves generally decrease because effective exploration is already built in and additional training is not worth the cost.

As expected, the optimal/homogeneous policies perform the best. Among them, there is no discernible difference between O-e and O-t, and thus no apparent cost due to estimating $\bar{\gamma}(x)$. Surprisingly, O-u, which assumes a uniform distribution, is hardly worse.

R-osl attains essentially optimal performance in many cases, provided that its confidence parameter C_0 is tuned. The last row in Figure 2 shows that total rewards can be significantly worse if C_0 is not well chosen. The high potential of R-osl is explained by its similarity to the optimal policy, as discussed earlier. The difference is that the optimal policy does not require parameter tuning.

CL and TS both outperform G and are similar to each other because they are both stochastic policies, choosing acceptance with the probability that they believe it is better than rejection. CL however requires tuning of its learning rate parameter whereas TS does not. ϵG slightly outperforms G at lower B_0 values that are insufficient for G.

6.3. Real Data Experiments

We now turn to two real-world datasets, the FICO Challenge dataset (FICO, 2018) and the COMPAS recidivism dataset (Angwin et al., 2016), for evaluating the policy of Section 5. The former comes from home equity line of credit (HELOC) applications together with an outcome variable indicating whether the borrower satisfactorily repaid the line of credit. Acceptance corresponds to approving the borrower for the HELOC. The COMPAS dataset, also used by Kilbertus et al. (2020), contains demographics and criminal histories of offenders, a recidivism risk score produced by the COMPAS tool, and an outcome variable indicating whether the offender was re-arrested within two years. Acceptance corresponds to releasing an offender on bail.

Each dataset is randomly permuted 1000 times and means and standard errors are computed from these permutations. Pre-processing steps are detailed in Appendix D.4.

Results Figure 3 plots the total rewards attained on the FICO and COMPAS datasets with $c = 0.8$ and $c = 0.6$ respectively (the latter choice conforms with Kilbertus et al. (2020)). Total rewards are computed without discounting ($\gamma = 1$); Appendix D.5 provides similar plots for $\gamma = 0.9995$. The shapes of the curves as functions of the initial training size B_0 are generally the same as in Figure 2, although there are some additional algorithms that benefit from having $B_0 > 0$.

The highest utility is again achieved by a policy in the homogeneous family, namely the policy of Section 5 (labelled H). Next in line are R-osl and C-nu. All three policies have at least one algorithm-specific parameter (discount factor $\bar{\gamma}$ for H, C_0 for R-osl, (N_p, ψ) for C-nu) as well as the online learning rate α that are tuned. The middle panels in Figures 3a and 3b show that H is less sensitive to its parameter, achieving good performance over a relatively wide range of $\bar{\gamma}$. In contrast for R-osl, the utility is much lower for C_0 grid values not equal to the best one. C-nu can display extreme sensitivity: In Figure 3b, $\psi = 1$ drops the total reward to

~ 100 (outside the plotted range), while in Figure 3a, the values not shown are in the negative hundreds.

A major difference compared to the finite-domain case is that the greedy policy (G) is more competitive.[2] Indeed, in Figure 3, H is the only policy that consistently exceeds the largest total reward achieved by G (dashed black line) as B_0 is allowed to increase. Conversely, CL, CL-VW, and ϵG never exceed the maximum reward of G, suggesting that they over-explore using an accceptance rate that is too high. The VW variant CL-VW outperforms CL on FICO (Figure 3a) except at $B_0 = 1$, justifying the alternative implementation.

7. Discussion

Optimal decision policies were presented for homogeneous and finite-domain cases of the selective labels problem. An extension of these policies was proposed for the general infinite-domain case and was shown to outperform several baselines with less parameter sensitivity. The policies account for the cost of learning as they seek to maximize utility. In doing so, they make deterministic decisions and become more stringent as more labels are observed, similar to UCB policies. They thus avoid potential objections to making consequential decisions non-deterministically, as noted by Kilbertus et al. (2020). On the other hand, Proposition 4 suggests a kind of "sequence unfairness": early-arriving individuals are subject to a more lenient policy, enjoying the "benefit of the doubt" in their true success probability.

Limitations The experiments in Section 6 have the following limitations:

1. The FICO and COMPAS datasets are treated as samples from the uncensored joint distribution of X, Y. However, as noted by Kilbertus et al. (2020), these datasets likely suffer themselves from selective labels and the true joint distribution can only be inferred from real-world exploration. This limitation highlights the need for datasets that are realistic and that ideally do not suffer from selective labels, or suffer only mildly in a correctable way, to support further research.

2. For COMPAS in particular, the present work does not consider fairness, or the possibility of finer-grained decisions such as supervised/unconditional release. Given the history of the COMPAS dataset in particular and high-stakes decision-making domains in general, some may argue that fairness should always be a consideration.

3. It may not be feasible to update policies after every observation.

4. More seriously, it is assumed that outcomes are observed immediately following an acceptance decision. In reality, there is often a long delay (two years in the cases of FICO and COMPAS).

5. The results in Figure 3 and for some of the baselines in Figure 2 are optimized over each policy's parameter and the online learning rate. This optimistically represents the potential of each policy. It appears to be common practice in contextual bandit papers (Bietti et al., 2020; Foster et al., 2018), where parameter selection appears to be a common challenge. In their "bake-off", Bietti et al. (2020) took advantage of having a large number (200) of datasets to select good parameter values for use on unseen datasets.

Future work Some of the limitations above may be more readily addressable. Notably, incorporation of fairness (limitation 2) would be an important extension to the problem formulation herein, possibly along the lines of Kilbertus et al. (2020); Bechavod et al. (2019). Non-binary decisions (e.g. supervised/unconditional release) as well as non-binary outcomes are also of interest. The effect of limitation 3 could be simulated in future experiments. In addition, Section 5 leaves open the development of decision policies that more directly tackle the infinite-domain case where the conditional success probability must be modelled. One goal would be to avoid having to select the parameter $\bar{\gamma}$, which would partly address limitation 5.

Acknowledgements

I would like to thank Eric Mibuari for sparking my interest in the problem and for multiple interesting discussions (particularly on limitation 3), Blossom Metevier for a very useful pointer to Bietti et al. (2020) and the contextual bandit literature, Jiri Navratil and Prasanna Sattigeri for a discussion on confidence estimation, Andrea Simonetto for a discussion leading to the "sequence unfairness" point, and Kush Varshney for a discussion on datasets and general encouragement. I would also like to thank the anonymous reviewers for their thoughtful feedback, in particular contributions to the discussion in Section 7 and the connection to the "one-armed bandit" of Weber (1992).

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pp. 2312–2320, 2011. URL https://proceedings.

[2] This echoes a finding from Bietti et al. (2020). Appendix D.5 shows that increasing c to 0.8 on COMPAS results in G being the best policy, due to the increased cost of exploration and decreased reward of better learning.

neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, pp. II–1638—-II–1646, 2014.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 23 May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. dataset available at https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv.

Athey, S. and Wager, S. Policy learning with observational data, 2017. arXiv e-print https://arxiv.org/abs/1702.02896.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. URL https://doi.org/10.1023/A:1013689704352.

Bastani, H., Bayati, M., and Khosravi, K. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 2020. URL https://doi.org/10.1287/mnsc.2020.3605.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, Z. S. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper/2019/file/084afd913ab1e6ea58b8ca73f6cb41a6-Paper.pdf.

Bertsekas, D. P. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Belmont, MA, USA, 2005.

Bietti, A., Agarwal, A., and Langford, J. A contextual bandit bake-off, 2020. arXiv e-print https://arxiv.org/abs/1802.04064.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 11–13 Apr 2011. URL http://proceedings.mlr.press/v15/chu11a.html.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 797–806, August 2017. URL http://doi.acm.org/10.1145/3097983.3098095.

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 582–593, 2020. URL https://doi.org/10.1145/3351095.3372851.

Creager, E., Madras, D., Pitassi, T., and Zemel, R. Causal modeling for fairness in dynamical systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

De-Arteaga, M., Dubrawski, A., and Chouldechova, A. Learning under selective labels in the presence of expert consistency. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018. URL https://arxiv.org/abs/1807.00905.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML)*, pp. 1097—1104, 2011.

Eckles, D. and Kaptein, M. Thompson sampling with the online bootstrap, 2014. arXiv e-print https://arxiv.org/abs/1410.4009.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAccT)*, pp. 160–171, 23–24 Feb 2018. URL http://proceedings.mlr.press/v81/ensign18a.html.

FICO. Explainable machine learning challenge, 2018. URL https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, pp. 586–594, 2010. URL https://proceedings.neurips.cc/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Foster, D., Agarwal, A., Dudik, M., Luo, H., and Schapire, R. Practical contextual bandits with regression oracles.

In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pp. 1539–1548, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/foster18a.html.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pp. 1929–1938, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/hashimoto18a.html.

Heidari, H., Nanda, V., and Gummadi, K. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 2692–2701, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/heidari19a.html.

Hernán, M. A. and Robins, J. M. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2020.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pp. 1389–1398, 2018. URL https://doi.org/10.1145/3178876.3186044.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 325–333, 2016. URL http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf.

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8895–8906, 2018. URL http://papers.nips.cc/paper/8105-balanced-policy-evaluation-and-learning.pdf.

Kallus, N. and Zhou, A. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2439–2448, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/kallus18a.html.

Kannan, S., Morgenstern, J., Roth, A., Waggoner, B., and Wu, Z. S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2231–2241, 2018.

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., and Valera, I. Fair decisions despite imperfect predictions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 277–287, August 2020. URL http://proceedings.mlr.press/v108/kilbertus20a.html.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 275—284, 2017. URL https://doi.org/10.1145/3097983.3098066.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, pp. 817–824, 2008. URL https://proceedings.neurips.cc/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 2071–2080, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/li17c.html.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3156–3164, 2018. URL http://proceedings.mlr.press/v80/liu18c.html.

Lum, K. and Isaac, W. To predict and serve? *Significance*, 13(5):14–19, 2016. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2016.00960.x.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 359—368, 2019. URL https://doi.org/10.1145/3287560.3287599.

Osband, I. and Roy, B. V. Bootstrapped thompson sampling and deep exploration, 2015. arXiv e-print https://arxiv.org/abs/1507.00300.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 7599–7609, 13–18 Jul 2020.

URL http://proceedings.mlr.press/v119/perdomo20a.html.

Rosenfeld, N., Hilgard, A., Ravindranath, S. S., and Parkes, D. C. From predictions to decisions: Using lookahead regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16 (52):1731–1755, 2015. URL http://jmlr.org/papers/v16/swaminathan15a.html.

Tsirtsis, S. and Gomez-Rodriguez, M. Decisions, counterfactual explanations and strategic behavior. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Weber, R. On the gittins index for multiarmed bandits. *Annals of Applied Probability*, 2(4):1024–1033, 1992.

Zhang, X., Khaliligarekani, M., Tekin, C., and Liu, M. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 15269–15278, 2019. URL https://proceedings.neurips.cc/paper/2019/file/7690dd4db7a92524c684e3191919eb6b-Paper.pdf.

Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellström, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

A. Proofs

A.1. Proof of Theorem 1

We start by deriving the dynamic programming recursion that governs the value function starting from initial state $(\mu_0, \nu_0) = (\mu, \nu)$,

$$V^{\Pi}(\mu, \nu) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i \Pi(\mu_i, \nu_i)(\mu_i - c) \,\middle|\, (\mu_0, \nu_0) = (\mu, \nu)\right], \tag{7}$$

where we have used the fact that $\mu_i = \mathbb{E}[\rho_i]$ is the mean of the posterior success probability ρ_i. This recursion is given in turn by the state transitions of (μ_i, ν_i). There are three possible transitions corresponding to an acceptance decision followed by success, acceptance followed by failure, and rejection (followed by no observation). Thus in terms of the number of successes σ_i, the state transitions are

$$(\sigma_{i+1}, \nu_{i+1}) = \begin{cases} (\sigma_i + 1, \nu_i + 1) & \text{with probability } \pi_i \mu_i \text{ and reward } 1 - c, \\ (\sigma_i, \nu_i + 1) & \text{with probability } \pi_i(1 - \mu_i) \text{ and reward } -c, \\ (\sigma_i, \nu_i) & \text{with probability } 1 - \pi_i \text{ and reward } 0, \end{cases}$$

where we have defined acceptance probability $\pi_i = \Pi(\mu_i, \nu_i)$ and again used the fact that μ_i is the success probability marginalized over the posterior. In terms of μ_i instead of σ_i, we have

$$(\mu_{i+1}, \nu_{i+1}) = \begin{cases} \left(\frac{\mu_i \nu_i + 1}{\nu_i + 1}, \nu_i + 1\right) & \text{with probability } \pi_i \mu_i \text{ and reward } 1 - c, \\ \left(\frac{\mu_i \nu_i}{\nu_i + 1}, \nu_i + 1\right) & \text{with probability } \pi_i(1 - \mu_i) \text{ and reward } -c, \\ (\mu_i, \nu_i) & \text{with probability } 1 - \pi_i \text{ and reward } 0. \end{cases} \tag{8}$$

A recursion for $V^{\Pi}(\mu, \nu)$ is now obtained by separating out the $i = 0$ term from (7), using (8) for the possible transitions to (μ_1, ν_1), and then reusing the definition of $V^{\Pi}(\mu, \nu)$ in (7). This yields

$$V^{\Pi}(\mu, \nu) = \pi(\mu - c) + \gamma\left[\pi\mu V^{\Pi}\left(\frac{\mu\nu + 1}{\nu + 1}, \nu + 1\right) + \pi(1 - \mu)V^{\Pi}\left(\frac{\mu\nu}{\nu + 1}, \nu + 1\right) + (1 - \pi)V^{\Pi}(\mu, \nu)\right], \tag{9}$$

where $\pi = \Pi(\mu, \nu)$ to simplify notation. The first term $\pi(\mu - c)$ is the immediate reward and the quantity in square brackets is the expected future reward, discounted by γ. Solving the previous equation for $V^{\Pi}(\mu, \nu)$ yields

$$V^{\Pi}(\mu, \nu) = \frac{\pi}{1 - \gamma + \gamma\pi}\left(\mu - c + \gamma\left[\mu V^{\Pi}\left(\frac{\mu\nu + 1}{\nu + 1}, \nu + 1\right) + (1 - \mu)V^{\Pi}\left(\frac{\mu\nu}{\nu + 1}, \nu + 1\right)\right]\right). \tag{10}$$

An optimal policy is obtained recursively by assuming that it is followed from state $\nu + 1$ onward, which replaces $V^{\Pi}(\cdot, \nu + 1)$ in (10) by the optimal value $V^*(\cdot, \nu + 1)$, and then maximizing the right-hand side of (10) with respect to the current action π (Bertsekas, 2005):

$$V^*(\mu, \nu) = \max_{\pi \in [0,1]} \frac{\pi}{1 - \gamma + \gamma\pi} \underbrace{\left(\mu - c + \gamma\left[\mu V^*\left(\frac{\mu\nu + 1}{\nu + 1}, \nu + 1\right) + (1 - \mu)V^*\left(\frac{\mu\nu}{\nu + 1}, \nu + 1\right)\right]\right)}_{\tilde{V}(\mu,\nu)}. \tag{11}$$

The key observation is that the dependence on π is confined to the first factor above and is moreover monotonically increasing or decreasing depending on the sign of $\tilde{V}(\mu, \nu)$. It follows that $V^*(\mu, \nu) = \max\{\tilde{V}(\mu, \nu), 0\}$, which proves the theorem.

A.2. Proof of Propositions 2 and 3

Both propositions are proven by induction over decreasing ν. Technically, the proofs are only for the approximations $V^N(\mu, \nu)$ to $V^*(\mu, \nu)$ described in Section 3. However by taking $N \to \infty$, $V^N \to V^*$ and the properties extend to V^* as well. The proofs also require the following lemma pertaining to convex functions.

Lemma 1. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be convex. Then for any x and $\alpha \in [0, 1]$, $\alpha f(x + (1 - \alpha)\delta) + (1 - \alpha)f(x - \alpha\delta)$ is non-decreasing in $\delta \geq 0$.*

This lemma can be proved graphically by drawing chords that intersect $f(x)$ at $x - \alpha\delta$ and $x + (1-\alpha)\delta$ for increasing δ. An algebraic proof is provided below.

Proof. Let $0 \leq \delta_1 \leq \delta_2$. By the convexity of f,

$$\begin{aligned} f(x + (1-\alpha)\delta_1) &\leq \left(1 - \frac{\delta_1}{\delta_2}\right) f(x) + \frac{\delta_1}{\delta_2} f(x + (1-\alpha)\delta_2), \\ f(x - \alpha\delta_1) &\leq \left(1 - \frac{\delta_1}{\delta_2}\right) f(x) + \frac{\delta_1}{\delta_2} f(x - \alpha\delta_2). \end{aligned}$$

Multiplying the first inequality by α, the second inequality by $1-\alpha$, and summing,

$$\alpha f(x + (1-\alpha)\delta_1) + (1-\alpha) f(x - \alpha\delta_1) \;\leq\; \left(1 - \frac{\delta_1}{\delta_2}\right) f(x) + \frac{\delta_1}{\delta_2}\left[\alpha f(x + (1-\alpha)\delta_2) + (1-\alpha) f(x - \alpha\delta_2)\right].$$

Since we also have

$$f(x) \leq \alpha f(x + (1-\alpha)\delta_2) + (1-\alpha) f(x - \alpha\delta_2),$$

the result follows, i.e.

$$\alpha f(x + (1-\alpha)\delta_1) + (1-\alpha) f(x - \alpha\delta_1) \leq \alpha f(x + (1-\alpha)\delta_2) + (1-\alpha) f(x - \alpha\delta_2).$$

□

For Proposition 2, the base case is $\nu = N+1$, for which $V^N(\mu, N+1) = V^*(\mu, \infty) = \max\{\mu - c, 0\}/(1-\gamma)$ is both non-decreasing and convex in μ. Appendix A.2.1 proves the inductive step, i.e. that $V^N(\mu, \nu+1)$ being non-decreasing and convex in μ implies the same for $V^N(\mu, \nu)$, with the help of Lemma 1.

For Proposition 3, the base case requires showing that $V^N(\mu, N) \geq V^N(\mu, N+1) = V^*(\mu, \infty)$, where $V^N(\mu, N)$ is obtained from (3) with $V^*(\mu, \infty)$ in place of $V^*(\mu, \nu+1)$. This calculation is shown in Appendix A.2.2. The inductive step follows in Appendix A.2.2, again using Lemma 1.

A.2.1. Proof of Proposition 2: Inductive step

We wish to show that $V^N(\mu, \nu)$ is non-decreasing and convex given that $V^N(\mu, \nu+1)$ has these properties. Since $V^N(\mu, \nu) = \max\{\tilde{V}(\mu, \nu), 0\}$ and $\tilde{V}(\mu, \nu) = \mu - c + \gamma\bar{V}(\mu, \nu)$, where

$$\bar{V}(\mu, \nu) = \mu V^N\left(\frac{\mu\nu + 1}{\nu + 1}, \nu + 1\right) + (1-\mu) V^N\left(\frac{\mu\nu}{\nu + 1}, \nu + 1\right),$$

it suffices to show that $\bar{V}(\mu, \nu)$ is non-decreasing and convex. This is because these properties are preserved under addition with the function $\mu - c$, which is increasing and convex, and under the pointwise maximum with the zero function (also convex).

To show that $\bar{V}(\mu, \nu)$ is non-decreasing in μ, let $\mu_1 \leq \mu_2$. Then

$$\begin{aligned} &\bar{V}(\mu_2, \nu) - \bar{V}(\mu_1, \nu) \\ &\quad= \mu_2 V^N\left(\frac{\mu_2\nu + 1}{\nu + 1}, \nu + 1\right) + (1-\mu_2) V^N\left(\frac{\mu_2\nu}{\nu + 1}, \nu + 1\right) \\ &\qquad - \mu_1 V^N\left(\frac{\mu_1\nu + 1}{\nu + 1}, \nu + 1\right) - (1-\mu_1) V^N\left(\frac{\mu_1\nu}{\nu + 1}, \nu + 1\right) \\ &\quad= (\mu_2 - \mu_1)\left[V^N\left(\frac{\mu_2\nu + 1}{\nu + 1}, \nu + 1\right) - V^N\left(\frac{\mu_2\nu}{\nu + 1}, \nu + 1\right)\right] \\ &\qquad + \mu_1\left[V^N\left(\frac{\mu_2\nu + 1}{\nu + 1}, \nu + 1\right) + V^N\left(\frac{\mu_1\nu}{\nu + 1}, \nu + 1\right) - V^N\left(\frac{\mu_2\nu}{\nu + 1}, \nu + 1\right) - V^N\left(\frac{\mu_1\nu + 1}{\nu + 1}, \nu + 1\right)\right] \\ &\qquad + \left[V^N\left(\frac{\mu_2\nu}{\nu + 1}, \nu + 1\right) - V^N\left(\frac{\mu_1\nu}{\nu + 1}, \nu + 1\right)\right]. \end{aligned}$$

In the final right-hand side above, the first and third quantities in square brackets are non-negative because of the inductive assumption that $V^N(\mu, \nu+1)$ is non-decreasing in μ. The second bracketed quantity is also shown to be non-negative by applying Lemma 1 to $V^N(\mu, \nu+1)$, assumed to be convex in μ, with

$$x = \frac{\nu(\mu_2+\mu_1)+1}{2(\nu+1)}, \quad \alpha = \frac{1}{2}, \quad \delta_1 = \frac{\nu(\mu_2-\mu_1)-1}{2(\nu+1)} \leq \delta_2 = \frac{\nu(\mu_2-\mu_1)+1}{2(\nu+1)}.$$

Thus $\bar{V}(\mu_2, \nu) - \bar{V}(\mu_1, \nu) \geq 0$ as required.

To show that $\bar{V}(\mu, \nu)$ is convex in μ, we require

$$\alpha \bar{V}(\mu_1, \nu) + (1-\alpha)\bar{V}(\mu_2, \nu) \geq \bar{V}(\alpha\mu_1 + (1-\alpha)\mu_2, \nu) \tag{12}$$

for $\alpha \in [0, 1]$. The left-hand side yields

$$\begin{aligned}
&\alpha \bar{V}(\mu_1, \nu) + (1-\alpha)\bar{V}(\mu_2, \nu) \\
&\quad = \alpha\mu_1 V^N\left(\frac{\mu_1\nu+1}{\nu+1}, \nu+1\right) + (1-\alpha)\mu_2 V^N\left(\frac{\mu_2\nu+1}{\nu+1}, \nu+1\right) \\
&\qquad + \alpha(1-\mu_1) V^N\left(\frac{\mu_1\nu}{\nu+1}, \nu+1\right) + (1-\alpha)(1-\mu_2) V^N\left(\frac{\mu_2\nu}{\nu+1}, \nu+1\right) \\
&\quad \geq (\alpha\mu_1 + (1-\alpha)\mu_2) V^N\left(\frac{\nu}{\nu+1}\frac{\alpha\mu_1^2 + (1-\alpha)\mu_2^2}{\alpha\mu_1 + (1-\alpha)\mu_2} + \frac{1}{\nu+1}, \nu+1\right) \\
&\qquad + (1 - \alpha\mu_1 - (1-\alpha)\mu_2) V^N\left(\frac{\nu}{\nu+1}\frac{\alpha(1-\mu_1)\mu_1 + (1-\alpha)(1-\mu_2)\mu_2}{1-\alpha\mu_1 - (1-\alpha)\mu_2}, \nu+1\right)
\end{aligned} \tag{13}$$

where the convexity of $V^N(\mu, \nu+1)$ has been applied separately to the second line and third line above (note $\alpha(1-\mu_1) + (1-\alpha)(1-\mu_2) = 1 - \alpha\mu_1 - (1-\alpha)\mu_2$). The right-hand side of (12) is

$$\begin{aligned}
\bar{V}(\alpha\mu_1 + (1-\alpha)\mu_2, \nu) &= (\alpha\mu_1 + (1-\alpha)\mu_2) V^N\left(\frac{\nu(\alpha\mu_1 + (1-\alpha)\mu_2)+1}{\nu+1}, \nu+1\right) \\
&\quad + (1 - \alpha\mu_1 - (1-\alpha)\mu_2) V^N\left(\frac{\nu(\alpha\mu_1 + (1-\alpha)\mu_2)}{\nu+1}, \nu+1\right).
\end{aligned} \tag{14}$$

The right-hand sides of (13) and (14) are both convex combinations of $V^N(\mu, \nu+1)$ with the same weights, which suggests using Lemma 1 (with $\alpha \leftarrow \alpha\mu_1 + (1-\alpha)\mu_2$) to compare them. With the two terms in (14) playing the roles of $f(x + (1-\alpha)\delta)$ and $f(x - \alpha\delta)$ in Lemma 1, we find

$$\begin{aligned}
x &= (\alpha\mu_1 + (1-\alpha)\mu_2)\frac{\nu(\alpha\mu_1 + (1-\alpha)\mu_2)+1}{\nu+1} + (1 - \alpha\mu_1 - (1-\alpha)\mu_2)\frac{\nu(\alpha\mu_1 + (1-\alpha)\mu_2)}{\nu+1} \\
&= \frac{\nu(\alpha\mu_1 + (1-\alpha)\mu_2)}{\nu+1} + \frac{\alpha\mu_1 + (1-\alpha)\mu_2}{\nu+1} \\
&= \alpha\mu_1 + (1-\alpha)\mu_2,
\end{aligned}$$

and a similar calculation with (13) yields the same value for x. Furthermore, comparing the arguments of the first terms in (13) and (14),

$$\frac{\nu}{\nu+1}\frac{\alpha\mu_1^2 + (1-\alpha)\mu_2^2}{\alpha\mu_1 + (1-\alpha)\mu_2} + \frac{1}{\nu+1} - \frac{\nu(\alpha\mu_1 + (1-\alpha)\mu_2)+1}{\nu+1} = \frac{\nu}{\nu+1}\frac{\alpha\mu_1^2 + (1-\alpha)\mu_2^2 - (\alpha\mu_1 + (1-\alpha)\mu_2)^2}{\alpha\mu_1 + (1-\alpha)\mu_2} \geq 0,$$

where the inequality is due to the convexity of the function $\mu \mapsto \mu^2$. This indicates that the δ corresponding to (13) (which will not be computed explicitly) is greater than or equal to the δ corresponding to (14). Lemma 1 then implies that the right-hand side of (13) is greater than or equal to the right-hand side of (14), thus completing the proof of (12). (Note that this proof of convexity only required $V^N(\mu, \nu+1)$ to be convex in μ, not necessarily non-decreasing.)

A.2.2. Proof of Proposition 3

First the base case is proven, i.e. $V^N(\mu, N) \geq V^N(\mu, N+1)$, where $V^N(\mu, N+1) = V^*(\mu, \infty) = \max\{\mu - c, 0\}/(1-\gamma)$ and $V^N(\mu, N)$ is given by recursion (3) (with V^* replaced by V^N). There are three cases corresponding to where the arguments on the right-hand side of (3), $(\mu N)/(N+1)$ and $(\mu N + 1)/(N+1)$, fall with respect to the threshold c.

Case $(\mu N + 1)/(N+1) \leq c$: Since this implies $(\mu N)/(N+1) < c$ and $\mu \leq c$, we have

$$V^N\left(\frac{\mu N + 1}{N+1}, N+1\right) = V^N\left(\frac{\mu N}{N+1}, N+1\right) = 0$$

and the right-hand side of (3) yields $V^N(\mu, N) = 0$. This is equal to $V^N(\mu, N+1) = 0$.

Case $(\mu N)/(N+1) > c$: This implies $(\mu N + 1)/(N+1) > c$ and $\mu > c$ as well. $V^N(\mu, N+1)$ is then a linear function over the interval $[(\mu N)/(N+1), (\mu N + 1)/(N+1)]$ and

$$\mu V^N\left(\frac{\mu N + 1}{N+1}, N+1\right) + (1-\mu) V^N\left(\frac{\mu N}{N+1}, N+1\right) = V^N(\mu, N+1) = \frac{\mu - c}{1-\gamma}.$$

Eq. (3) then gives

$$V^N(\mu, N) = \max\left\{\mu - c + \gamma \times \frac{\mu - c}{1-\gamma}, 0\right\} = \frac{\mu - c}{1-\gamma} = V^N(\mu, N+1).$$

Case $(\mu N)/(N+1) \leq c < (\mu N + 1)/(N+1)$: Only one of the $V^N(\cdot, N+1)$ terms in (3) is non-zero, resulting in

$$V^N(\mu, N) = \max\left\{\mu - c + \frac{\gamma\mu}{1-\gamma}\left(\frac{\mu N + 1}{N+1} - c\right), 0\right\} = \max\left\{\underbrace{(\mu - c)\left(1 + \frac{\gamma\mu}{1-\gamma}\right) + \frac{\gamma\mu(1-\mu)}{(1-\gamma)(N+1)}}_{\tilde{V}(\mu, N)}, 0\right\}.$$

In comparison,

$$V^N(\mu, N+1) = \max\left\{\tilde{V}(\mu, N+1), 0\right\}, \quad \tilde{V}(\mu, N+1) = \frac{\mu - c}{1-\gamma}.$$

Subtracting,

$$\begin{aligned}
\tilde{V}(\mu, N) - \tilde{V}(\mu, N+1) &= (\mu - c)\frac{\gamma(\mu - 1)}{1-\gamma} + \frac{\gamma\mu(1-\mu)}{(1-\gamma)(N+1)} \\
&= \frac{\gamma(1-\mu)}{1-\gamma}\left(c - \mu + \frac{\mu}{N+1}\right) \\
&\geq 0
\end{aligned}$$

because $(\mu N)/(N+1) \leq c$ for this case. It follows that $V^N(\mu, N) \geq V^N(\mu, N+1)$.

Now for the inductive step, assume that $V^N(\mu, \nu) \geq V^N(\mu, \nu + 1)$. Then

$$\begin{aligned}
V^N(\mu, \nu - 1) &= \max\left\{\mu - c + \gamma\left[\mu V^N\left(\mu + \frac{1-\mu}{\nu}, \nu\right) + (1-\mu) V^N\left(\mu - \frac{\mu}{\nu}, \nu\right)\right], 0\right\} \\
&\geq \max\left\{\mu - c + \gamma\left[\mu V^N\left(\mu + \frac{1-\mu}{\nu}, \nu + 1\right) + (1-\mu) V^N\left(\mu - \frac{\mu}{\nu}, \nu + 1\right)\right], 0\right\} \\
&\geq \max\left\{\mu - c + \gamma\left[\mu V^N\left(\mu + \frac{1-\mu}{\nu + 1}, \nu + 1\right) + (1-\mu) V^N\left(\mu - \frac{\mu}{\nu + 1}, \nu + 1\right)\right], 0\right\} \\
&= V^N(\mu, \nu),
\end{aligned}$$

where the second inequality follows from the convexity of $V^N(\mu, \nu + 1)$ in μ (Proposition 2) and application of Lemma 1 with $x = \mu$, $\alpha = \mu$, and $\delta_1 = 1/\nu > \delta_2 = 1/(\nu + 1)$.

A.3. Proof of Proposition 4

As in the proofs of Propositions 2 and 3, we work with the approximation $V^N(\mu, \nu)$ to the optimal value function and then take $N \to \infty$. The proposition is proven by deriving a piecewise-linear upper bound $U(\mu, \nu)$ on $V^N(\mu, \nu)$. The μ-intercept of $U(\mu, \nu)$, i.e., the largest μ for which $U(\mu, \nu) = 0$, then provides a lower bound on c_ν, the μ-intercept of $V^N(\mu, \nu)$. Piecewise linearity allows the μ-intercept of $U(\mu, \nu)$ to be expressed straightforwardly in closed form.

We again use induction over decreasing ν to derive the upper bound $U(\mu, \nu)$. For the base case $\nu = N+1$, $V^N(\mu, N+1) = V^*(\mu, \infty) = \max\{\mu - c, 0\}/(1-\gamma)$ is already piecewise-linear, so we take

$$U(\mu, N+1) = V^N(\mu, N+1) = \max\left\{\frac{\mu - c}{1-\gamma}, 0\right\}. \tag{15}$$

Note that $U(\mu, N+1)$ is of the form

$$U(\mu, \nu) = \max\left\{\frac{1-c}{1-\gamma}\frac{\mu - \underline{c}_\nu}{1-\underline{c}_\nu}, 0\right\} \tag{16}$$

with μ-intercept $\underline{c}_{N+1} = c$.

For the inductive step, assume $V^N(\mu, \nu+1) \le U(\mu, \nu+1)$ and that $U(\mu, \nu+1)$ has the form in (16) with $\underline{c}_{\nu+1} \le c$. Then from recursion (3) and using the fact that $V^N(\mu, \nu+1)$ is non-decreasing in μ (Proposition 2),

$$\begin{aligned} V^N(\mu, \nu) &= \max\left\{\mu - c + \gamma\left[\mu V^N\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^N\left(\frac{\mu\nu}{\nu+1}, \nu+1\right)\right], 0\right\} \\ &\le \max\left\{\mu - c + \gamma V^N\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right), 0\right\} \\ &\le \max\left\{\mu - c + \gamma U\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right), 0\right\} := U(\mu, \nu). \end{aligned} \tag{17}$$

It must now be shown that the upper bound $U(\mu, \nu)$ defined above is also of the form in (16) with intercept $\underline{c}_\nu \le c$. First, since $U(\mu, \nu+1)$ is piecewise-linear and non-decreasing in μ, so too is $U(\mu, \nu)$ as these properties are preserved by (17). Moreover, the leftmost piece is identically zero, as in (16), and its right endpoint is the intercept $\underline{c}_\nu$. Second, there are two possibilities at $\underline{c}_\nu$:

$$U\left(\frac{\underline{c}_\nu \nu + 1}{\nu+1}, \nu+1\right) > 0, \tag{18a}$$

$$U\left(\frac{\underline{c}_\nu \nu + 1}{\nu+1}, \nu+1\right) = 0. \tag{18b}$$

If the latter case (18b) is true, then since $\underline{c}_\nu$ is by definition the largest value such that $U(\underline{c}_\nu, \nu) = 0$, (17) implies $\underline{c}_\nu = c$. This however leads to a contradiction because $(c\nu + 1)/(\nu+1) > c$ while $\underline{c}_{\nu+1} \le c$ by assumption, and hence substitution into (16) shows that case (18a) is actually true. Given (18a) then, we must have

$$U\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) = \frac{1-c}{1-\gamma}\frac{(\mu\nu+1)/(\nu+1) - \underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} > 0 \quad \forall\, \mu \ge \underline{c}_\nu$$

since $U(\mu, \nu+1)$ is non-decreasing. Equation (17) can therefore be rewritten as

$$U(\mu, \nu) = \max\left\{\mu - c + \gamma\frac{1-c}{1-\gamma}\frac{(\mu\nu+1)/(\nu+1) - \underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}}, 0\right\}, \tag{19}$$

which is a non-decreasing piecewise-linear function with two pieces, like (16). Furthermore, we must have $\underline{c}_\nu \le c$, since (19) shows that $U(\mu, \nu) > 0$ for $\mu > c$. Lastly, we use (17) or (19) to check

$$U(1, \nu) = 1 - c + \gamma\frac{1-c}{1-\gamma} = \frac{1-c}{1-\gamma},$$

in agreement with (16). We conclude that $U(\mu, \nu)$ is also given by (16) for some $\underline{c}_\nu \le c$.

The next step is to derive a recursion for $\underline{c}_\nu$ in terms of $\underline{c}_{\nu+1}$, toward obtaining an explicit expression for $\underline{c}_\nu$. The intercept $\underline{c}_\nu$ is the value of μ for which the first term on the right-hand side of (19) is equal to zero. Hence

$$\begin{aligned}\underline{c}_\nu - c + \gamma \frac{1-c}{1-\gamma} \frac{(\underline{c}_\nu \nu + 1)/(\nu+1) - \underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} &= 0,\\ \underline{c}_\nu - c + \gamma \frac{1-c}{1-\gamma} \frac{1}{1-\underline{c}_{\nu+1}} \left(\frac{\nu}{\nu+1}(\underline{c}_\nu - \underline{c}_{\nu+1}) + \frac{1-\underline{c}_{\nu+1}}{\nu+1} \right) &= 0.\end{aligned}$$

Multiplying through by $(1-\gamma)(\nu+1)$ and adding and subtracting c in the numerator,

$$(1-\gamma)(\nu+1)(\underline{c}_\nu - c) + \gamma(1-c)\left(\nu \frac{\underline{c}_\nu - c + c - \underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} + 1 \right) = 0.$$

Rearranging to solve for $c - \underline{c}_\nu$,

$$\begin{aligned}\left[(1-\gamma)(\nu+1) + \gamma\nu \frac{1-c}{1-\underline{c}_{\nu+1}}\right](c - \underline{c}_\nu) &= \gamma(1-c)\left(\nu \frac{c-\underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} + 1\right),\\ \left[\nu + 1 - \gamma\left(\nu\frac{c-\underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} + 1\right)\right](c-\underline{c}_\nu) &= \gamma\left(\nu\frac{c-\underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} + 1\right)(1-c),\\ c - \underline{c}_\nu &= \frac{\gamma\left(\nu\frac{c-\underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} + 1\right)}{\nu+1-\gamma\left(\nu\frac{c-\underline{c}_{\nu+1}}{1-\underline{c}_{\nu+1}} + 1\right)}(1-c), \qquad (20)\end{aligned}$$

which is the desired recursion.

For the case $\nu = N$, we have $\underline{c}_{N+1} = c$ so (20) simplifies to

$$c - \underline{c}_N = \frac{\gamma}{N+1-\gamma}(1-c). \qquad (21)$$

For $\nu < N$, based on (20) and (21), we postulate that $c - \underline{c}_\nu$ has the general form

$$c - \underline{c}_\nu = \frac{P_\nu(\gamma)}{\nu+1-P_\nu(\gamma)}(1-c), \quad \nu = N, N-1, \dots, \qquad (22)$$

where $P_\nu(\gamma)$ is a polynomial in γ, and $P_N(\gamma) = \gamma$. From (22), we also obtain

$$\begin{aligned}1 - \underline{c}_\nu = 1 - c + c - \underline{c}_\nu = \left(1 + \frac{P_\nu(\gamma)}{\nu+1-P_\nu(\gamma)}\right)(1-c) &= \frac{\nu+1}{\nu+1-P_\nu(\gamma)}(1-c),\\ \frac{c-\underline{c}_\nu}{1-\underline{c}_\nu} &= \frac{P_\nu(\gamma)}{\nu+1}. \qquad (23)\end{aligned}$$

Substituting (23) into (20),

$$c - \underline{c}_\nu = \frac{\gamma\left(\frac{\nu}{\nu+2}P_{\nu+1}(\gamma) + 1\right)}{\nu+1-\gamma\left(\frac{\nu}{\nu+2}P_{\nu+1}(\gamma)+1\right)}(1-c).$$

Comparing the above to (22), we find

$$P_\nu(\gamma) = \gamma\left(\frac{\nu}{\nu+2}P_{\nu+1}(\gamma) + 1\right). \qquad (24)$$

Thus if $P_{\nu+1}(\gamma)$ is a polynomial, so too is $P_\nu(\gamma)$.

An explicit expression for $P_\nu(\gamma)$ can be obtained from the recursion in (24), and hence for $c - \underline{c}_\nu$ in (22) as well. From (24), it can be seen that $P_\nu(\gamma)$ is a polynomial of degree $N + 1 - \nu$. Because of the recursive multiplication by $\nu/(\nu+2)$, we postulate that its coefficients are given by *rising factorials*, defined as

$$(\nu)_p = \begin{cases} \nu(\nu+1)\dots(\nu+p-1), & p > 0,\\ 1, & p = 0.\end{cases}$$

Specifically, let

$$P_\nu(\gamma) = \gamma \sum_{p=0}^{N-\nu} \frac{(\nu)_p}{(\nu+2)_p} \gamma^p. \tag{25}$$

Substituting this into the right-hand side of (24) to verify:

$$\begin{aligned}
\gamma \left(\frac{\nu}{\nu+2} P_{\nu+1}(\gamma) + 1 \right) &= \gamma \left(\frac{\nu}{\nu+2} \gamma \sum_{p=0}^{N-\nu-1} \frac{(\nu+1)_p}{(\nu+3)_p} \gamma^p + 1 \right) \\
&= \gamma \left(1 + \sum_{p=0}^{N-\nu-1} \frac{\nu(\nu+1)_p}{(\nu+2)(\nu+3)_p} \gamma^{p+1} \right) \\
&= \gamma \sum_{p=0}^{N-\nu} \frac{(\nu)_p}{(\nu+2)_p} \gamma^p,
\end{aligned}$$

in agreement with (25).

We now take $N \to \infty$ and recognize the resulting infinite power series in (25) as a Gaussian hypergeometric function:

$$P_\nu(\gamma) = \gamma \cdot {}_2F_1(1, \nu; \nu+2; \gamma). \tag{26}$$

Substituting (26) into (22) and using $c - \underline{c}_\nu$ to upper bound $c - c_\nu$ establishes the first upper bound in the proposition.

The second, looser upper bound in the proposition comes from combining two upper bounds on the hypergeometric function. For the first bound, the ratio of rising factorials is bounded by 1 to yield

$${}_2F_1(1, \nu; \nu+2; \gamma) = \sum_{p=0}^{\infty} \frac{(\nu)_p}{(\nu+2)_p} \gamma^p \le \sum_{p=0}^{\infty} \gamma^p = \frac{1}{1-\gamma}. \tag{27}$$

The second upper bound results from writing

$$\frac{(\nu)_p}{(\nu+2)_p} = \frac{\nu(\nu+1)}{(\nu+p)(\nu+p+1)} = \nu(\nu+1) \left(\frac{1}{\nu+p} - \frac{1}{\nu+p+1} \right), \quad p \ge 2,$$

where the first equality is due to cancellation. Hence

$$\begin{aligned}
{}_2F_1(1, \nu; \nu+2; \gamma) &= 1 + \frac{\nu}{\nu+2}\gamma + \nu(\nu+1) \sum_{p=2}^{\infty} \left(\frac{1}{\nu+p} - \frac{1}{\nu+p+1} \right) \gamma^p \\
&= 1 + \frac{\nu}{\nu+2}\gamma + \frac{\nu(\nu+1)}{\nu+2}\gamma^2 + \sum_{p=2}^{\infty} \frac{\gamma^{p+1} - \gamma^p}{\nu+p+1}.
\end{aligned}$$

The hypergeometric function is then bounded by its limit as $\gamma \to 1$, as follows:

$${}_2F_1(1, \nu; \nu+2; \gamma) \le 1 + \frac{\nu}{\nu+2} + \frac{\nu(\nu+1)}{\nu+2} = \nu + 1. \tag{28}$$

The second upper bound in the proposition follows from (27) and (28).

A.4. Proof of Proposition 5

This proposition is also proven by induction over decreasing ν. For the base case $\nu = N+1$,

$$V^N(\mu, N+1) = V^*(\mu, \infty) = \max\left\{ \frac{\mu - c}{1-\gamma}, 0 \right\} \le V^*(\mu, N+1) \le U(\mu, N+1) = \max\left\{ \frac{1-c}{1-\gamma} \frac{\mu - \underline{c}_{N+1}}{1 - \underline{c}_{N+1}}, 0 \right\},$$

where the first inequality is due to Proposition 3 and the second due to the upper bound $U(\mu,\nu)$ (16) established in the proof of Proposition 4. Looking at the piecewise-linear expressions for $V^N(\mu, N+1)$ and $U(\mu, N+1)$, the largest difference between them occurs at $\mu = c$. Hence

$$U(\mu, N+1) - V^N(\mu, N+1) \leq \frac{1-c}{1-\gamma}\frac{c-\underline{c}_{N+1}}{1-\underline{c}_{N+1}} - 0.$$

Using (23),

$$U(\mu, N+1) - V^N(\mu, N+1) \leq \frac{1-c}{1-\gamma}\frac{P_{N+1}(\gamma)}{N+2}.$$

Summarizing the base case, we have

$$V^N(\mu, N+1) \leq V^*(\mu, N+1) \leq U(\mu, N+1) \leq V^N(\mu, N+1) + \frac{1-c}{1-\gamma}\frac{P_{N+1}(\gamma)}{N+2} \quad \forall\, \mu \in [0,1]. \tag{29}$$

For the inductive step, assume that a slight generalization of (29) holds,

$$V^N(\mu, \nu+1) \leq V^*(\mu, \nu+1) \leq V^N(\mu, \nu+1) + \gamma^{N-\nu}\frac{1-c}{1-\gamma}\frac{P_{N+1}(\gamma)}{N+2} \quad \forall\, \mu \in [0,1], \tag{30}$$

where the last term is further discounted by $\gamma^{N-\nu}$. Then from recursion (3) and using (30),

$$\begin{aligned}
V^N(\mu,\nu) &= \max\left\{\mu - c + \gamma\left[\mu V^N\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^N\left(\frac{\mu\nu}{\nu+1}, \nu+1\right)\right], 0\right\}\\
&\leq \max\left\{\mu - c + \gamma\left[\mu V^*\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^*\left(\frac{\mu\nu}{\nu+1}, \nu+1\right)\right], 0\right\}\\
&= V^*(\mu,\nu)\\
&\leq \max\left\{\mu - c + \gamma\left[\mu V^N\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^N\left(\frac{\mu\nu}{\nu+1}, \nu+1\right) + \gamma^{N-\nu}\frac{1-c}{1-\gamma}\frac{P_{N+1}(\gamma)}{N+2}\right], 0\right\}\\
&\leq \max\left\{\mu - c + \gamma\left[\mu V^N\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^N\left(\frac{\mu\nu}{\nu+1}, \nu+1\right)\right], 0\right\} + \gamma^{N+1-\nu}\frac{1-c}{1-\gamma}\frac{P_{N+1}(\gamma)}{N+2}\\
&= V^N(\mu,\nu) + \gamma^{N+1-\nu}\frac{1-c}{1-\gamma}\frac{P_{N+1}(\gamma)}{N+2}.
\end{aligned} \tag{31}$$

The last inequality follows from $\max\{a+b, 0\} \leq \max\{a, 0\} + b$ for $b \geq 0$. This completes the induction. The proposition results after substituting (26) for $P_{N+1}(\gamma)$ into (31), rewriting $(1-c)/(1-\gamma) = V^*(1,\nu)$, and rearranging.

A.5. Proof of Theorem 6

To formalize the claim that the optimal finite-domain policy consists of parallel optimal homogeneous policies, define

$$V^{\Pi}(x; \mu_0(x), \nu_0(x)) = \mathbb{E}\left[\Pi(x_0)(\rho_0(x_0) - c) + \sum_{i>0: x_i = x}\gamma^i \Pi(x_i)(\rho_i(x_i) - c) \,\middle|\, x_0 = x, \rho_0(x) \sim B(\mu_0(x), \nu_0(x))\right] \tag{32}$$

as the expected discounted sum of rewards conditioned on $x_0 = x$ and restricted to $x_i = x$ thereafter. Here, $\rho_i(x)$ is the posterior success probability at x before individual i arrives, which has a beta distribution with parameters $\mu_i(x)$, $\nu_i(x)$, and $\mu_0(x)$, $\nu_0(x)$ represent the initial state. It can be shown that the unrestricted sum of rewards in (1) is a weighted sum of $V^{\Pi}(x; \mu_0(x), \nu_0(x))$ over $x \in \mathcal{X}$ (the weights are related to the expected time of first occurrence of $x_i = x$). Hence it suffices to maximize each $V^{\Pi}(x; \mu_0(x), \nu_0(x))$ independently.

The value function $V^{\Pi}(x; \mu_0(x), \nu_0(x))$ (32) obeys a similar recursion as in (9). Specifically, the first term at time index $i = 0$, $\pi(\mu - c)$, is the same (modulo additional notation), and the quantity in square brackets in (9) is also the same because the same state transitions (of either (μ, ν) or $(\mu_i(x), \nu_i(x))$) occur with the same probabilities. The difference is that the time of next occurrence of $x_i = x$ is random, which affects the discount factor in front of the square brackets in (9). Define

$I(x)$ to be the time index of the first occurrence of $x_i = x$ after $i = 0$. Conditioned on $I(x) = i$, the discount factor is γ^i. Taking an expectation over $I(x)$, the recursion for $V^\Pi(x; \mu_0(x), \nu_0(x))$ is therefore

$$V^\Pi(x; \mu_0(x), \nu_0(x)) = \pi(\mu_0(x) - c) + \mathbb{E}\left[\gamma^{I(x)}\right] \left[\pi \mu_0(x) V^\Pi\left(x; \frac{\mu_0(x)\nu_0(x) + 1}{\nu_0(x) + 1}, \nu_0(x) + 1\right)\right.$$
$$\left. + \pi(1 - \mu_0(x)) V^\Pi\left(x; \frac{\mu_0(x)\nu_0(x)}{\nu_0(x) + 1}, \nu_0(x) + 1\right) + (1 - \pi) V^\Pi(x; \mu_0(x), \nu_0(x))\right],$$

where $\pi = \Pi(\mu_0(x), \nu_0(x))$. We define $\bar{\gamma}(x) := \mathbb{E}\left[\gamma^{I(x)}\right]$ to be the *effective discount factor* at x. Since $I(x)$ follows a geometric distribution with parameter equal to $p(x)$, the probability of $X = x$, we have

$$\bar{\gamma}(x) = \sum_{i=1}^{\infty} \gamma^i p(x)(1 - p(x))^{i-1} = \frac{\gamma p(x)}{1 - \gamma(1 - p(x))}$$

as in (4).

B. Additional Examples of Optimal Homogeneous Value Functions

Figure 4 shows additional examples of optimal value function approximations $V^N(\mu, \nu)$ for $\gamma = 0.95$ to complement Figure 1.

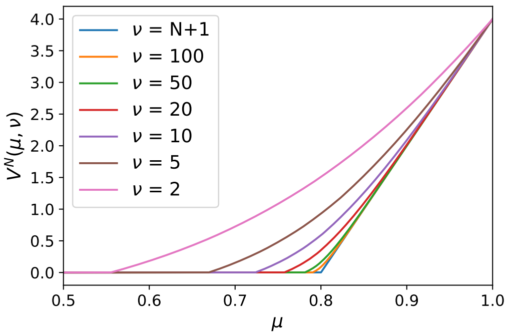


Figure 4. Optimal value function approximations $V^N(\mu, \nu)$ for $N = 1000$, $c = 0.8$, and $\gamma = 0.95$.

C. The Homogeneous Case with Undiscounted Average Reward

This appendix briefly considers the case of *undiscounted average* reward,

$$\lim_{N \to \infty} \mathbb{E}\left[\frac{1}{N} \sum_{i=0}^{N-1} \pi_i(\rho_i - c)\right], \tag{33}$$

in the homogeneous setting. The same POMDP approach is followed, where ρ_i refers to the beta posterior distribution for the success probability, parametrized by belief state (μ_i, ν_i). Let $\pi_i = \Pi(\mu_i, \nu_i)$ be the acceptance probability given by a policy Π for state (μ_i, ν_i); the expected immediate reward is then $\pi_i(\mu_i - c)$, recalling that $\mu_i = \mathbb{E}[\rho_i]$. Define

$$V_i^\Pi(\mu_i, \nu_i) = \frac{1}{N} \sum_{j=i}^{N-1} \mathbb{E}\left[\pi_j(\mu_j - c) \mid \mu_i, \nu_i\right] \tag{34}$$

to be the sum of rewards, divided by N, starting from individual i under policy Π. We wish to maximize $V_0^\Pi(\mu_0, \nu_0)$ in the limit $N \to \infty$. As the number of observations $\nu_i \to \infty$, we again have $\mu_i \to \rho$ by the law of large numbers and optimal reward $V_i^*(\mu, \infty) = \max\{\mu - c, 0\}$.

Equation (34) can be rewritten as a recursion using the same state transition probabilities as in (8):

$$V_i^{\Pi}(\mu_i, \nu_i) = \frac{1}{N}\pi_i(\mu_i - c) + \pi_i \mu_i V_{i+1}^{\Pi}\left(\frac{\mu_i \nu_i + 1}{\nu_i + 1}, \nu_i + 1\right) + \pi_i(1-\mu_i)V_{i+1}^{\Pi}\left(\frac{\mu_i \nu_i}{\nu_i + 1}, \nu_i + 1\right) + (1-\pi_i)V_{i+1}^{\Pi}(\mu_i, \nu_i).$$

Taking the limit $N \to \infty$, the first term vanishes and the sample index i again ceases to matter, i.e., $V_{i+1}^{\Pi} \to V_i^{\Pi} = V^{\Pi}$ and the subscript i is dropped elsewhere. The result can be rearranged to yield

$$\pi\left(V^{\Pi}(\mu,\nu) - \mu V^{\Pi}\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) - (1-\mu)V^{\Pi}\left(\frac{\mu\nu}{\nu+1}, \nu+1\right)\right) = 0.$$

There are two cases corresponding to choices of actions: Either $\pi = \Pi(\mu,\nu) = 0$, which stops the state evolution and results in zero reward, $V^{\Pi}(\mu,\nu) = 0$, or $\pi > 0$ and the value function satisfies

$$V^{\Pi}(\mu,\nu) = \mu V^{\Pi}\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^{\Pi}\left(\frac{\mu\nu}{\nu+1}, \nu+1\right). \tag{35}$$

Below it is shown that the choice $\pi > 0$ leads to non-negative value $V^{\Pi}(\mu,\nu) \geq 0$ and is hence preferred in all states (μ,ν).

Theorem 7. *Policies that maximize undiscounted infinite-horizon average reward accept individuals with positive probability* $\Pi(\mu,\nu) > 0$ *in all belief states* (μ,ν).

Proof. In each belief state (μ,ν), there are two choices for the acceptance probability $\pi = \Pi(\mu,\nu)$: either stop ($\pi = 0$) with zero reward $V^{\Pi}(\mu,\nu) = 0$, or accept with some positive probability, in which case $V^{\Pi}(\mu,\nu)$ is given by (35). To determine the optimal action by dynamic programming, we assume that an optimal policy is used from state $\nu+1$ onward, thus replacing $V^{\Pi}(\cdot,\nu+1)$ by $V^*(\cdot,\nu+1)$ on the right-hand side of (35). It follows that $\pi > 0$ is optimal if this right-hand side is non-negative. This in turn is true if $V^*(\mu,\nu+1)$ is convex in μ and non-negative, since Jensen's inequality would imply

$$V^{\Pi}(\mu,\nu) = \mu V^*\left(\frac{\mu\nu+1}{\nu+1}, \nu+1\right) + (1-\mu)V^*\left(\frac{\mu\nu}{\nu+1}, \nu+1\right) \geq V^*(\mu,\nu+1) \geq 0. \tag{36}$$

It is now shown by induction over decreasing ν that $V^*(\mu,\nu)$ is convex in μ and non-negative for all ν, implying by the previous argument that $\Pi(\mu,\nu) > 0$ is optimal for all states. More precisely, we again consider approximations $V^N(\mu,\nu)$ to $V^*(\mu,\nu)$, initialized by setting $V^N(\mu,N+1) = V^*(\mu,\infty) = \max\{\mu - c, 0\}$. By taking $N \to \infty$, the proof extends to optimal policies.

The base case $\nu = N+1$ is simply given by the initialization $V^N(\mu,N+1) = \max\{\mu - c, 0\}$, since this is a convex and non-negative function. It then suffices to establish convexity for $\nu = N, N-1, \ldots$ since (36) would then show that $V^N(\mu,\nu)$ is non-increasing in ν, not just non-negative. This inductive step corresponds exactly with the proof of convexity of the function $\bar{V}(\mu,\nu)$ in the proof of Proposition 2 (Appendix A.2.1). □

Theorem 7 shares a similar spirit with the exploring policies in (Kilbertus et al., 2020), which assign positive acceptance probability to all subsets of $\mathcal{X}$ with positive probability under $p(x)$. It clearly contrasts with Theorem 1 for the case of discounted total reward, where stopping sets are optimal.

Theorem 7 however does not provide further guidance on selecting a policy. It does not even distinguish between an always-accept policy $\Pi(\mu,\nu) \equiv 1$ and a stochastic one, $\Pi(\mu,\nu) = \pi \in (0,1)$, which spends a geometrically distributed amount of time in state ν before eventually moving to $\nu+1$. Intuitively, this seems to be because the lack of a discount factor means that any short-term cost incurred in learning the parameter ρ is trumped by eventual long-term reward. Indeed, one can conceive of the following two-phase N-step policy (with $N \to \infty$): The first "explore" phase learns ρ using a number of samples N_1 that increases to infinity but sublinearly in N, for example using the always-accept policy $\Pi(\mu,\nu) \equiv 1$. The second $(N - N_1)$-step phase simply "exploits" this knowledge using the threshold policy $\mathbb{1}(\mu > c)$. Future work could consider the analysis of these and similar policies.

D. Additional Experimental Details and Results

D.1. Re-Implementations and Modifications of Baselines

D.1.1. Consequential Learning (CL, CLVW)

As mentioned in Section 6.1, due to unavailabilty of code, the CL algorithm (Kilbertus et al., 2020, Alg. 1) was re-implemented for the case of no fairness penalty ($\lambda = 0$) and policy updates after every observation ($N = 1$). The latter also implies no mini-batches ($B = M = 1$).

Given $N = 1$, the update to the policy parameters θ has a simple form. From a combination of Kilbertus et al. (2020, eq. (9), (10)) (and using their notation), the stochastic gradient approximation is given by

$$\begin{aligned}\nabla_{\theta_t} v_P(\pi_{\theta_t}) = \nabla_{\theta_t} u\left(\pi_{\theta_t}, \pi_{\theta_{t-1}}\right) &\approx \frac{y_i - c}{\pi_{\theta_{t-1}}(D = 1 \mid x_i)} \mathbb{E}_{D \sim \pi_{\theta_t}}\left[D \nabla_{\theta_t} \log \pi_{\theta_t}(D \mid x_i)\right] \\ &= \frac{y_i - c}{\pi_{\theta_{t-1}}(D = 1 \mid x_i)} \nabla_{\theta_t} \pi_{\theta_t}(D = 1 \mid x_i).\end{aligned}$$

The last equality is due to D being Bernoulli (with parameter $\pi_{\theta_t}(D = 1 \mid x_i)$) and contrasts with Kilbertus et al. (2020), who sample D instead of evaluating the expectation. Substituting in the gradient of the logistic policy from Kilbertus et al. (2020),

$$\nabla_{\theta_t} v_P(\pi_{\theta_t}) \approx \frac{y_i - c}{\pi_{\theta_{t-1}}(D = 1 \mid x_i)} \pi_{\theta_t}(D = 1 \mid x_i)\left(1 - \pi_{\theta_t}(D = 1 \mid x_i)\right) x_i.$$

Since θ_t is initialized to θ_{t-1}, we have

$$\nabla_{\theta_t} v_P(\pi_{\theta_{t-1}}) \approx (y_i - c)\left(1 - \pi_{\theta_{t-1}}(D = 1 \mid x_i)\right) x_i$$

and hence the update

$$\theta_t = \theta_{t-1} + \alpha(y_i - c)\left(1 - \pi_{\theta_{t-1}}(D = 1 \mid x_i)\right) x_i. \tag{37}$$

A similar expression holds for the semi-logistic policy from (Kilbertus et al., 2020). However, only the results of the logistic policy are reported as it is found to be better than semi-logistic.

To initialize the policy, Kilbertus et al. (2020) prescribe training a logistic predictive model for Y on fully labelled examples. Herein this is done on the B_0 initial training examples and also in an online fashion via the stochastic gradient update

$$\theta_t = \theta_{t-1} + \alpha\left(y_i - \pi_{\theta_{t-1}}(D = 1 \mid x_i)\right) x_i. \tag{38}$$

The updates in (37) and (38) are “plain” stochastic gradient updates, whereas VW uses a more sophisticated algorithm. For this reason, a VW version of Consequential Learning (abbreviated CLVW) was also implemented. Update (38) is replaced by VW’s update for a standard logistic regression model. For (37), its similarity with (38) is exploited by defining sample weights

$$w_i = \begin{cases} c\frac{1 - \pi_{\theta_{t-1}}(D=1 \mid x_i)}{\pi_{\theta_{t-1}}(D=1 \mid x_i)}, & y_i = 0, \\ 1 - c, & y_i = 1 \end{cases}$$

to make (37) equivalent to weighted logistic regression with weights w_i.

Modifications for finite domains: The core CL algorithm applies with no changes after one-hot encoding X and instantiating parameters θ_x for each value $x \in \mathcal{X}$. The initialization of θ can be simplified by observing that logistic regression on the B_0 initial training examples corresponds to equating the predicted probability at x, $1/(1 + e^{-\theta_x})$, with the empirical success probability, $\mu_{B_0}(x)$. Inverting this equation, θ_x is thus initialized to $\log(\mu_{B_0}(x)/(1 - \mu_{B_0}(x)))$.

D.1.2. RegCB-Optimistic (R-o, R-osl)

The VW implementation of RegCB-Optimistic (R-o) did not perform as well in experiments on the FICO and COMPAS datasets (see Appendix D.5). It is believed that this is due to not taking advantage of the assumption in the selective labels problem that rejection has zero cost. Thus a specialized version of R-o that does exploit this assumption was implemented, referred to as R-osl. As discussed in Section 6.1, due to the zero-cost assumption for rejection, the decision policy reduces to determining whether the UCB on the estimated conditional mean $\hat{\mu}(x)$ exceeds the threshold c.

To compute the UCB on $\hat{\mu}(x)$, the BINSEARCH algorithm in Foster et al. (2018, Alg. 3) is used with the following modifications (in the notation of Foster et al. (2018)):

1. Only acceptances are considered from the history H (line 3), i.e. a' is restricted to 1 and $a = 1$ (line 4). The regressor $f(x', 1)$ (which corresponds to $\hat{\mu}(x)$) is trained to predict the outcomes y_i, i.e. $r' = y_i$ and the loss function is
$$R(f) = \sum_{i:a_i=1} (f(x_i, 1) - y_i)^2. \tag{39}$$
2. The algorithm is terminated as soon as it is known whether the UCB exceeds c. Thus at the beginning, if the current regressor f predicts $f(x, 1) > c$, then the decision is to accept since the UCB can only be larger. More generally, the algorithm terminates with acceptance when the lower bound z_L on the UCB exceeds c for the first time, and terminates with rejection when the upper bound z_H on the UCB falls below c.

Modifications for finite domains: R-o was not run in the finite-domain experiments due to its poorer performance. For R-osl, the conditional mean $\mu(x)$ can be computed directly without resorting to a model, and the UCB on $\mu(x)$ has a simple closed form so that BINSEARCH is not needed. Indeed, it can be seen from (39) that $R(f)$ is minimized by $f(x, 1) = \mu(x) = \sigma(x)/\nu(x)$, i.e., the empirical mean conditioned on x where $\sigma(x)$ is the number of successes and $\nu(x)$ the number of acceptances. Furthermore, it is straightforward to show that the excess loss of a regressor f compared to $\mu(x)$ is
$$R(f) - \min_f R(f) = \sum_{x\in\mathcal{X}} \nu(x)(f(x, 1) - \mu(x))^2.$$
Given a confidence parameter C_0, the UCB at x is defined by Foster et al. (2018) as
$$\max \ f(x, 1) \quad \text{s.t.} \quad R(f) - \min_f R(f) \le C_0.$$
The maximizing solution is to set $f(x', 1) = \mu(x')$ for $x' \neq x$ and maximize $f(x, 1)$ subject to $\nu(x)(f(x, 1) - \mu(x))^2 \le C_0$. This yields
$$\mu(x) + \sqrt{\frac{C_0}{\nu(x)}} \tag{40}$$
as the UCB on $\mu(x)$. The policy is thus to accept if (40) is greater than c and reject otherwise.

D.2. Supervised Learning Methods

All of the compared policies rely on supervised learning methods for various purposes and these were trained online using VW. The greedy and homogeneous policies use online logistic regression, both for the conditional mean estimator $\hat{\mu}(x)$ as well as the bootstrap estimators $\hat{\mu}_1(x), \dots, \hat{\mu}_K(x)$. CL and CLVW perform online updates as described in Section D.1.1. For the VW contextual bandit algorithms (ϵG, B-g, C-nu), the method for policy learning was set to doubly robust, except for R-o which uses regression with the squared loss. R-osl also uses online linear regression with the squared loss, both to update the regressor $f(x, 1)$ as acceptances are made, and to temporarily update the regressor with virtual examples of varying weights w in the BINSEARCH algorithm (lines 6, 7, 12).

Only linear models (linear or logistic regression) were tested in the experiments herein, as also done in Kilbertus et al. (2020). This is by no means a limitation however as nonlinear models can be substituted straightforwardly, especially those with efficient online algorithms.

For the finite-domain case, the conditional means $\mu(x)$ can be well-estimated by the respective empirical means and a model is not needed. Following the Bayesian approach in Section 3 of placing a beta prior on the success probability $\rho(x)$, these empirical means are slightly modified by adding pseudo-counts from the prior. Specifically, the conditional mean estimate before individual i arrives is $\mu_i(x) = \sigma_i(x)/\nu_i(x)$, where the initial numbers of pseudo-successes $\sigma_0(x) = 1$ and pseudo-acceptances $\nu_0(x) = 2$ correspond to a Beta$(1, 1)$ prior, and $\sigma_i(x)$ and $\nu_i(x)$ are the numbers of successes and acceptances observed at x plus the pseudo-counts. The above method of updating $\mu(x)$ is used for all policies. In particular for TS, the sampling probability $\Pr(\rho_i(x) > c)$ assumes that $\rho_i(x)$ follows a beta distribution with parameters $\sigma_i(x)$ and $\nu_i(x) - \sigma_i(x)$.

D.3. Algorithm Parameters and Tuning

Algorithm-specific parameters Tables 1 and 2 list the parameters specific to each algorithm and the values explored in the experiments. Fewer parameters are needed for the finite-domain experiments (Table 1), since e.g. the optimal policies (O-t, O-e, O-u) compute or estimate effective discount factors $\bar{\gamma}(x)$ themselves, bagging is replaced by Thompson sampling which is parameter-free, and for R-osl, the BinSearch algorithm is not needed. For the contextual bandit algorithms in VW (ϵG, B-g, C-nu, R-o), the parameter ranges are the same as in Bietti et al. (2020). All other parameters are kept at their defaults except for the online learning rate discussed below. For parameters with multiple values listed, the results shown in Section 6 and Appendix D.5 are optimized with respect to these values, except for plots depicting sensitivity to these parameters. Each algorithm has zero or one of these tuning parameters, except for C-nu which has two (N_p, ψ).

Table 1. Algorithm-specific parameters and ranges for finite-domain experiments.

Algorithm	Parameter	Values
Optimal (O-t, O-e, O-u)	order N of approximation $V^N(\mu, \nu)$	1000
Consequential Learning (CL)	learning rate α	$0.1, 0.3, 1, 3, 10, 30, 100$
ϵ-Greedy (ϵG)	exploration probability ϵ	$0.01, 0.02, 0.05$
RegCB-Optimistic Selective Labels (R-osl)	confidence width C_0	$0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1$

Table 2. Algorithm-specific parameters and ranges for FICO and COMPAS experiments.

Algorithm	Parameter	Values
Homogeneous (H)	discount factor $\bar{\gamma}$	$0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8, 0.5$
	number of bootstrap estimators K	10
	order N of approximation $V^N(\mu, \nu)$	1000
Consequential Learning (CL)	learning rate α	$0.05, 0.1, 0.2, 0.5, 1$
Consequential Learning VW (CLVW)	learning rate α	$1, 2, 5, 10, 20, 50$
ϵ-Greedy (ϵG)	exploration probability ϵ	$0.01, 0.02, 0.05$
Greedy Bagging (B-g)	number of policies N_p	$4, 8, 16$
Online Cover, no uniform exploration (C-nu)	number of policies N_p	$4, 8, 16$
	ψ	$0.01, 0.1, 1$
RegCB-Optimistic (R-o)	confidence width C_0	$0.001, 0.01, 0.1$
RegCB-Optimistic Selective Labels (R-osl)	confidence width C_0	$0.005, 0.01, 0.02, 0.05, 0.1$
	BinSearch precision α	0.01

Online learning rate For the real-data experiments (FICO, COMPAS), all policies (including the greedy policy) make use of online learning as discussed in Section D.2. The learning rate α is tuned over the range $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$, except for CL and CLVW, which have their own α ranges shown in Table 2. The results shown in Section 6 and Appendix D.5 always correspond to the best of these learning rates.

Initial training size B_0 Total reward is plotted as a function of the initial training size B_0 in Section 6 and Appendix D.5. For the finite-domain experiments, $B_0 \in \{0, 1, 2, 5, 10\} \times |\mathcal{X}|$. For the FICO and COMPAS experiments, $B_0 \in \{1, 10, 20, 30, 40, 50\}$, except for the greedy policy for which the range is extended to 100 to locate where the total reward saturates.

D.4. Data Pre-Processing

FICO First, 588 rows with all entries missing (values of -9) are removed. For the outcome variable, “Good” and “Bad” values are encoded as $y_i = 1$ and $y_i = 0$ respectively. For the `MSinceMostRecentDelq` feature, special values of -7 appear to mean more than 84 months (7 years) since the most recent delinquency, based on correlation with the outcome variable. These -7 values are thus replaced by the maximum value in the data plus 1. Similarly for the `MSinceMostRecentInqexcl7days` feature, values of -8 appear to mean more than 24 months since the most recent credit inquiry and are replaced by the maximum plus 1. On the other hand, `MSinceMostRecentInqexcl7days` values of -7 appear to mean an inquiry within the last 7 days and are thus replaced by 0 (months). For `MaxDelq2PublicRecLast12M`, values greater than 7 (other) were imputed as 7 (current and never delinquent) based on the corresponding values in `MaxDelqEver`. All remaining special values of -8 and -9 are imputed with the mean of the corresponding feature.

COMPAS The same CSV file and filtering of rows are used as in ProPublica's analysis (Angwin et al., 2016). Features included are as follows: demographics `sex`, `age`, `age_cat`, criminal history `priors_count`, `juv_fel_count`, `juv_misd_count`, `juv_other_count`, current charge degree `c_charge_degree`, and COMPAS score `decile_score`, `score_text`. Like in Kilbertus et al. (2020), race was not used. The outcome variable is `two_year_recid`, where re-arrest is encoded as $y_i = 0$ since it is a negative outcome.

After the above pre-processing is done, categorical features are one-hot encoded. Then all features are standardized, as this was found to improve online learning performance.

D.5. Additional Results

For the finite-domain experiments discussed in Section 6.2, Figures 5 and 6 display the complete set of results for $|\mathcal{X}| \in \{2, 3, 4, 5, 7, 10\}$ and $c \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. Discounted total rewards (discount factor $\gamma = 0.999$) are plotted as a function of the initial training size B_0, as in the first two rows of Figure 2. The overall patterns are the same as in Figure 2.

For the experiments on the FICO and COMPAS datasets discussed in Section 6.3, Figure 7 is the same as Figure 3 except that it also includes R-o, which is seen to be less competitive than the other methods.

Figure 8 shows discounted total rewards computed with discount factor $\gamma = 0.9995$, as opposed to $\gamma = 1$ in Figures 3 and 7. $\gamma = 0.9995$ is chosen so that the sum of truncated discount weights is less than 1% of the total sum for FICO and 5% for COMPAS, similar to the finite-domain experiments in Section 6.2. The curves have more of a downward tilt as functions of B_0. This is expected because of the slightly greater value placed on early time indices, and hence slightly greater cost to exploration. Otherwise the conclusions are the same as in Figures 3 and 7.

Figure 9 plots discounted total rewards on the COMPAS dataset with threshold $c = 0.8$. In this case, the curves for the greedy (G), homogeneous (H), and R-osl policies coincide, so no policy outperforms greedy.

Table 3 lists the best parameter values corresponding to the results shown in Figures 3, 7, 8, and 9.

Table 3. Best parameter values in FICO and COMPAS experiments. α refers to the learning rate and the rejection pass is discussed in Section 6.1. See Table 2 for definitions of the other parameters.

Algorithm	Parameter	Best values					
		FICO		COMPAS $c = 0.6$		COMPAS $c = 0.8$	
		$\gamma = 1$	$\gamma = 0.9995$	$\gamma = 1$	$\gamma = 0.9995$	$\gamma = 1$	$\gamma = 0.9995$
Greedy (G)	α	0.5	0.01	0.5	0.5	0.01	0.01
Homogeneous (H)	$\bar{\gamma}$	0.99	0.99	0.95	0.95	0.5	0.5
	α	1	0.5	0.5	0.5	0.01	0.01
CL	α	0.2	0.5	0.5	0.5	0.5	0.5
CLVW	α	5	5	5	5	2	5
ϵ-Greedy (ϵG)	ϵ	0.01	0.01	0.01	0.01	0.05	0.05
	α	0.05	0.05	0.02	0.02	0.1	0.1
	rejection pass	yes	yes	yes	yes	no	no
Bagging (B-g)	N_p	4	4	4	4	4	4
	α	0.05	0.1	0.1	0.05	0.2	0.2
	rejection pass	yes	no	no	no	no	no
Cover (C-nu)	N_p	4	4	4	4	4	4
	ψ	0.1	0.01	0.1	0.01	0.01	0.01
	α	0.02	0.02	0.1	0.1	1	1
	rejection pass	yes	yes	no	no	no	no
R-o	C_0	0.01	0.1	0.001	0.1	0.001	0.01
	α	0.05	0.05	0.05	0.02	0.1	0.1
	rejection pass	yes	no	no	no	yes	no
R-osl	C_0	0.02	0.02	0.02	0.02	0.1	0.1
	α	0.2	0.2	0.2	0.2	0.02	0.02

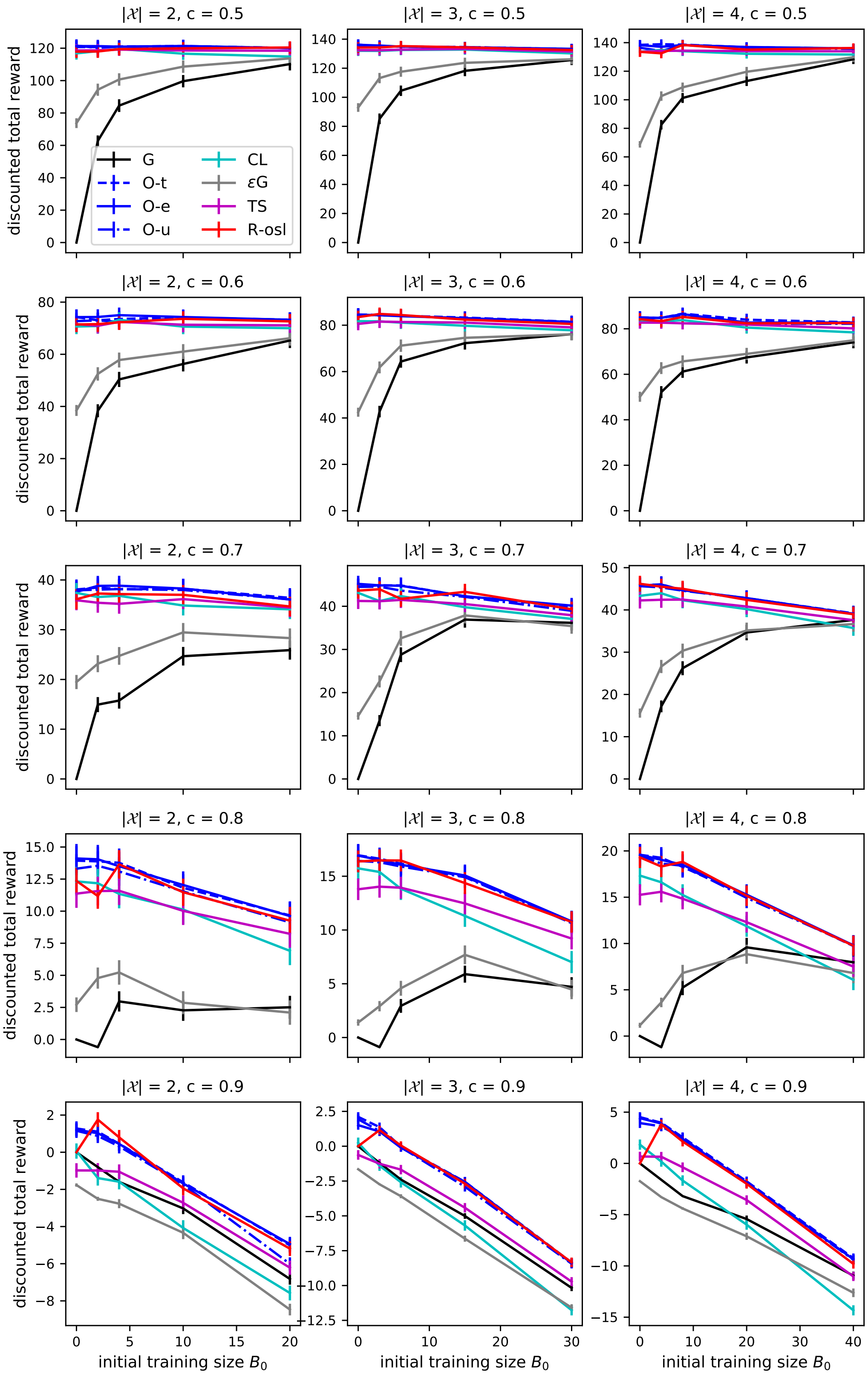


Figure 5. Discounted total rewards (discount factor $\gamma = 0.999$) on finite domains $\mathcal{X}$ of cardinality $2, 3, 4$.

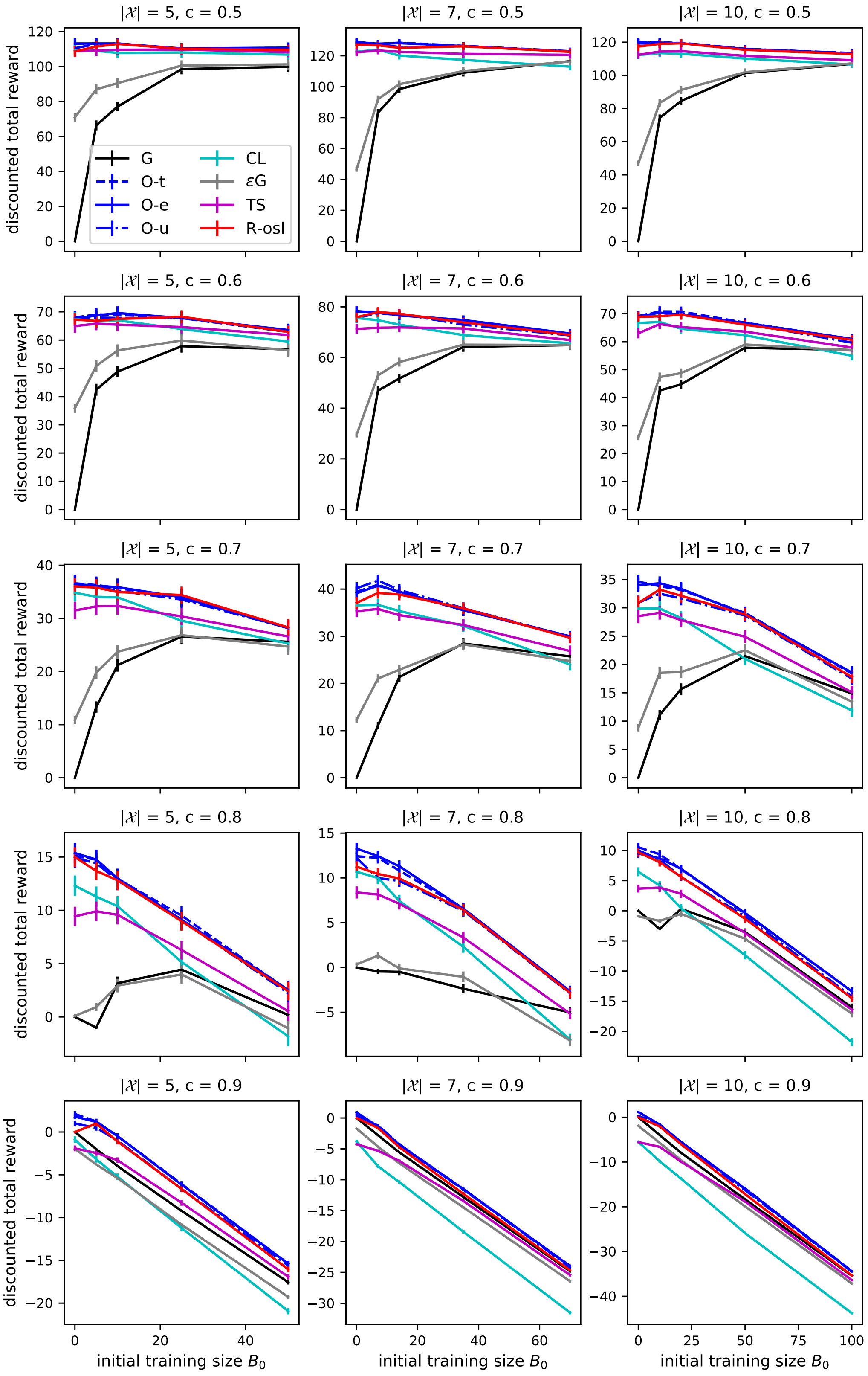
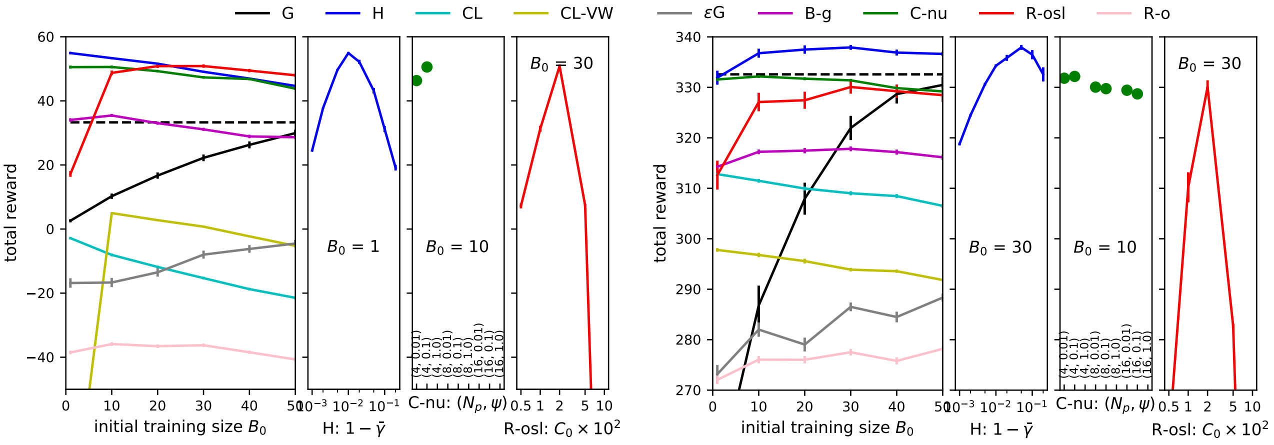


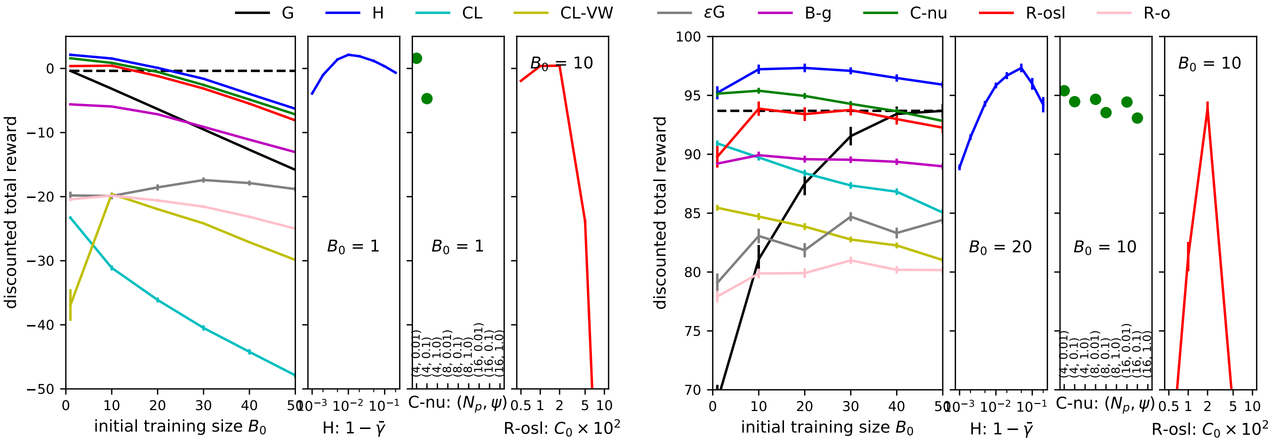
Figure 6. Discounted total rewards (discount factor $\gamma = 0.999$) on finite domains $\mathcal{X}$ of cardinality $5, 7, 10$.



(a) FICO, $c = 0.8$

(b) COMPAS, $c = 0.6$

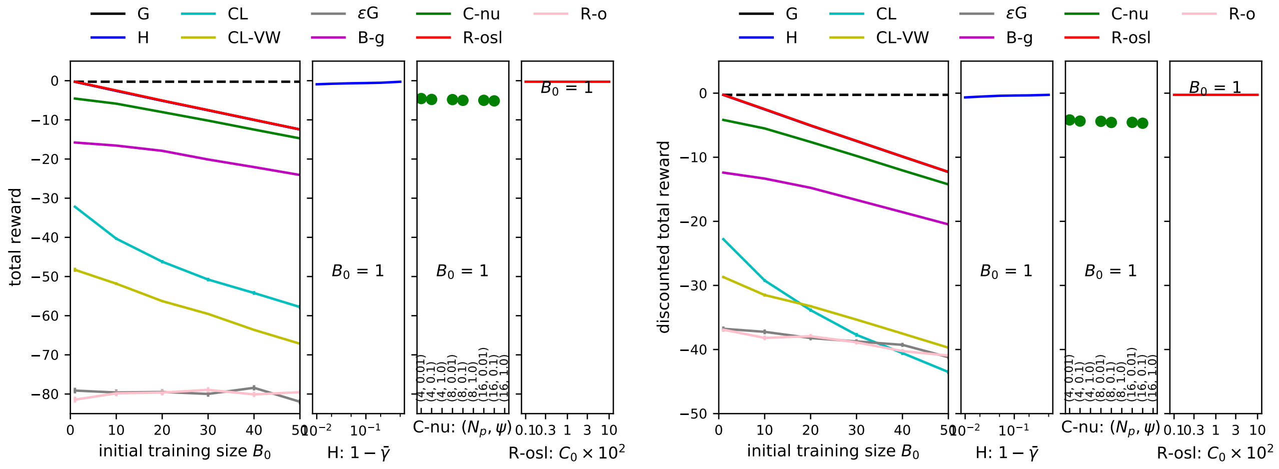
Figure 7. Total rewards (discount factor $\gamma = 1$) on real-world datasets. The dashed black line indicates the maximum total reward achieved by the greedy policy (G) at $B_0 = 80$ for FICO and $B_0 = 90$ for COMPAS.



(a) FICO, $c = 0.8$

(b) COMPAS, $c = 0.6$

Figure 8. Discounted total rewards (discount factor $\gamma = 0.9995$) on real-world datasets. The dashed black line indicates the maximum total reward achieved by the greedy policy (G) at $B_0 = 1$ for FICO and $B_0 = 50$ for COMPAS.



(a) $\gamma = 1$

(b) $\gamma = 0.9995$

Figure 9. Discounted total rewards on COMPAS dataset with $c = 0.8$. The curves for greedy (G), homogeneous (H), and R-osl coincide.