Optimistic Feasible Search for Closed-Loop Fair Threshold Decision-Making

Du Wenzhang
Dept. of Computer Engineering
Mahanakorn University of Technology, International College (MUTIC)
Bangkok, Thailand
dqswordman@gmail.com

***Abstract*—Closed-loop decision-making systems (e.g., lending, screening, or recidivism risk assessment) often operate under fairness and service constraints while simultaneously inducing feedback effects: decisions change who appears in the future, yielding non-stationary data and potentially amplifying disparities. We study online learning of a one-dimensional threshold policy from bandit feedback under demographic parity (DP) and (optionally) service-rate constraints. The learner observes only a scalar score each round and selects a threshold; reward and constraint signals are revealed for the chosen threshold only.**

We propose Optimistic Feasible Search (OFS), a simple grid-based method that maintains confidence bounds for reward and constraint residuals for each candidate threshold. At each round, OFS selects the threshold that appears feasible under confidence bounds and, among those, maximizes optimistic reward; if no threshold appears feasible, OFS selects the threshold minimizing optimistic violation. This design directly targets feasible high-utility thresholds and is particularly effective for low-dimensional, interpretable policy classes where discretization is natural.

We evaluate OFS on (i) a synthetic closed-loop benchmark with stable contraction dynamics and (ii) two semi-synthetic closed-loop benchmarks grounded in German Credit and COMPAS, constructed by training a score model and feeding group-dependent acceptance decisions back into population composition. Across all environments, OFS achieves substantially higher reward with dramatically smaller cumulative constraint violation than unconstrained and primal–dual bandit baselines. Moreover, OFS is near-oracle: its steady-state reward gap to the best feasible fixed threshold is below 0.003 across benchmarks. All experiments are reproducible and organized with double-blind friendly relative outputs.

***Index Terms*—Trustworthy and Reliable AI, Reinforcement Learning, Ethics and Regulation in AI, Interpretable and Explainable AI**

I. Introduction

Algorithmic decision systems increasingly mediate access to opportunities and resources in high-stakes settings such as lending and criminal justice. A core difficulty is that these systems are closed-loop: decisions affect the future population that the system will encounter (selection effects, behavioral responses, or measurement shifts), creating non-stationarity and potentially reinforcing disparities. At the same time, regulations and organizational policies often require fairness constraints such as demographic parity (DP) [1]–[3] or related notions [4], [5].

This paper focuses on a minimal but practically relevant closed-loop learning problem: a decision-maker observes a scalar score (e.g., a model logit) and chooses a threshold θ (accept if score $\geq \theta$). Threshold policies are widely used for their transparency and ease of auditing. However, learning a good threshold online under constraints is challenging because the environment is non-stationary and feedback is bandit-style.

Goal. We aim to maximize long-run utility while keeping constraint violations small, using only bandit feedback. We specifically target the regime where the policy class is low-dimensional (here 1D thresholds), so a discrete search strategy with rigorous uncertainty estimates is viable.

Contributions.

- **Method:** We propose Optimistic Feasible Search (OFS), an optimism-based feasible screening method for constrained closed-loop threshold learning, inspired by optimism in bandits [6]. OFS is simple, interpretable, and works directly on discretized thresholds (Section III).
- **Benchmarks:** We provide a synthetic closed-loop benchmark (MVE-S) with stable contraction dynamics and two semi-synthetic benchmarks derived from German Credit and COMPAS via score-model-based logits and feedback-induced population shifts (Section V).
- **Evidence:** Using a unified experimental suite and fully reproducible relative outputs, OFS dominates strong baselines in reward and cumulative violations and is near-oracle across all benchmarks (Section VI, Tables I, II and Figs. 1–3).

Related context. Our work connects constrained online learning [7]–[9] and bandit/zeroth-order optimization [10]–[12]. It is also motivated by growing recognition of performative or feedback-driven effects in prediction and decision pipelines [13]. Compared to many fairness-aware learning approaches that assume i.i.d. data [3], we emphasize closed-loop non-stationarity and provide benchmark constructions to study it.

II. Problem Setup

At each round $t = 1, \ldots, T$, the environment produces a batch of n individuals (we use n=256 in experiments) with a sensitive attribute $a \in \{0, 1\}$, a scalar score s, and an outcome label $y \in \{0, 1\}$. The decision-maker selects a threshold policy

$$\pi_\theta(s) = \mathbb{I}\{s \geq \theta\}, \tag{1}$$

yielding decisions $d \in \{0, 1\}$ (accept/reject). A bandit feedback signal is observed: reward and constraint residuals for the chosen threshold.

A. Utility

We consider a standard acceptance utility with false-positive cost $c_{\text{fp}} > 0$:

$$r_t(\theta) = \mathbb{E}\big[d \cdot (y - c_{\text{fp}}(1 - y))\big], \tag{2}$$

estimated by the batch mean. This captures the trade-off between approving beneficial cases (y=1) and incurring a cost when approving harmful cases (y=0).

B. Constraints

We consider two classes of constraints, expressed via residuals $g(\theta)$ where feasibility means $g(\theta) \leq 0$.

Demographic Parity (DP). Let group-wise acceptance rates be

$$\text{acc}_g(\theta) = \mathbb{P}(d = 1 \mid a = g), \quad g \in \{0, 1\}. \tag{3}$$

The DP gap is $\Delta_{\text{DP}}(\theta) = |\text{acc}_0(\theta) - \text{acc}_1(\theta)|$. We enforce

$$g_{\text{dp}}(\theta) = \Delta_{\text{DP}}(\theta) - \varepsilon \leq 0. \tag{4}$$

Service-rate constraint. To avoid degenerate always-reject or always-accept solutions, we optionally enforce an overall acceptance-rate window:

$$\text{acc}(\theta) = \mathbb{P}(d = 1) \in [\alpha_{\min}, \alpha_{\max}], \tag{5}$$

encoded as two residuals $g_{\text{srv,low}}(\theta) = \alpha_{\min} - \text{acc}(\theta)$ and $g_{\text{srv,high}}(\theta) = \text{acc}(\theta) - \alpha_{\max}$.

C. Closed-loop dynamics and performance metrics

Crucially, the environment is closed-loop: the distribution of (s, y, a) at time t depends on previous decisions, creating non-stationarity. We evaluate:

- **Tail reward / tail DP gap:** mean over the last K_{tail} iterations (K_{tail}=200), then averaged across random seeds.
- **Cumulative violation:** sum over t of the positive parts of residuals,

$$V_T = \sum_{t=1}^{T} \sum_{j} \max\{g_{j,t}, 0\}. \tag{6}$$

  This metric penalizes persistent or repeated constraint violations and is robust to stochastic fluctuations.

We also report bootstrap confidence intervals (CIs) across seeds for key differences (Table III).

III. Optimistic Feasible Search (OFS)

OFS targets low-dimensional policy classes where discretization is natural. Let $\{\theta_i\}_{i=1}^{K_\theta}$ be a uniform grid on $[\theta_{\min}, \theta_{\max}]$. For each grid point, OFS maintains empirical means of reward and constraint residuals along with a confidence radius.

Algorithm 1 Optimistic Feasible Search (OFS)

Input: grid $\{\theta_i\}_{i=1}^{K_\theta}$, confidence δ, exploration ϵ_{exp}
Initialize $n_i \leftarrow 0$, $\hat{r}_i \leftarrow 0$, $\hat{g}_{i,j} \leftarrow 0$ for all i, j
for $t = 1$ to T **do**
  Compute $b_i(t) = \sqrt{\frac{c \log(((t+1)K_\theta)/\delta)}{\max(1, n_i)}}$ for all i
  $\mathcal{F}_t \leftarrow \{i : \hat{g}_{i,j} + b_i(t) \leq 0 \ \forall j\}$
  With prob. ϵ_{exp}, pick i_t uniformly in $\{1, \ldots, K_\theta\}$ (explore)
  Else if $\mathcal{F}_t \neq \emptyset$, pick $i_t \in \arg\max_{i \in \mathcal{F}_t} (\hat{r}_i + b_i(t))$
  Else pick $i_t \in \arg\min_i \sum_j \max(\hat{g}_{i,j} + b_i(t), 0)$
  Play threshold θ_{i_t}; observe reward r_t and residuals $g_{t,j}$
  Update $n_{i_t} \leftarrow n_{i_t} + 1$ and empirical means $\hat{r}_{i_t}, \hat{g}_{i_t,j}$
end for

A. Algorithm

Let $\hat{r}_i(t)$ be the empirical mean reward for θ_i after it has been played $n_i(t)$ times. Similarly, let $\hat{g}_{i,j}(t)$ denote the empirical mean of residual j. We define a confidence radius

$$b_i(t) = \sqrt{\frac{c \, \log\left(\frac{(t+1)K_\theta}{\delta}\right)}{\max(1, n_i(t))}}, \tag{7}$$

with constant $c > 0$ and confidence $\delta \in (0, 1)$.

At round t, define the optimistic feasible set

$$\mathcal{F}_t = \Big\{ i : \hat{g}_{i,j}(t) + b_i(t) \leq 0, \ \forall j \Big\}. \tag{8}$$

OFS selects:

- If $\mathcal{F}_t \neq \emptyset$: choose $i_t \in \arg\max_{i \in \mathcal{F}_t} \big(\hat{r}_i(t) + b_i(t)\big)$.
- Else: choose $i_t \in \arg\min_i \sum_j \max\big(\hat{g}_{i,j}(t) + b_i(t), 0\big)$.

We also use a small ϵ_{exp}-greedy exploration probability to ensure coverage.

B. Discussion and intuition

OFS differs from standard primal–dual approaches in a crucial way: instead of optimizing a Lagrangian with dual multipliers that can oscillate under non-stationarity, OFS explicitly screens feasibility using confidence bounds and focuses search on thresholds that are plausibly feasible. When feasibility is currently uncertain, OFS naturally allocates samples to reduce uncertainty in regions that are near the boundary.

C. Theoretical perspective (sketch)

A full closed-loop analysis can be complex, but in our benchmarks the dynamics are designed to be stable and contractive. This yields a useful perspective: each fixed threshold induces a well-defined steady-state distribution, and samples gathered when repeatedly choosing a threshold concentrate around its steady-state means.

Assumption (contractive closed-loop). For each fixed threshold θ_i, the environment state evolves via a contraction mapping with a unique stationary distribution, and observations have bounded noise (sub-Gaussian) around steady-state means. This is satisfied by our synthetic benchmark by construction and approximately captured by the clipped, stable update rules in our semi-synthetic benchmark.

Proposition (informal). Under the above assumption, with high probability $1-\delta$, OFS achieves (i) sublinear cumulative violation $V_T = \tilde{O}(\sqrt{TK_\theta})$ and (ii) near-optimal reward relative to the best feasible grid threshold, up to standard statistical and discretization errors. The proof follows a standard optimism argument: confidence bounds hold uniformly over arms; OFS selects an arm whose optimistic feasibility implies true feasibility once sampled sufficiently; when infeasible, OFS minimizes optimistic violation, driving exploration toward feasibility.

We emphasize that our experiments validate these behaviors empirically and show that OFS is consistently near-oracle (Table II).

IV. Baselines

We compare against two strong bandit baselines commonly used for constrained online optimization:

Unconstrained bandit gradient descent (U). We apply a one-point zeroth-order gradient estimator [10] to the loss $-r_t(\theta)$, ignoring constraints.

Primal–Dual bandit gradient descent (PD). We optimize a Lagrangian with dual ascent on constraint residuals [8], [9]. We include a small PD hyperparameter grid for the semi-synthetic tasks and report a tuning table to address baseline sensitivity concerns (Table IV).

V. Closed-loop Benchmarks

A. Synthetic benchmark (MVE-S)

We use a synthetic closed-loop environment where group-specific score distributions evolve with acceptance decisions. At each round, a group $a \in \{0,1\}$ is sampled and the score is Gaussian with mean $\mu_a(t)$. Labels are generated via a sigmoid model. The decision threshold affects group acceptance rates, which then update the state (μ_0, μ_1) through a stable contraction dynamics. This benchmark isolates closed-loop feedback in a controlled setting while keeping the policy class (thresholds) simple.

B. Semi-synthetic benchmarks (German Credit and COMPAS)

To ground our closed-loop study in real tabular datasets, we construct semi-synthetic environments using German Credit and COMPAS [14], [15]. In German, we use sex as the sensitive attribute and define $y = 1$ as good credit; in COMPAS, we use sex and define $y = 1$ as non-recidivism. We preprocess each dataset with one-hot encoding and standardization, then train a simple logistic score model to produce scalar logits s. Closed-loop dynamics are induced by composition shift: for each group we maintain a mixture over a "high-score" pool and a "low-score" pool, and the mixture weight is updated as a stable function of group-wise acceptance rates. Intuitively, higher acceptance can increase the fraction of "high-score" individuals observed in the future, capturing selection effects while remaining simple and reproducible.

VI. Experiments

A. Protocol and implementation details

We run 10 random seeds for each main experiment. We report tail means over the last K_{tail}=200 iterations per seed, then report mean±std across seeds (Table I). Cumulative violation sums positive parts of constraint residuals over the full horizon (DP-only for MVE-S; DP + service constraints for German/COMPAS). We also report bootstrap CIs (Table III). Oracle tradeoff curves are computed by evaluating fixed thresholds over long horizons and selecting the best feasible point (Fig. 3, Table II).

B. Main results

Table I summarizes tail reward, tail DP gap, and cumulative violations on all three benchmarks. OFS consistently achieves the highest reward while keeping DP gaps well below the required ε and dramatically reducing cumulative violations. On the semi-synthetic tasks, unconstrained learning incurs severe violations (due to both DP and service constraints), while PD improves utility but still exhibits substantially larger cumulative violation than OFS.

C. Learning dynamics on MVE-S

Fig. 1 shows representative learning curves on the synthetic benchmark. OFS rapidly finds a feasible high-utility threshold: DP gap decreases and cumulative violations grow slowly, while reward remains high. In contrast, the unconstrained baseline improves utility only when it allows large DP violations, and PD tends to trade reward for partial constraint control.

D. Semi-synthetic learning dynamics

We next evaluate on German Credit and COMPAS semi-synthetic closed-loop environments. Fig. 2 shows that OFS keeps DP gaps small and service violations low, yielding far smaller cumulative violation while achieving high reward. PD can achieve strong reward but exhibits substantially larger violation, especially under service constraints, which is precisely what our cumulative-violation metric captures.

E. Near-oracle performance

To assess absolute optimality, we compute an oracle tradeoff curve by sweeping fixed thresholds and selecting the best feasible point. Table II reports oracle best feasible reward and the steady-state suboptimality gaps. Across all benchmarks, OFS is near-oracle: its reward gap to the best feasible fixed threshold is below 0.003. Fig. 3 visualizes the reward–DP tradeoff curves and highlights the feasible region.

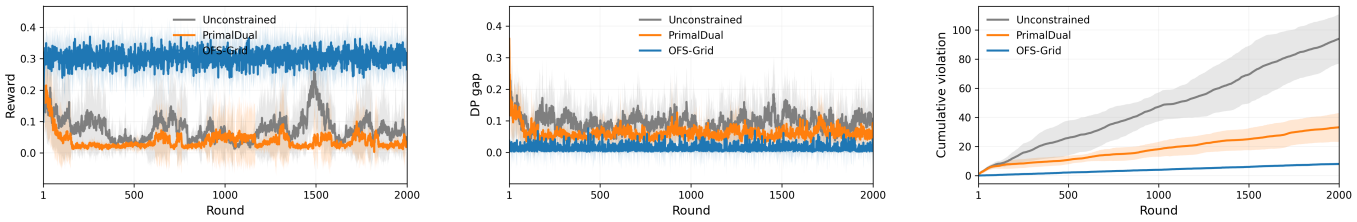
F. Statistical significance via bootstrap CIs

We compute bootstrap confidence intervals across seeds for key performance differences (e.g., OFS reward improvement over PD; PD vs. OFS in DP gap and violation). Table III shows that OFS improvements are statistically robust across benchmarks.

TABLE I

MAIN RESULTS. METRICS ARE COMPUTED AS TAIL MEAN OVER THE LAST K_{TAIL}=200 ITERATIONS PER SEED, THEN REPORTED AS MEAN±STD OVER 10 SEEDS. CUMULATIVE VIOLATION SUMS THE POSITIVE PARTS OF THE CONSTRAINTS OVER THE FULL HORIZON (DP ONLY FOR MVE-S; DP + ACCEPT-RATE CONSTRAINTS FOR GERMAN/COMPAS).

Task	Algorithm	Reward ↑	DP gap ↓	Cum. viol. ↓
MVE-S ($\varepsilon = 0.06$)	Unconstrained	0.0657±0.0373	0.0953±0.0230	93.88±16.93
	PrimalDual	0.0287±0.0155	0.0581±0.0141	33.14±9.96
	OFS-Grid	**0.3068**±0.0051	**0.0125**±0.0014	**7.98**±0.76
German ($\varepsilon = 0.02$, $a \in [0.30, 0.99]$)	Unconstrained	0.1127±0.0306	0.0448±0.0078	622.10±18.63
	PrimalDual	0.3929±0.0993	0.0309±0.0057	179.10±34.55
	OFS-Grid	**0.4546**±0.0064	**0.0131**±0.0021	**31.13**±1.54
COMPAS ($\varepsilon = 0.03$, $a \in [0.30, 0.99]$)	Unconstrained	0.1377±0.0391	0.1399±0.0133	921.37±12.37
	PrimalDual	0.6681±0.0225	0.0396±0.0046	161.49±12.92
	OFS-Grid	**0.7556**±0.0071	**0.0101**±0.0023	**26.53**±1.05



(a) Reward (b) DP gap (c) Cumulative violation

Fig. 1. MVE-S closed-loop learning dynamics (mean±std over seeds). OFS achieves higher reward with dramatically smaller DP gap and cumulative violation.

TABLE II

ORACLE BEST FEASIBLE PERFORMANCE (STEADY-STATE) AND SUBOPTIMALITY GAPS. ORACLE IS OBTAINED BY SWEEPING FIXED THRESHOLDS θ AND SELECTING THE BEST FEASIBLE POINT; GAPS ARE ORACLE REWARD MINUS EACH ALGORITHM'S STEADY-STATE REWARD (SMALLER IS BETTER).

Task	Oracle reward	Oracle DP	Oracle acc	Gap(U)	Gap(PD)	Gap(OFS)
MVE-S	0.3168	0.0062	0.9964	0.2571	0.2911	**0.0026**
German	0.4656	0.0153	0.9877	0.3303	0.0382	**0.0028**
COMPAS	0.7665	0.0157	0.9849	0.5950	0.0050	**0.0012**

TABLE III

BOOTSTRAP CONFIDENCE INTERVALS (95%) FOR KEY DIFFERENCES, COMPUTED OVER SEEDS USING PAIRED BOOTSTRAP.

Task	ΔReward (OFS–PD)	ΔDP (PD–OFS)	ΔViol. (PD–OFS)
MVE-S	0.2781 [0.2683, 0.2858]	0.0457 [0.0375, 0.0544]	25.16 [20.05, 32.29]
German	0.0617 [0.0051, 0.1261]	0.0178 [0.0149, 0.0211]	147.98 [127.88, 170.01]
COMPAS	0.0875 [0.0714, 0.1027]	0.0295 [0.0263, 0.0326]	134.96 [127.30, 142.90]

TABLE IV

PRIMALDUAL (PD) HYPERPARAMETER TUNING FOR SEMI-SYNTHETIC TASKS (SMALL GRID).

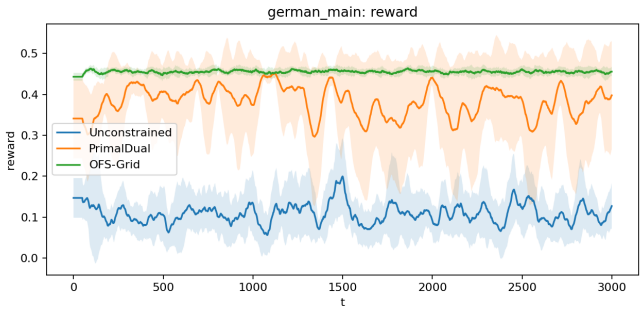
Dataset	Selected (η, μ)	Selection rule
German	(0.03, 0.10)	infeasible-min-violation
COMPAS	(0.03, 0.40)	infeasible-min-violation

G. Baseline tuning: Primal–Dual hyperparameters

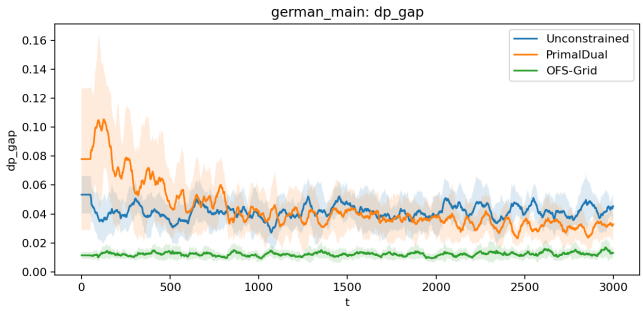
A common critique for primal–dual baselines is sensitivity to step sizes and dual updates. We therefore include a small PD tuning grid and report results (Table IV). Even after tuning, OFS maintains superior reward–violation tradeoffs on the semi-synthetic tasks. We will release the full tuning grid and experiment logs upon publication.

H. Robustness: stress sweeps and strict infeasibility

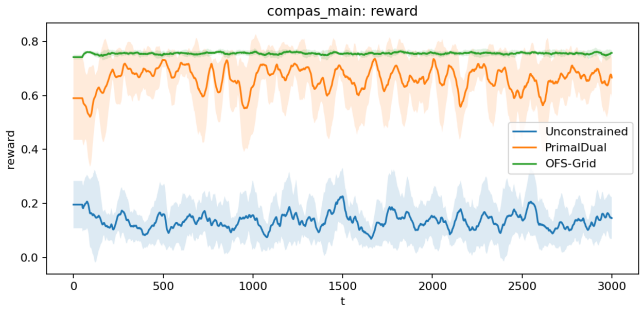
We evaluate robustness by sweeping key parameters (cost, fairness tolerance, feedback strength) and computing win-rates for OFS under feasibility-aware selection rules. Table V reports stress win-rates. For each stress configuration, we declare the winner as the feasible algorithm with the highest tail reward; if no algorithm is feasible, the winner is the one with the smallest cumulative violation. We also include a strict infeasible configuration where constraints are intentionally conflicting; in this regime, no method can be perfectly feasible, so the goal is to minimize cumulative violation while maintaining utility. Table VI summarizes strict infeasible results.



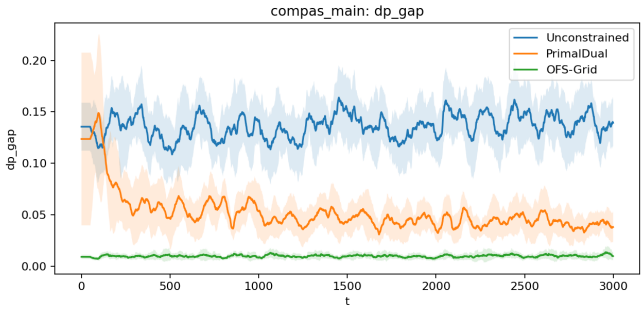
(a) German: reward



(b) German: DP gap

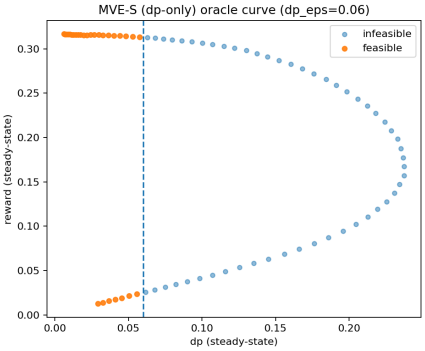


(c) COMPAS: reward

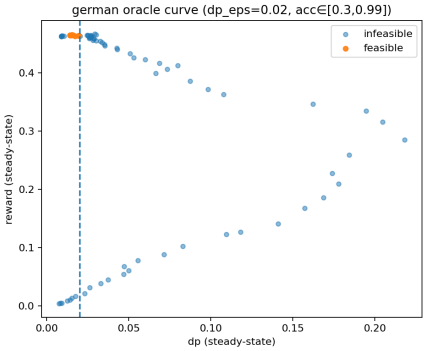


(d) COMPAS: DP gap

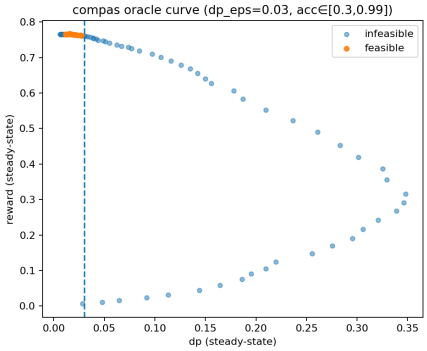
Fig. 2. Semi-synthetic closed-loop learning curves on German Credit and COMPAS. OFS consistently reduces constraint violations while preserving high reward.



(a) MVE-S



(b) German



(c) COMPAS

Fig. 3. Oracle tradeoff curves (fixed threshold steady-state evaluation). Each curve is obtained by sweeping thresholds and estimating steady-state reward and DP gap; the oracle best feasible point is the feasible point with maximal reward. OFS operates online yet approaches oracle-quality feasible thresholds.

TABLE V

STRESS SWEEP WIN RATE (PERCENTAGE OF STRESS SETTINGS WHERE THE METHOD IS THE WINNER). VALUES ARE TAKEN FROM STRESS_WINRATE_*.JSON; MISSING ENTRIES ARE REPORTED AS 0.0.

Task	OFS-Grid	PrimalDual
MVE-S	88.9%	11.1%
German	100.0%	0.0%
COMPAS	100.0%	0.0%

TABLE VI

STRICT INFEASIBLE REGIME WITH TIGHTER CONSTRAINTS ($\varepsilon = 0.01$, $\text{acc} \in [0.35, 0.85]$) FOR SEMI-SYNTHETIC TASKS. METRICS ARE REPORTED USING THE SAME PROTOCOL AS THE MAIN RESULTS.

Setting	Algorithm	Reward ↑	DP gap ↓	Cum. viol. ↓
German (strict)	Unconstrained	0.1127±0.0306	0.0448±0.0078	797.41±17.95
	PrimalDual	0.4115±0.0138	0.0469±0.0064	**383.10**±14.02
	OFS-Grid	**0.4546**±0.0064	**0.0131**±0.0021	445.69±0.92
COMPAS (strict)	Unconstrained	0.1377±0.0391	0.1399±0.0133	1130.38±11.32
	PrimalDual	0.6519±0.0251	0.0408±0.0043	476.69±3.81
	OFS-Grid	**0.7556**±0.0071	**0.0101**±0.0023	**440.62**±1.09

VII. Discussion and Limitations

Why OFS works well here. Closed-loop fairness problems often suffer from instability and oscillations when constraints are enforced indirectly. OFS explicitly searches for thresholds that appear feasible under uncertainty, which is particularly effective for low-dimensional, interpretable policy classes (thresholds) where discretization provides a transparent search space. In our experiments, OFS quickly identifies the feasible region and then refines reward within it, matching near-oracle performance.

Limitations. First, our current focus is on one-dimensional thresholds; extending optimism-based feasible screening to higher-dimensional policy classes may require structured discretization or surrogate modeling. Second, we focus on DP (and a service-rate constraint); other fairness notions (e.g., equalized odds [4]) may require different constraint estimators and could change tradeoffs. Third, our semi-synthetic benchmark captures composition shift via a simple feedback model; real-world feedback mechanisms can be more complex and may involve strategic behavior [13].

VIII. Conclusion

We presented OFS, an optimism-based method for closed-loop threshold learning under constraints. Across a synthetic benchmark and two semi-synthetic benchmarks derived from German Credit and COMPAS, OFS achieves higher reward and lower cumulative violations than strong baselines while remaining near-oracle. Our pipeline is fully reproducible and uses double-blind friendly relative outputs, making it straightforward to validate and extend these results.

References

[1] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, Cambridge, MA, USA, Jan. 2012, pp. 214–226.

[2] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Sydney, NSW, Australia, Aug. 2015, pp. 259–268.

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 60–69.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 30*, Barcelona, Spain, Dec. 2016, pp. 3323–3331.

[5] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*, ser. LIPIcs, vol. 67, Berkeley, CA, USA, Jan. 2017, pp. 43:1–43:23.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multi-armed bandit problem," *Machine Learning*, vol. 47, no. 2–3, pp. 235–256, May 2002.

[7] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the 20th International Conference on Machine Learning (ICML)*, Washington, DC, USA, Aug. 2003, pp. 928–936.

[8] M. Mahdavi, R. Jin, and T. Yang, "Trading regret for efficiency: online convex optimization with long-term constraints," *Journal of Machine Learning Research*, vol. 13, pp. 2503–2528, 2012.

[9] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, Sydney, NSW, Australia, Aug. 2017, pp. 22–31.

[10] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Vancouver, BC, Canada, Jan. 2005, pp. 385–394.

[11] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, Mar. 1992.

[12] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[13] J. Perdomo, T. Žrnić, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, Virtual (originally Vienna, Austria), Jul. 2020, pp. 7599–7609.

[14] D. Dua and C. Graff, "Uci machine learning repository," University of California, Irvine, 2017, accessed Dec. 24, 2025.

[15] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," ProPublica, May 2016, accessed Dec. 24, 2025.