A Boosting Approach to Reinforcement Learning

Nataly Brukhim
Princeton University
nbrukhim@cs.princeton.edu

Elad Hazan
Princeton University
Google AI Princeton
ehazan@cs.princeton.edu

Karan Singh
Carnegie Mellon University
karansingh@cmu.edu

Abstract

Reducing reinforcement learning to supervised learning is a well-studied and effective approach that leverages the benefits of compact function approximation to deal with large-scale Markov decision processes. Independently, the boosting methodology (e.g. AdaBoost) has proven to be indispensable in designing efficient and accurate classification algorithms by combining inaccurate *rules-of-thumb*.

In this paper, we take a further step: we reduce reinforcement learning to a sequence of weak learning problems. Since weak learners perform only marginally better than random guesses, such subroutines constitute a weaker assumption than the availability of an accurate supervised learning oracle. We prove that the sample complexity and running time bounds of the proposed method do not explicitly depend on the number of states.

While existing results on boosting operate on convex losses, the value function over policies is non-convex. We show how to use a non-convex variant of the Frank-Wolfe method for boosting, that additionally improves upon the known sample complexity and running time even for reductions to supervised learning.

1 Introduction

In reinforcement learning, Markov decision processes (MDP) model the mechanism of learning from rewards, as opposed to examples. Although the case of tabular MDPs is well understood, the main challenge in applying RL in the real-world is the size of the state space in practical domains.

This challenge of finding efficient and provable algorithms for MDPs with large state space is the focus of our study. Various techniques have been suggested and applied to cope with very large MDPs. One class of approaches attempts to approximate either the value or the transition function of the underlying MDP by using a parametric function class. Such approaches invariably make strong *realizability assumptions* to produce global optimality guarantees. Another class of approaches, *so-called* direct methods, produces a near-optimal policy that maximizes the expected return from a given policy class. To deal with the challenge of large (possibly innumerable) policy classes, a popular strategy [23] is to the frame policy search as a sequence of supervised learning problems. Such approaches yield global optimality guarantees under state coverage assumptions without reliance on realizability, and have inspired practical adaptations for sampling-based policy search.

In this paper, we study another methodology to derive provable algorithms for reinforcement learning: ensemble methods for aggregating weak or approximate algorithms into substantially more accurate solutions. Our proposal extends the methodology of boosting, typically used to solve supervised learning instances [31], to reinforcement learning. A typical boosting algorithm (e.g. AdaBoost)

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

	Supervised weak learner	Online weak learner
Episodic model	$1/\alpha^4\varepsilon^5$	$1/\alpha^2\varepsilon^3$
Rollouts w. ν-resets	$1/\alpha^4\varepsilon^6$	$1/\alpha^2\varepsilon^4$

Table 1: Sample complexity of the proposed algorithms for different α-weak learning models (supervised & online) and modes of accessing the MDP (rollouts & rollouts with reset distribution ν), in terms of ϵ and α, suppressing other terms. This work is the first to introduce a reduction of RL to *weak* supervised learning. See Theorem 7 for details.

	Supervised strong learner	
	This work (Theorem 8)	CPI [23]
Episodic model	$1/\varepsilon^3$	$1/\varepsilon^4$
Rollouts w. ν-resets	$1/\varepsilon^4$	$1/\varepsilon^4$

Table 2: Compared to previous work [23], the table shows sample complexity of the proposed algorithm for a strong ($\alpha = 1$) supervised learning model and different modes of accessing the MDP.

iteratively constructs a near-optimal classifier by combining computationally cheap, yet inaccurate *rules-of-thumb*. Unlike RL reductions to supervised learning which assume the existence of an efficient and accurate classification or regression procedure, the proposed algorithms builds on learning algorithms that perform only ever-so-slightly better than a random guess, and which thus may be produced cheaply both in computational and statistical terms.

Concretely, we assume access to a weak learner: an efficient sample-based procedure that is capable of generating an approximate solution to any weighted multi-class objective over a fixed policy class. We describe an algorithm that iteratively calls this procedure on carefully constructed new objectives, and aggregates the solution into a single policy. We prove that after sufficiently many iterations, our resulting policy has competitive global gurantees on performacnce. Interestingly, unlike boosting algorithms for regression and classification, our resulting aggregation of weak learners is non-linear.

1.1 Challenges and techniques

Reinforcement learning is quite different from supervised learning and several difficulties have to be circumvented for boosting to work. Among the challenges that the reinforcement learning setting presents, consider the following,

(a) The value function is not a convex or concave function of the policy. This is true even in the tabular case, and even more so if we use a parameterized policy class.

(b) The transition matrix is unknown, or prohibitively large to manipulate for large state spaces. This means that even evaluation of a policy cannot be exact, and can only be computed approximately.

(c) It is unrealistic to expect a weak learner that attains near-optimal value for a given linear objective over the policy class. At most one can hope for a multiplicative and/or additive approximation of the overall value.

Our approach overcomes these challenges by applied several new as well as recently developed techniques. To overcome the nonconvexity of the value function, we use a novel variant of the Frank-Wolfe optimization algorithm that simultaneously delivers on two guarantees. First, it finds a first order stationary point with near-optimal rate. Secondly, if the objective happens to admit a certain gradient domination property, an important generalization of convexity, it also guarantees near optimal value. The application of the nonconvex Frank-Wolfe method is justified due to previous recent investigation of the policy gradient algorithm [2, 1], which identified conditions under which the value function is gradient dominated.

The second information-theoretic challenge of the unknown transition function is overcome by careful algorithmic design: our boosting algorithm requires only samples of the transitions and rewards, obtained by rollouts on the MDP.

The third challenge is perhaps the most difficult to overcome. Thus far, the use of the Frank-Wolfe method in reinforcement learning did not include a multiplicative approximation, which is critical for our application. We adapt the techniques used for boosting in online convex optimization [18] with a multiplicative weak learner to our setting, by non-linearly aggregating (using a 2-layer network) the weak learners. This aspect is perhaps of general interest to boosting algorithm design, which is mostly based on linear aggregation.

1.2 Our contributions

Our main contribution is a novel efficient boosting algorithm for reinforcement learning. Our techniques apply in various settings and the sample complexity bounds of all of our results are summarized in Tables 1 and 2.

The input to this algorithm is a weak learning method capable of approximately solving a weighted multi-class problem instance over a certain policy class. The output of the algorithm is a policy which does not belong to the original class considered, hence being an instance of *improper* learning. It is rather a non-linear aggregation of policies from the original class, according to a two-layer neural network. This is a result of the two-tier structure of our algorithm: an outer loop of non-convex Frank-Wolfe method, and an inner loop of online convex optimization based boosting. The final policy comes with provable global optimality guarantees.

Beyond novelty of techniques, an important contribution (Table 1) of our work is to highlight the quantitative difference in guarantees that depend on the mode of accessing the MDP (episodic rollouts vs. access to an exploratory reset distrbution) and the nature of the weak learners (online vs statistical), thus indicating that some algorithmic choices may be preferable compared to others in terms of speed of convergence and sample complexity.

As with existing reductions to supervised learning [23], these global convergence guarantees happen under appropriate state coverage assumptions either via access to a reset distribution that has some overlap with the state distribution of the optimal policy, or by constraining the policy class to policies that explore sufficiently. Yet another contribution of our work is to show an improved sample complexity result in the latter setting, *even when considering reductions to supervised learning instances*. This improvement in convergence in well-studied settings is documented in Table 2.

1.3 Related work

Reinforcement learning approaches for dealing with large-scale MDPs rely on function approximation [34]. Such function approximation may be performed on the underlying conditional probability of transition (e.g. [33, 20]) or the value function (e.g. [36, 35]). The provable guarantees in such methods come at the cost of strong realizability assumptions. In contrast, the so-called direct approaches attempt policy search over an appropriate policy class [2, 1], and rely on making making incremental updates, such as Conservative Policy Iteration (CPI) [23, 32], and Policy Search by Dynamic Programming (PSDP)[5]. These provide convergence guarantees under appropriate state coverage assumptions comparable to ones made in this work.

Our boosting approach for provable RL builds on the vast literature of boosting for supervised learning [31], and recently online learning [26, 11, 12, 6, 21, 22]. One of the crucial techniques important for our application is the extension of boosting to the online convex optimization setting, with bandit information [9], and critically with a multiplicative weak learner [18]. This latter technique implies a non-linear aggregation of the weak learners. Non-linear boosting was only recently investigated in the context of classification [4], where it was shown to potentially enable significantly more efficient boosting. Another work on boosting in the context of control of dynamical systems [3]. However, this work critically requires knowledge of the underlying dynamics (transitions) and makes convexity assumptions, which we do not, and cannot cope with a multiplicative approximate weak learner.

The Frank-Wolfe algorithm is extensively used in machine learning, see e.g. [19], references therein, and recent progress in stochastic Frank-Wolfe methods [15, 27, 10, 38]. Recent literature has applied a variant of this algorithm to reinforcement learning in the context of state space exploration [17].

2 Preliminaries

Optimization. We say that a differentiable function $f : \mathcal{K} \mapsto \mathbb{R}$ over some domain $\mathcal{K} \subset \mathbb{R}^d$ is L-smooth with respect to some norm $\|\cdot\|_*$ if for every $x, y \in \mathcal{K}$ we have $\left|f(y) - f(x) - \nabla f(x)^\top (y-x)\right| \leq \frac{L}{2}\|x-y\|_*^2$. We define the projection $\Gamma : \mathbb{R}^{|A|} \to \Delta_A$, with respect to a set A, where Δ_A denotes the probability simplex over A. For any $x \in \mathbb{R}^{|A|}$, $\Gamma[x] = \arg\min_{y \in \Delta_A} \|x - y\|$. An important generalization of the property of convexity we use henceforth is that of gradient domination.

Definition 1 (Gradient Domination)**.** A function $f : \mathcal{K} \to \mathbb{R}$ is said to be $(\kappa, \tau, \mathcal{K}_1, \mathcal{K}_2)$-locally gradient dominated (around $\mathcal{K}_1$ by $\mathcal{K}_2$) if for all $x \in \mathcal{K}_1$, it holds that

$$\max_{y \in \mathcal{K}} f(y) - f(x) \;\leq\; \kappa \cdot \max_{y \in \mathcal{K}_2} \left\{\nabla f(x)^\top (y - x)\right\} + \tau.$$

Markov decision process. An infinite-horizon discounted Markov Decision Process (MDP) $\mathcal{M} = (S, A, P, r, \gamma, d_0)$ is specified by: a state space S, an action space A, a transition model P where $P(s'|s,a)$ denotes the probability of immediately transitioning to state s' upon taking action a at state s, a reward function $r : S \times A \to [0,1]$ where $r(s,a)$ is the immediate reward associated with taking action a at state s, a discount factor $\gamma \in [0,1)$; a starting state distribution d_0 over S. For any infinite-length state-action sequence (hereafter, called a trajectory), we assign the following value $V(\varsigma = (s_0, a_0, s_1, a_1, \dots)) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$. The agent interacts with the MDP through the choice of stochastic policy $\pi : S \to \Delta_A$ it executes. The execution of such a policy induces a distribution over trajectories $\varsigma = (s_0, a_0, \dots)$ as $P(\varsigma|\pi) = d_0(s_0) \prod_{t=0}^{\infty} (P(s_{t+1}|s_t, a_t)\pi(a_t|s_t))$. Using this description we can associate a state $V^\pi(s)$ and state-action $Q^\pi(s,a)$ value function with any policy π. For an arbitrary distribution d over S, define:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| \pi, s_0 = s, a_0 = a\right],$$

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[Q^\pi(s,a)|\pi, s\right],\; V_d^\pi = \mathbb{E}_{s_0 \sim d}\left[V^\pi(s)|\pi\right].$$

Here the expectation is with respect to the randomness of the trajectory induced by π in $\mathcal{M}$. When convenient, we shall use V^π to denote $V_{d_0}^\pi$, and V^* to denote $\max_\pi V^\pi$.

Similarly, to any policy π, one may ascribe a (discounted) state-visitation distribution $d^\pi = d_{d_0}^\pi$.

$$d_d^\pi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \sum_{\varsigma : s_t = s} P(\varsigma|\pi, s_0 \sim d)$$

Modes of Accessing the MDP. We henceforth consider two modes of accessing the MDP, that are standard in the reinforcement learning literature, and provide different results for each.

The first natural access model is called the **episodic rollout setting.** This mode of interaction allows us to execute a policy, stop and restart at any point, and do this multiple times.

Another interaction model we consider is called **rollout with ν-restarts.** This is similar to the episodic setting, but here the agent may draw from the MDP a trajectory seeded with an initial state distribution $\nu \neq d_0$. This interaction model was considered in prior work on policy optimization [23, 2]. The motivation for this model is two-fold: first, ν can be used to incorporate priors (or domain knowledge) about the state coverage of the optimal policy; second, ν provides a mechanism to incorporate exploration into policy optimization procedures.

2.1 Weak learning

Our boosting algorithms henceforth call upon weak learners to generate weak policies. We formalize the notion of a weak learner next. We consider two types of weak learners, and give different end results based on the different assumptions: weak supervised and weak online learners. In the discussion below, let π_{Rand} be a uniformly random policy, i.e. $\forall (s,a) \in S \times A, \pi_{Rand}(a|s) = 1/|A|$. The formal definition and results for the online setting are deferred to the appendix. In what follows we define the supervised weak learning model.

The natural way to define weak learning is an algorithm whose performance is always slight better than that of random policy, one that chooses an action uniformly at random at any given state. However, in general no learner can outperform a random learner over all label distributions. This motivates the literature on agnostic boosting [24, 8, 18] that defines a weak learner as one that can approximate the best policy in a given policy class.

Definition 2 (Weak Supervised Learner)**.** Let $\alpha \in (0, 1]$. Consider a class $\mathcal{L}$ of linear loss functions $\ell : \mathbb{R}^A \to \mathbb{R}$, a family $\mathbb{D}$ of distributions that are supported over $S \times \mathcal{L}$, and policy class Π. A weak supervised learning algorithm, for every $\varepsilon, \delta > 0$, given $m(\varepsilon, \delta) = \frac{\log|\mathcal{W}|}{\varepsilon^2} \log \frac{1}{\delta}$ samples D_m from any distribution $\mathcal{D} \in \mathbb{D}$ outputs a policy $\mathcal{W}(D_m) \in \Pi$ such that with probability $1 - \delta$,

$$\mathbb{E}_{(s,\ell)\sim\mathcal{D}}\left[\ell(\mathcal{W}(D_m))\right] \;\leq\; \alpha \min_{\pi^* \in \Pi} \mathbb{E}_{(s,\ell)\sim\mathcal{D}}\left[\ell(\pi^*(s))\right] + (1-\alpha)\, \mathbb{E}_{(s,\ell)\sim\mathcal{D}}\left[\ell(\pi_{Rand}(s))\right] + \varepsilon.$$

Note that the weak learner outputs a policy in Π which is approximately competitive against the class Π. As an additional relaxation, instead of requiring that the weak learning guarantee holds for all distributions, in our setup, it will be sufficient that the weak learning assumption holds over *natural* distributions. Specifically, we define a class of *natural* distributions $\mathbb{D}$, such that $\mathcal{D} \in \mathbb{D}$ if and only if there exists some $\pi \in \Pi$ such that, $\mathcal{D}(s) = \int_\ell \mathcal{D}(s, \ell) d\mu(\ell) = d^\pi(s)$. In particular, while a *natural* distribution may have arbitrary distribution over labels, its marginal distribution over states must be realizable as the state distribution of some policy in Π over the MDP $\mathcal{M}$. Therefore, the complexity of weak learning adapts to the complexity of the MDP itself. As an extreme example, in stochastic contextual bandits where policies do not affect the distribution of states (say d_0), it is sufficient that the weak learning condition holds with respect to all couplings of a single distribution d_0.

3 Algorithm & Main Results

In this section we describe our RL boosting algorithm. Here we focus on the case where a supervised weak learning is provided. The online weak learners variant of our result is detailed in the appendix. We next define several definitions and algorithmic subroutines required for our method.

3.1 Policy aggregation

For a base class of policies Π, our algorithm incrementally builds a more expressive policy class by aggregating base policies via both linear combinations and non-linear transformations. In effect, the algorithm produces a finite-width depth-2 circuit over some subset of the base policy class. That is, our approach can be thought of as an aggregation of base policies, which forms a 2-layer neural network, as depicted in Figure 1. The leaves of the tree are the policies $\pi \in \Pi$ the base policy class. These are then linearly aggregated to form the first layer of the tree, denoted $\tilde{\pi}_1, \tilde{\pi}_2$ in Figure 1.

Next, each linear combination of policies in the overall aggregation undergoes a projection operation. The projection may be viewed as a non-linear activation function, such as ReLU, in deep learning terms. Note that the projection of any function from S to $\mathbb{R}^{|A|}$ produces a policy, i.e. a mapping from states to distributions over actions. In the analysis of our algorithm we give a particular projection operation $\Gamma[\cdot]$ which allows us to yield the desired guarantees.

Definition 3 (Policy Projection)**.** Given $\tilde{\pi} : S \to \mathbb{R}^{|A|}$, define a projected policy $\pi = \Gamma[\tilde{\pi}]$ to be a policy such that simultaneously for all $s \in S$, it holds that $\pi(\cdot|s) = \Gamma\left[\tilde{\pi}(s)\right]$.

Definition 4 (Policy Tree)**.** A *Policy Tree* $\mathbb{\Pi} \subseteq S \to \Delta_A$ with respect to $\Pi \subseteq S \to \Delta_A$ some base policy class, and $N, T \in \mathbb{N}$, is a linear combination of T projected policies $\Gamma[\tilde{\pi}]$, where each $\tilde{\pi}$ is a linear combination of N base policies $\pi \in \Pi$.

This final definition describes the set of possible outputs of the boosting procedure. It is important that the policy that the boosting algorithm outputs can be evaluated efficiently. In the appendix we show it is indeed the case (see Lemma 12). Hereafter, we refer to a Policy Tree with respect to Π, N and T, as $\mathbb{\Pi}$ for $N, T = O(\mathrm{poly}(|A|, (1-\gamma)^{-1}, \varepsilon^{-1}, \alpha^{-1}, \log \delta^{-1}))$ specified later.

3.2 Main results

Next, we give the main results of our RL boosting algorithm via weak supervised learning, specified in Algorithm 1.

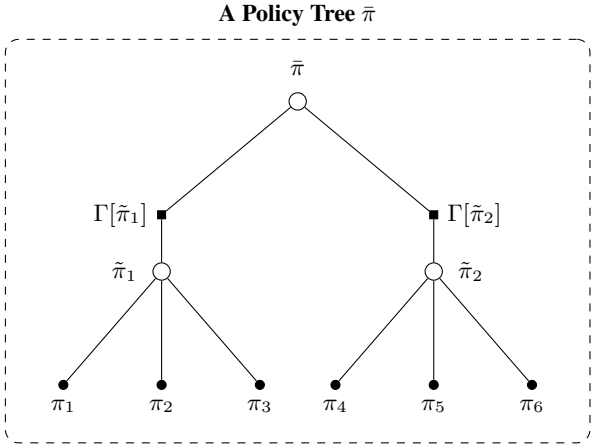

Figure 1: The figure illustrates a Policy Tree hierarchy (see Definition 4), output of the boosting procedure specified in Algorithm 1. Specifically, it is obtained by setting $N = 3$ on the inner loop of Internal Boost (Algorithm 2), and $T = 2$ on the main booster (Algorithm 1). Overall we get all base policies $\pi_1, ..., \pi_6 \in \Pi$ on the lower level, to form the Policy Tree $\bar{\pi} \in \mathbb{\Pi}$.

Algorithm 1 RL Boosting

1: **Input**: number of iterations T, initial state distribution μ, and P, N, M parameters for Internal Boost.
2: Initialize a policy $\pi_0 \in \Pi$ arbitrarily.
3: **for** $t = 1$ **to** T **do**
4: Run Internal Boost (Algorithm 2) with distribution μ and policy π_t to obtain π'_t.
5: Update $\pi_t = (1 - \eta_{1,t})\pi_{t-1} + \eta_{1,t}\pi'_t$.
6: **end for**
7: Run each policy π_t for P rollouts to compute an empirical estimate $\widehat{V^{\pi_t}}$ of the expected return.
8: **return** $\bar{\pi} := \pi_{t'}$ where $t' = \arg\max_t \widehat{V^{\pi_t}}$.

To state the results, we need the following definitions. The first generalizes the policy completeness notion from [32]. It may be seen as the policy-equivalent analogue of inherent bellman error [28]. Intuitively, it measures the degree to which a policy in Π can best approximate the bellman operator in an average sense with respect to the state distribution induced by a policy from $\mathbb{\Pi}$.

Definition 5 (Policy Completeness)**.** For any initial state distribution μ, and policy classes $\Pi, \mathbb{\Pi}$, define $\mathcal{E}_\mu = \max_{\pi \in \mathbb{\Pi}} \min_{\pi^* \in \Pi} \mathbb{E}_{s \sim d^\pi_\mu} \left[\max_{a \in A} Q^\pi(s, a) - Q^\pi(s, \cdot)^\top \pi^*(\cdot|s) \right]$.

Definition 6 (Distribution Mismatch)**.** Let $\pi^* = \arg\max_\pi V^\pi$, and ν a fixed initial state distribution (see section 2). Define the following distribution mismatch coefficients: $C_\infty = \max_{\pi \in \mathbb{\Pi}} \left\| d^{\pi^*}/d^\pi \right\|_\infty, D_\infty = \left\| d^{\pi^*}/\nu \right\|_\infty$.

The above notion of the distribution mismatch coefficient is often useful to characterize the exploration problem faced by policy optimization algorithms. We now give the main result for the output of our RL boosting algorithm, assuming supervised weak learners.

Theorem 7. *Algorithm 1 samples $T(MN + P)$ episodes of length $\tilde{O}(\frac{1}{1-\gamma})$ with probability $1 - \delta$. In the episodic model, with $\mu = d_0$, for $\eta_{1,t} = \min\{1, \frac{2C_\infty}{t}\}$, $T = O\left(\frac{C^2_\infty}{(1-\gamma)^3 \varepsilon}\right)$, $N = \left(\frac{16|A|C_\infty}{(1-\gamma)^2 \alpha \epsilon}\right)^2$, $M = m\left(\frac{(1-\gamma)^2 \alpha \varepsilon}{C_\infty |A|}, \frac{\delta}{NT}\right)$, with probability $1 - \delta$, $V^* - V^\pi \leq \frac{C_\infty \mathcal{E}}{1-\gamma} + \varepsilon$.*

In the ν-reset model, with $\mu = \nu$, for $\eta_{1,t} = \sqrt{\frac{8\gamma(1-\gamma)^2}{|A|^2 T}}$, $T = \frac{8D^2_\infty}{(1-\gamma)^6 \varepsilon^2}$, $N = \left(\frac{16|A|D_\infty}{(1-\gamma)^3 \alpha \epsilon}\right)^2$, $M = m\left(\frac{(1-\gamma)^3 \alpha \varepsilon}{8|A|D_\infty}, \frac{\delta}{2NT}\right)$, with probability $1 - \delta$, $V^* - V^\pi \leq \frac{D_\infty \mathcal{E}_\nu}{(1-\gamma)^2} + \varepsilon$.

Sample complexities: *If $m(\varepsilon, \delta) = \frac{\log |\mathcal{W}|}{\varepsilon^2} \log \frac{1}{\delta}$ for some measure of weak learning complexity $|\mathcal{W}|$,*

Algorithm 2 Internal Boost

1: **Input**: number of iterations N, number of episodes M, initial policy π, initial state distribution μ.
2: Set $\tilde{\pi}_0$ to be an arbitrary policy in Π.
3: **for** $n = 1$ **to** N **do**
4: Execute π with μ via Algorithm 3 for M episodes, to get $D_n = \{(s_i, \widehat{Q_i})_{i=1}^M\}$.
5: Modify D_n to produce a new dataset $D'_n = \{(s_i, f_i)\}_{i=1}^M$, such that for all $i \in [m]$:

$$f_i = \frac{1}{\beta}\left(y_i - \tilde{\pi}_n(\cdot|s_i)\right), \quad y_i = \underset{y \in \mathbb{R}^{|A|}}{\arg\min}\{-\widehat{Q}_i^\top y + G \min_{z \in \Delta_A} \|z - y\| + \frac{\|\tilde{\pi}_n(\cdot|s_i) - y\|^2}{2\beta}\}$$

where $G = \frac{A}{1-\gamma}, \beta = \frac{2\gamma}{(1-\gamma)^3}$ and $f_i, \widehat{Q}_i \in \mathbb{R}^{|A|}$.
6: Let $\mathcal{A}_n$ be the policy chosen by the weak learning oracle when given data set $D'_{t,n}$.
7: Update

$$\tilde{\pi}_n = (1 - \eta_{2,n})\tilde{\pi}_{n-1} + \frac{\eta_{2,n}}{\alpha}\mathcal{A}_n.$$

8: **end for**
9: **return** $\Gamma[\tilde{\pi}_N]$.

Algorithm 3 Trajectory Sampler: samples a state $s \sim d^\pi$, and an unbiased estimate of Q_s^π

1: Sample state $s_0 \sim \mu$, action $a' \sim \mathcal{U}(A)$ uniformly.
2: Sample $s \sim d^\pi$ as follows: at every timestep h, with probability γ, act according to π; else, accept s_h as the sample and proceed to Step 3.
3: Take action a' at state s_h, then continue to execute π, and use a termination probability of $1 - \gamma$. Upon termination, set $R(s_h, a')$ as the *undiscounted* sum of rewards from time h onwards.
4: Define the vector $\widehat{Q^\pi_{s_h}}$, such that for all $a \in A$, $\widehat{Q^\pi_{s_h}}(a) = |A| \cdot R(s_h, a') \cdot \mathbb{I}_{a=a'}$.
5: **return** $(s_h, \widehat{Q^\pi_{s_h}})$.

the algorithm samples $\tilde{O}\left(\frac{C_\infty^6 |A|^4 \log|\mathcal{W}|}{(1-\gamma)^{11}\alpha^4\varepsilon^5}\right)$ *episodes in the episodic model, and* $\tilde{O}\left(\frac{D_\infty^6 |A|^4 \log|\mathcal{W}|}{(1-\gamma)^{18}\alpha^4\varepsilon^6}\right)$ *in the ν-reset model.*

Theorem 7 above pertains to the case where a weak learning algorithm is available. However, another main result is given by considering the simpler approach of reduction of RL to a *strong* supervised learning algorithm. In particular, when running our main boosting algorithm, we can replace the call to Internal Boost (in Line 4 of Algorithm 1) with a call to a *strong* supervised learning algorithm. By a similar analysis to that of Theorem 7 we obtain the following corollary.

Corollary 8. *Let* $m(\varepsilon, \delta) = \frac{\log|\mathcal{W}|}{\varepsilon^2}\log\frac{1}{\delta}$ *for some measure of weak learning complexity* $|\mathcal{W}|$*. When run with a supervised learning oracle (Definition 2 with* $\alpha = 1$*, i.e.* $N = 1$*) as the Internal boosting, Algorithm 1 samples* $\tilde{O}\left(\frac{C_\infty^3 \log|\mathcal{W}|}{\varepsilon^3}\right)$ *episodes in the episodic model, and* $\tilde{O}\left(\frac{D_\infty^4 \log|\mathcal{W}|}{\varepsilon^4}\right)$ *in the ν-reset model, to guarantee* $V^* - V^\pi \leq \frac{C_\infty \mathcal{E}}{1-\gamma} + \varepsilon$ *with probability* $1 - \delta$ *in the episodic model and* $V^* - V^\pi \leq \frac{D_\infty \mathcal{E}_\nu}{(1-\gamma)^2} + \varepsilon$ *in the ν-reset model.*

We note that this result is an improvement over previous results in terms of sample complexity requirement of the algorithm. In particular, in [23], Theorem 4.4 and Corollary 4.5 achieve the same guarantee using $O(1/\varepsilon^4)$ samples regardless of the MDP access model. Briefly, CPI utilizes $1/\varepsilon^2$ calls to an ε-optimal supervised learning oracle (each call needing $1/\varepsilon^2$ samples) to reach a ε-local optima of the value function. Under requisite state coverage assumptions, this translates to ε-function value suboptimality. Indeed, such mode of analysis via first arguing for convergence to a local optima for the CPI algorithm can be shown to be tight. The improvement in our case for the episodic access model comes from the insight that it is possible to make direct claims on the function value sub-optimality (second part of Theorem 9), bypassing the need for making a claim on the local optimality, in the gradient-dominated case.

3.3 Trajectory sampler

In Algorithm 3 we describe an episodic sampling procedure, that is used in our sample-based RL boosting algorithms described above. For a fixed initial state distribution μ, and any given policy π, we apply the following sampling procedure: start at an initial state $s_0 \sim \mu$, and continue to act thereafter in the MDP according to any policy π, until termination. With this process, it is straightforward to both sample from the state visitation distribution $s \sim d^\pi$, and to obtain unbiased samples of $Q^\pi(s, \cdot)$; see Algorithm 3 for the detailed process.

4 Sketch of the analysis

Non-convex Frank-Wolfe. We give an abstract high-level procedural template that the previously introduced RL boosters operate in. This is based on a variant of the Frank-Wolfe optimization technique [14], adapted to non-convex and gradient dominated function classes (see Definition 1). The Frank-Wolfe (FW) method assumes oracle access to a black-box linear optimizer, denoted $\mathcal{O}$, and utilizes it by iteratively making oracle calls with modified objectives, in order to solve the harder task of convex optimization. Analogously, boosting algorithms often assume oracle access to a "weak" learner, which are utilized by iteratively making oracle calls with modified objective, in order to obtain a "strong" learner, with boosted performance. In the RL setting, the objective is in fact non-convex, but exhibits gradient domination. By adapting Frank-Wolfe technique to this setting, we will in subsequent section obtain guarantees for the algorithms given in Section 3.**Oracle:** Denote by $\mathcal{O}$ a black-box oracle to an $(\epsilon_0, \mathcal{K}_2)$-approximate linear optimizer over a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ such that for any given $v \in \mathbb{R}^d$, we have $v^\top \mathcal{O}(v) \geq \max_{u \in \mathcal{K}_2} v^\top u - \epsilon_0$.

Algorithm 4 Non-convex Frank-Wolfe

1: Input: $T > 0$, objective f, linear optimizer $\mathcal{O}$, rate η_t.
2: Choose $x_0 \in \mathcal{K}$ arbitrarily.
3: **for** $t = 1, \ldots, T$ **do**
4: Call $z_t = \mathcal{O}(\nabla_{t-1})$, where $\nabla_{t-1} = \nabla f(x_{t-1})$. Set $x_t = (1 - \eta_t)x_{t-1} + \eta_t z_t$.
5: **end for**
6: **return** $\bar{x} := x_{t'}$ where $t' = \arg\min_t \nabla_t^\top (z_t - x_t)$.

Theorem 9. *Let $f : \mathcal{K} \to \mathbb{R}$ be L-smooth in some norm $\|\cdot\|_*$, bounded for all $x \in \mathcal{K}$, $|f(x)| \leq H$ for some $H > 0$, and let the diameter of $\mathcal{K}$ in $\|\cdot\|_*$ be D. Then, for a $(\epsilon_0, \mathcal{K}_2)$-linear optimization oracle $\mathcal{O}$, and $\eta_t = \eta = \sqrt{\frac{4H}{LD^2T}}$, the output $\bar{x}$ of Algorithm 4 satisfies*

$$\max_{u \in \mathcal{K}_2} \nabla f(\bar{x})^\top (u - \bar{x}) \leq \sqrt{\frac{2HLD^2}{T}} + \epsilon_0; \max_{x^* \in \mathcal{K}} f(x^*) - f(\bar{x}) \leq \frac{2\kappa^2 \max\{LD^2, H\}}{T} + \tau + \kappa\epsilon_0$$

Furthermore, if f is $(\kappa, \tau, \mathcal{K}_1, \mathcal{K}_2)$-locally gradient-dominated and $x_0, \ldots x_T \in \mathcal{K}_1$, then the output $\bar{x}$ of Algorithm 4 where $\eta_t = \min\{1, \frac{2\kappa}{t}\}$ satisfies the bound on the right.

We sketch the high-level ideas of the proof of our main result, stated in Theorem 7, and refer the reader to the appendix for the formal proof. We will establish an equivalence between RL Boosting (Algorithm 1) and the variant of the Frank-Wolfe algorithm (Algorithm 4). This abstraction allows us to obtain the novel convergence guarantees given in Theorem 7. Throughout the analysis, we use the notation $\nabla_\pi V^\pi$ to denote the gradient of the value function with respect to the $|S| \times |A|$-sized representation of the policy π, namely the functional gradient of V^π.

Internal-boosting weak learners. The Frank-Wolfe algorithm utilizes an inner gradient optimization oracle as a subroutine. To implement this oracle using approximate optimizers, we utilize yet another variant of the FW method as "internal-boosting" for the weak learners, by employing an adapted analysis of [18] that is stated in Claim 10 below. Let $\mathcal{D}_t$ be the distribution induced by the trajectory sampler in round t.

Claim 10. *Let $\beta = \sqrt{1/\alpha N}$, $\eta_{2,n} = \min\{2/n, 1\}$. π'_t produced by Algorithm 1 satisfies $\max_{\pi \in \Pi} \mathbb{E}_{(s,Q)\sim\mathcal{D}_t} \left[Q^\top \pi(s)\right] - \mathbb{E}_{(s,Q)\sim\mathcal{D}_t} \left[Q^\top \pi'_t(s)\right] \leq (2|A|/(1-\gamma)\alpha) \left(\varepsilon + 2/\sqrt{N}\right)$.*

From weak learning to linear optimization, Next, we give an important observation which allows us to re-state the guarantee in the previous subsection in terms of linear optimization over functional gradients. The key observation here is that the expensive optimizing procedure for $(\nabla_\pi V^\pi)^\top \pi'$, which in particular requires iterating over all states in S, can be instead replaced with sampling from an appropriate distribution $\mathcal{D}$ (via Algorithm 3). These sample pairs $(s, \widehat{Q^\pi}(s,\cdot))$ could then be fed to our weak learning algorithm, which guarantees generalization.

Lemma 11. *Applying Algorithm 3 for any given policy π yields an unbiased estimate of the gradient, such that for any π', $(\nabla_\pi V_\mu^\pi)^\top \pi' = \mathbb{E}_{(s,\widehat{Q^\pi}(s,\cdot))\sim\mathcal{D}}\left[\widehat{Q^\pi}(s,\cdot)^\top \pi'(\cdot|s)\right]/(1-\gamma)$, where $\pi'(\cdot|s) \in \Delta_A$, $\mathcal{D}$ is the distribution induced on the outputs of Algorithm 3, for the policy π and initial distribution μ.*

5 Experiments

The primary contribution of the present work is theoretical. Nevertheless, we empirically test our proposal with the experiment designed to elicit qualitative properties of the proposed algorithm, instead of aiming to achieve the state-of-the-art. To validate our results, we check if the proposed algorithm is indeed capable of boosting the accuracy of concrete instantiations of weak learners. We use depth-3 decision trees, with the implementation adapted from Scikit-Learn [29], as our base weak learner. This choice of weak learner is particularly suitable for boosting, because it is an impoverished policy class in a representational sense and hence it is reasonable to expect that it may do only slightly better than random guessing with respect to the classification loss. We consider the performance of the boosting algorithm (Algorithm 1) across multiple rounds of boosting or number of weak learners to that of supervised-learning-based policy iteration; the computational burden of the algorithm scales linearly with the latter. Throughout all the experiments, we used $\eta = 0.9$. To speed up computation, the plots below were generated by retaining the 3 most recent policies of every iteration in the policy mixture. We evaluated these on the CartPole and the LunarLander environments. The results demonstrate the proposed RL boosting algorithm succeeds in maximizing rewards while using few weak learners (equivalently, within a few rounds of boosting).

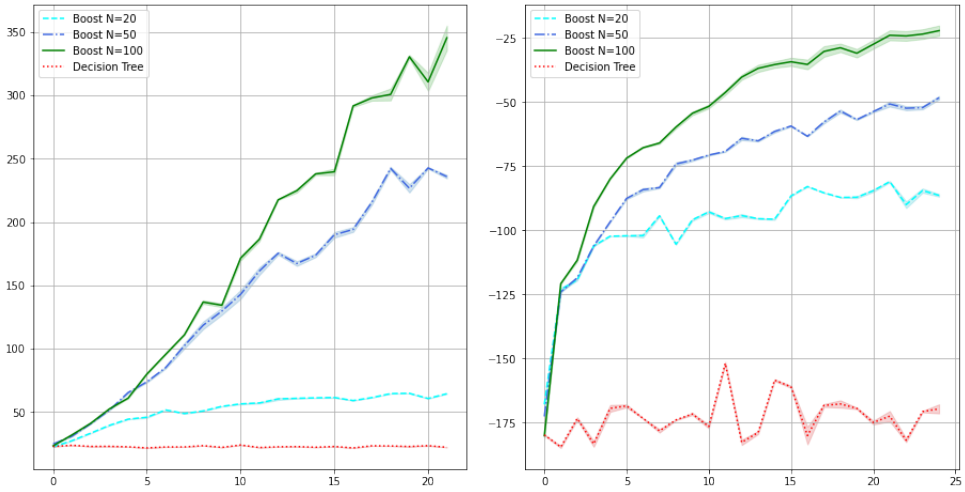

Figure 2: Reward trajectory for the CartPole (left) and the LunarLander (right) environments of the proposed boosting algorithm for $N = 20, 50, 100$ number of base weak learners is compared to supervised-learning-based policy iteration (decision tree) above. The x-axis corresponds to T number of iterations, and for each $t \in [T]$, reward is computed over 100 episodes of interactions. The confidence interval is plotted over 3 such runs.

6 Conclusions

Building on recent advances in boosting for online convex optimization and bandits, we have described a boosting algorithm for reinforcement learning over large state spaces with provable guarantees. We see this as a first attempt at using a tried-and-tested methodology from supervised learning to RL.

References

[1] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

[2] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

[3] Naman Agarwal, Nataly Brukhim, Elad Hazan, and Zhou Lu. Boosting for control of dynamical systems. In *International Conference on Machine Learning*, pages 96–103. PMLR, 2020.

[4] Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. Boosting simple learners. *arXiv preprint arXiv:2001.11704*, 2020.

[5] J Andrew Bagnell, Sham Kakade, Andrew Y Ng, and Jeff G Schneider. Policy search by dynamic programming. In *Advances in Neural Information Processing Systems*, 2003.

[6] Alina Beygelzimer, Satyen Kale, and Haipeng Luo. Optimal and adaptive algorithms for online boosting. In *International Conference on Machine Learning*, pages 2323–2331, 2015.

[7] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[8] Nataly Brukhim, Xinyi Chen, Elad Hazan, and Shay Moran. Online agnostic boosting via regret minimization. In *Advances in Neural Information Processing Systems*, 2020.

[9] Nataly Brukhim and Elad Hazan. Online boosting with bandit feedback. In *Algorithmic Learning Theory*, pages 397–420. PMLR, 2021.

[10] Lin Chen, Christopher Harshaw, Hamed Hassani, and Amin Karbasi. Projection-free online optimization with stochastic gradient: From convexity to submodularity. In *International Conference on Machine Learning*, pages 814–823, 2018.

[11] Shang-Tse Chen, Hsuan-Tien Lin, and Chi-Jen Lu. An online boosting algorithm with theoretical justifications. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1873–1880, 2012.

[12] Shang-Tse Chen, Hsuan-Tien Lin, and Chi-Jen Lu. Boosting with online binary learners for the multiclass bandit problem. In *International Conference on Machine Learning*, pages 342–350, 2014.

[13] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

[14] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[15] Hamed Hassani, Mahdi Soltanolkotabi, and Amin Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems*, pages 5841–5851, 2017.

[16] Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

[17] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

[18] Elad Hazan and Karan Singh. Boosting for online convex optimization. *arXiv preprint arXiv:2102.09305*, 2021.

[19] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[21] Young Hun Jung, Jack Goetz, and Ambuj Tewari. Online multiclass boosting. In *Advances in neural information processing systems*, pages 919–928, 2017.

[22] Young Hun Jung and Ambuj Tewari. Online boosting algorithms for multi-label ranking. In *International Conference on Artificial Intelligence and Statistics*, pages 279–287, 2018.

[23] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[24] Varun Kanade and Adam Kalai. Potential-based agnostic boosting. In *Advances in neural information processing systems*, pages 880–888, 2009.

[25] Alessandro Lazaric and Rémi Munos. Hybrid stochastic-adversarial on-line learning. In *Conference on Learning Theory*, 2009.

[26] Christian Leistner, Amir Saffari, Peter M Roth, and Horst Bischof. On robustness of on-line boosting-a competitive study. In *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1362–1369. IEEE, 2009.

[27] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *arXiv preprint arXiv:1804.09554*, 2018.

[28] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[30] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic and constrained adversaries. *arXiv preprint arXiv:1104.5070*, 2011.

[31] Robert E Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

[32] Bruno Scherrer and Matthieu Geist. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2014.

[33] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

[34] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[35] Yuanhao Wang, Ruosong Wang, and Sham M Kakade. An exponential lower bound for linearly-realizable mdps with constant suboptimality gap. *arXiv preprint arXiv:2103.12690*, 2021.

[36] Gellert Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, and Csaba Szepesvári. On query-efficient planning in mdps under linear realizability of the optimal state-value function. *arXiv preprint arXiv:2102.02049*, 2021.

[37] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[38] Jiahao Xie, Zebang Shen, Chao Zhang, Hui Qian, and Boyu Wang. Stochastic recursive gradient-based methods for projection-free online learning. *arXiv preprint arXiv:1910.09396*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Notation: List of Symbols

Weak Learning and Boosting

α	Weak learning parameter
T	Number of boosting iterations
N	Number of internal-boosting iterations
M	Number of internal-boosting episodes
$\Gamma[\cdot]$	Policy projection
Π	Policy class
$\mathbb{\Pi}$	Policy-Tree class (w.r.t Π, Γ, N and T)

Markov Decision Process

S	State space
A	Action space
Δ_A	Probability simplex over actions
$Q^\pi(s,a)$	Q function
$V^\pi(s)$	Value function
$d(s_0)$	Initial state distribution
$d_d^\pi(s)$	State-visitation distribution w.r.t π, d
γ	Discount factor
$\mathcal{E}_\mu(\mathbb{\Pi}, \Pi)$	Policy completeness
μ, ν	Used for different initial state distributions
C_∞	Distribution mismatch if $\mu = d_0$
D_∞	Distribution mismatch if $\mu = \nu \neq d_0$

Optimization

$\mathcal{K}$	Decision set
L	Smoothness of the objective
H	Upper bound on the range of function value
D	Upper bound on Euclidean diameter

B Appendix

It is important that the policy that the boosting algorithm outputs can be evaluated efficiently. Towards that end, we give the following claim.

Claim 12. *For any $\pi \in \mathbb{\Pi}(\Pi, N, T)$, $\pi(\cdot|s)$ for any $s \in S$ can be evaluated using TN base policy evaluations and $O(T \times (NA + A \log A))$ arithmetic and logical operations.*

Proof. Since $\pi \in \mathbb{\Pi}(\Pi, N, T)$, it is composed of TN base policies. Producing each aggregated function takes NA additions and multiplications; there are T of these. Each projection takes time equivalent to sorting $|A|$ numbers, due to a water-filling algorithm [13]; these are also T in number. The final linear transformation takes an additional TA operations. □

C RL Boosting via Weak Online Learning

The second model of weak learning we consider requires a stronger assumption, but will give us better sample and oracle complexity bounds henceforth.

Definition 13 (Weak Online Learner)**.** Let $\alpha \in (0,1)$. Consider a class $\mathcal{L}$ of linear loss functions $\ell : \mathbb{R}^A \to \mathbb{R}$. A weak online learning algorithm, for every $M > 0$, incrementally for each timestep computes a policy $\mathcal{W}_m \in \Pi$ and then observes the state-loss pair $(s, \ell_t) \in S \times \mathcal{L}$ such that

$$\sum_{m=1}^{M} \ell_m(\mathcal{W}_m(s_m)) \;\geq\; \alpha \max_{\pi^* \in \Pi} \sum_{m=1}^{M} \ell_m(\pi^*(s_m)) + (1-\alpha) \sum_{m=1}^{M} \ell_m(\pi_{Rand}(s_m)) - R_{\mathcal{W}}(M).$$

Assumption 1 (Weak Online Learning)**.** *The booster has access to a weak online learning oracle (Definition 13) over the policy class* Π*, for some* $\alpha \in (0,1)$.

Remark 14. A similar remark about *natural* distributions applies to the online weak learner. In particular, it is sufficient the guarantee in 13 holds for arbitrary sequence of loss functions with high probability over the sampling of the state from d^π for some $\pi \in \mathbb{\Pi}$. Although stronger than supervised weak learning, this oracle can be interpreted as a relaxation of the online weak learning oracle considered in [8, 9, 18]. A similar model of hybrid adversarial-stochastic online learning was considered in [30, 25, 7]. In particular, it is known [25] that unlike online learning, the capacity of a hypothesis class for this model is governed by its VC dimension (vs. Littlestone dimension).

Algorithm 5 RL Boosting via Weak Online Learning

Initialize a policy $\pi_0 \in \Pi$ arbitrarily.
for $t = 1$ **to** T **do**
  Initialize online weak learners $\mathcal{W}_1, \dots \mathcal{W}^N$.
  for $m = 1$ **to** M **do**
    Execute π_{t-1} once with initial state distribution μ via Algorithm 3, to get $(s_{t,m}, \widehat{Q}_{t,m})$.
    Choose $\tilde{\pi}_{t,m,0} \in \Pi$ arbitrarily.
    for $n = 1$ **to** N **do**
      Set $\tilde{\pi}_{t,m,n} = (1-\eta_{2,n})\tilde{\pi}_{t,m,n-1} + \frac{\eta_{2,n}}{\alpha}\mathcal{W}^n$.
    end for
    Pass to each $\mathcal{W}^n$ the following loss linear $f_{t,m,n}$:

$$f_{t,m,n} = \frac{1}{\beta}\left(y_{t,m,n} - \tilde{\pi}_{t,m,n}(\cdot|s_i)\right).$$

    where $G = \frac{A}{1-\gamma}, \beta = \frac{2\gamma}{(1-\gamma)^3}$ and $f_i, \widehat{Q}_i \in \mathbb{R}^{|A|}$

$$y_i = \arg\min_{y \in \Delta_A} \{-\widehat{Q}_{t,m}^\top y + G \min_{z \in \Delta_A} \|z - y\| + \frac{\|\tilde{\pi}_{t,m,n}(\cdot|s_{t,m}) - y\|^2}{2\beta}\}$$

  end for
  Declare $\pi'_t = \frac{1}{M}\sum_{m=1}^{M} \Gamma\left[\tilde{\pi}_{t,m,N}\right]$.
  Choose $\eta_{1,t} = \min\{1, \frac{2C_\infty}{t}\}$ if $\mu = d_0$ else set $\eta_{1,t} = \sqrt{\frac{8\gamma(1-\gamma)^2}{|A|^2 T}}$.
  Update $\pi_t = (1-\eta_{1,t})\pi_{t-1} + \eta_{1,t}\pi'_t$.
end for
Run each policy π_t for P rollouts to compute an empirical estimate $\widehat{V^{\pi_t}}$ of the expected return.
return $\bar{\pi} := \pi_{t'}$ where $t' = \arg\max_t \widehat{V^{\pi_t}}$.

Theorem 15. *Algorithm 5 samples* TM *episodes of length* $\frac{1}{1-\gamma}\log\frac{TM}{\delta}$ *with probability* $1-\delta$*. In the episodic model, Algorithm 5 guarantees as long as* $T = \frac{16C_\infty^2}{(1-\gamma)^3\varepsilon}$, $N = \left(\frac{16|A|C_\infty}{(1-\gamma)^2\alpha\epsilon}\right)^2$, $M = \max\left\{\frac{1000|A|^2C_\infty^2}{(1-\gamma)^4\varepsilon^2\alpha^2}\log^2 T\delta, \frac{8|A|C_\infty R_{\mathcal{W}}(M)}{(1-\gamma)^2\alpha\varepsilon}\right\}$,$\mu = d_0$*, we have with probability* $1-\delta$

$$V^* - V^\pi \;\leq\; C_\infty \frac{\mathcal{E}(\mathbb{\Pi}, \Pi)}{1-\gamma} + \varepsilon$$

In the ν-reset model, Algorithm 1 guarantees as long as $T = \frac{100D_\infty^2}{(1-\gamma)^6\varepsilon^2}$, $N = \left(\frac{20|A|D_\infty}{(1-\gamma)^3\alpha\epsilon}\right)^2$, $M = \max\left\{\left(\frac{40|A|D_\infty}{(1-\gamma)^3\alpha\varepsilon}\log\frac{T}{\delta}\right)^2, \frac{10|A|D_\infty R_{\mathcal{W}}(M)}{(1-\gamma)^3\alpha\varepsilon}\right\}$,$\mu = \nu$, we have with probability $1-\delta$

$$V^* - V^\pi \;\le\; D_\infty \frac{\mathcal{E}_\nu(\mathbb{\Pi}, \Pi)}{(1-\gamma)^2} + \varepsilon$$

If $R_{\mathcal{W}}(M) = \sqrt{M \log|\mathcal{W}|}$ for some measure of weak learning complexity $|\mathcal{W}|$, the algorithm samples $\tilde{O}\left(\frac{C_\infty^4|A|^2\log|\mathcal{W}|}{(1-\gamma)^7\alpha^2\varepsilon^3}\right)$ episodes in the episodic model, and $\tilde{O}\left(\frac{D_\infty^4|A|^2\log|\mathcal{W}|}{(1-\gamma)^{12}\alpha^2\varepsilon^4}\right)$ in the ν-reset model.

D Analysis for Boosting with Weak Supervised Learning (Proof of Theorem 7)

Theorem (Formal version of Theorem 7)**.** *Algorithm 1 samples TMN episodes of length $\frac{1}{1-\gamma}\log\frac{TMN}{\delta}$ with probability $1-\delta$. In the episodic model, Algorithm 1 guarantees as long as $T = \frac{16C_\infty^2)}{(1-\gamma)^3\varepsilon}$, $N = \left(\frac{16|A|C_\infty}{(1-\gamma)^2\alpha\epsilon}\right)^2$, $M = m\left(\frac{(1-\gamma)^2\alpha\varepsilon}{8C_\infty|A|}, \frac{\delta}{NT}\right)$,$\mu = d_0$, we have with probability $1-\delta$*

$$V^* - V^\pi \;\le\; C_\infty \frac{\mathcal{E}(\mathbb{\Pi}, \Pi)}{1-\gamma} + \varepsilon$$

In the ν-reset model, Algorithm 1 guarantees as long as $T = \frac{8D_\infty^2}{(1-\gamma)^6\varepsilon^2}$, $N = \left(\frac{16|A|D_\infty}{(1-\gamma)^3\alpha\epsilon}\right)^2$, $M = m\left(\frac{(1-\gamma)^3\alpha\varepsilon}{8|A|D_\infty}, \frac{\delta}{2NT}\right)$,$\mu = \nu$, we have with probability $1-\delta$

$$V^* - V^\pi \;\le\; D_\infty \frac{\mathcal{E}_\nu(\mathbb{\Pi}, \Pi)}{(1-\gamma)^2} + \varepsilon$$

If $m(\varepsilon, \delta) = \frac{\log|\mathcal{W}|}{\varepsilon^2}\log\frac{1}{\delta}$ for some measure of weak learning complexity $|\mathcal{W}|$, the algorithm samples $\tilde{O}\left(\frac{C_\infty^6|A|^4\log|\mathcal{W}|}{(1-\gamma)^{11}\alpha^4\varepsilon^5}\right)$ episodes in the episodic model, and $\tilde{O}\left(\frac{D_\infty^6|A|^4\log|\mathcal{W}|}{(1-\gamma)^{18}\alpha^4\varepsilon^6}\right)$ in the ν-reset model.

Proof of Theorem 7. The broad scheme here is to utilize an equivalence between Algorithm 1 and Algorithm 4 on the function V^π (or V_ν^π in the ν-reset model), to which Theorem 9 applies.

To this end, firstly, note V^π is $\frac{1}{1-\gamma}$-bounded. Define a norm $\|\cdot\|_{\infty,1} : \mathbb{R}^{|S|\times|A|} \to \mathbb{R}$ as $\|x\|_{1,\infty} = \max_{s\in S}\sum_{a\in A}|x_{s,a}|$. Further, observe that for any policy $\pi : S \to \Delta_A$, $\|\pi\|_{\infty,1} = 1$. The following lemma specifies the smoothness of V^π in this norm.

Lemma 16. *V^π is $\frac{2\gamma}{(1-\gamma)^3}$-smooth in the $\|\cdot\|_{\infty,1}$ norm.*

To be able to interpret Algorithm 1 as an instantiation of the algorithmic template Algorithm 4 presents, we need to show that π_t' (Line 3-10) serves as an approximate linear optimizer for $\nabla V^{\pi_{t-1}}$. This will imply that the iterates produced by the two algorithms coincide. Indeed, Claim 17 demonstrates that π_t' serves a linear optimizer over gradients of the function V^π; the suboptimality specifies ϵ_0.

Claim 17. *Let $\beta = \sqrt{\frac{1}{\alpha N}}$, and $\eta_{2,n} = \min\{\frac{2}{n}, 1\}$. Then, for any t, π_t' produced by Algorithm 1 satisfies with probability $1-\delta$*

$$\max_{\pi\in\Pi}(\nabla V_\mu^{\pi_{t-1}})^\top(\pi - \pi_t') \;\le\; \frac{2|A|}{(1-\gamma)^2\alpha}\left(\frac{2}{\sqrt{N}} + \varepsilon_W\right)$$

Finally, observe that it is by construction that $\pi_t \in \mathbb{\Pi}$. Therefore, in terms of the previous section, $\mathcal{K}$ is the class of all policies, $\mathcal{K}_1 = \mathbb{\Pi}$, $\mathcal{K}_2 = \Pi$.

In the episodic model, we wish to invoke the second part of Theorem 9. The next lemma establishes gradient-domination properties of V^π to support this.

Lemma 18. *V^π is $\left(C_\infty, \frac{1}{1-\gamma}C_\infty\mathcal{E}(\Pi,\mathbb{\Pi}),\mathbb{\Pi},\Pi\right)$-gradient dominated, i.e. for any $\pi\in\mathbb{\Pi}$:*

$$V^* - V^\pi \;\le\; C_\infty\left(\frac{1}{1-\gamma}\mathcal{E}(\Pi,\mathbb{\Pi}) + \max_{\pi'\in\Pi}(\nabla V^\pi)^\top(\pi'-\pi)\right)$$

Deriving κ,τ from the above lemma along with ϵ_0 from Claim 17, as a consequence of the second part of Theorem 9, we have with probability $1-NT\delta$

$$\begin{aligned}V^* - V^{\bar\pi} \;\le\; & C_\infty\frac{\mathcal{E}(\mathbb{\Pi},\Pi)}{1-\gamma} + \frac{4C_\infty^2}{(1-\gamma)^3T} + \frac{4|A|C_\infty}{(1-\gamma)^2\alpha\sqrt{N}} \\ & + \frac{2|A|C_\infty}{(1-\gamma)^2\alpha}\varepsilon_W.\end{aligned}$$

Similarly, in the ν-reset model, the first part of Theorem 9 provides a local-optimality guarantee for V_ν^π. Lemma 19 provides a bound on the function-value gap (on V^π) provided such local-optimality conditions.

Lemma 19. *For any $\pi\in\mathbb{\Pi}$, we have*

$$V^* - V^\pi \;\le\; \frac{1}{1-\gamma}D_\infty\left(\frac{1}{1-\gamma}\mathcal{E}_\nu(\Pi,\mathbb{\Pi}) + \max_{\pi'\in\Pi}(\nabla V_\nu^\pi)^\top(\pi'-\pi)\right).$$

Again, using the bound on $\max_{\pi'\in\Pi}(\nabla V_\nu^{\bar\pi})^\top(\pi'-\bar\pi)$ Theorem 9 provides, we have that with probability $1-2NT\delta$

$$\begin{aligned}V^* - V^{\bar\pi} \;\le\; & \frac{D_\infty\mathcal{E}_\nu(\mathbb{\Pi},\Pi)}{(1-\gamma)^2} + \frac{2D_\infty}{(1-\gamma)^3\sqrt{T}} \\ & + \frac{2|A|D_\infty}{(1-\gamma)^3\alpha}\left(\frac{2}{\sqrt{N}} + \varepsilon_W\right) \\ & + \frac{48|A|D_\infty}{(1-\gamma)^3\sqrt{P}}\log\frac{1}{\delta}\end{aligned}$$

□

E Analysis for Boosting with Weak Online Learning (Proof of Theorem 15)

Proof of Theorem 15. Similar to the proof of Theorem 7, we establish an equivalence between Algorithm 1 and Algorithm 4 on the function V^π (or V_ν^π in the ν-reset model), to which Theorem 9 applies provided smoothness (see Lemma 16).

Indeed, Claim 20 demonstrates π'_t serves a linear optimizer over gradients of the function V^π, and provides a bound on ϵ_0. As before, observe that it is by construction that $\pi_t\in\mathbb{\Pi}$.

Claim 20. *Let $\beta=\sqrt{\frac{1}{\alpha N}}$, and $\eta_{2,n}=\min\{\frac{2}{n},1\}$. Then, for any t, π'_t produced by Algorithm 5 satisfies with probability $1-\delta$*

$$\max_{\pi\in\Pi}(\nabla V_\mu^{\pi_{t-1}})^\top(\pi-\pi'_t) \;\le\; \frac{2|A|}{(1-\gamma)^2\alpha}\left(\frac{2}{\sqrt{N}} + \frac{R_{\mathcal{W}}(M)}{M} + \sqrt{\frac{16\log\delta^{-1}}{M}}\right)$$

In the episodic model, one may combine the second part of Theorem 9, which provides a bound on function-value gap for gradient dominated functions, which Lemma 18 guarantees, to conclude with probability $1-T\delta$

$$\begin{aligned}V^* - V^{\bar\pi} \;\le\; & \frac{C_\infty\mathcal{E}(\mathbb{\Pi},\Pi)}{1-\gamma} + \frac{4C_\infty^2(\mathbb{\Pi})}{(1-\gamma)^3T} + \frac{4|A|C_\infty}{(1-\gamma)^2\alpha\sqrt{N}} \\ & + \frac{2|A|C_\infty}{(1-\gamma)^2\alpha}\frac{R_{\mathcal{W}}(M)}{M} + \frac{8|A|C_\infty\log\delta^{-1}}{(1-\gamma)^2\alpha\sqrt{M}}.\end{aligned}$$

Similarly, in the ν-reset model, Lemma 19 provides a bound on the function-value gap provided local-optimality conditions, which the first part of Theorem 9 provides for. Again, with probability $1 - T\delta$

$$V^* - V^{\bar{\pi}} \;\leq\; \frac{D_\infty \mathcal{E}_\nu(\mathbb{\Pi}, \Pi)}{(1-\gamma)^2} + \frac{2D_\infty}{(1-\gamma)^3}\left(\frac{1}{\sqrt{T}} + \frac{|A|}{\alpha}\Big(\frac{2}{\sqrt{N}} + \frac{R_{\mathcal{W}}(M)}{M} + \frac{4\log\delta^{-1}}{\sqrt{M}}\Big) + \frac{24|A|}{\sqrt{P}}\log\frac{1}{\delta}\right).$$

□

F Proofs of Supporting Claims

F.1 Guarantees on the sampling algorithm

Proof of Lemma 11. Recall $\nabla_\pi V^\pi$ denotes the gradient with respect to the $|S| \times |A|$-sized representation of the policy π – the functional gradient. Then, using the policy gradient theorem [37, 34], it is given by,

$$\frac{\partial V_\mu^\pi}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_\mu^\pi(s) Q^\pi(s,a). \tag{1}$$

The following sources of randomness are at play in the sampling algorithm (Algorithm 3): the distribution d^π (which encompasses the discount-factor-based random termination, the transition probability, and the stochasticity of π), and the uniform sampling over A. For a fixed s, π, denote by $\mathcal{Q}_s^\pi$ as the distribution over $\widehat{Q^\pi}(s,\cdot) \in \mathbb{R}^A$, induced by all the aforementioned randomness sources. To conclude the claim, observe that by construction

$$\mathbb{E}_{\mathcal{Q}^\pi(s,\cdot)}[\widehat{Q^\pi}(s,\cdot)|\pi, s] = Q^\pi(s,\cdot). \tag{2}$$

□

F.2 Non-convex Frank-Wolfe method (Theorem 9)

Proof of Theorem 9. **Non-convex general case.** Note that for any timestep t, it holds due to smoothness that

$$f(x_t) = f(x_{t-1} + \eta(z_t - x_{t-1})) \tag{3}$$

$$\geq\; f(x_{t-1}) + \eta\nabla_{t-1}^\top(z_t - x_{t-1}) - \eta^2\frac{L}{2}D^2. \tag{4}$$

Let $t' = \arg\min_t f(x_t) - f(x_{t-1})$. Note that by telescoping over function-value differences across successive iterates, we get

$$f(x_{t'}) - f(x_{t'-1}) \leq \frac{1}{T}\Big(f(x_T) - f(x_0)\Big) \leq \frac{2H}{T}.$$

Combining with (4), and plugging in η, we get

$$\begin{aligned}\nabla_{t'-1}^\top(z_{t'} - x_{t'-1}) \;&\leq\; \eta LD^2/2 + \frac{2H}{T\eta} \\ &\leq\; \sqrt{\frac{2LD^2H}{T}}.\end{aligned}$$

To conclude the claim for the non-convex general case, observe that since $z_{t'} = \mathcal{O}(\nabla_{t'-1})$, it follows by the oracle definition that

$$\max_{u\in\mathcal{K}_2} \nabla_{t'-1}^\top u \;\leq\; \nabla_{t'-1}^\top z_{t'} + \epsilon_0.$$

Gradient-dominated case. Let $x^* = \arg\max_{x \in \mathcal{K}} f(x)$ and let $h_t = f(x^*) - f(x_t)$.

$$\begin{aligned}
h_t &\leq h_{t-1} - \eta_t \nabla_{t-1}^\top (z_t - x_{t-1}) + \eta_t^2 \frac{L}{2} D^2 \\
&\quad \text{(by smoothness)} \\
&\leq h_{t-1} - \eta_t \max_{y \in \mathcal{K}_2} \eta_t \nabla_{t-1}^\top (y - x_{t-1}) + \eta_t^2 \frac{L}{2} D^2 + \eta_t \epsilon_0 \\
&\quad \text{(by oracle guarantee)} \\
&\leq h_{t-1} - \frac{\eta_t}{\kappa} (f(x^*) - f(x_{t-1})) + \eta_t^2 \frac{L}{2} D^2 + \eta_t \left(\epsilon_0 + \frac{\tau}{\kappa} \right) \\
&\quad \text{(by gradient domination)} \\
&= \left(1 - \frac{\eta_t}{\kappa} \right) h_{t-1} + \eta_t^2 \frac{L}{2} D^2 + \eta_t \left(\epsilon_0 + \frac{\tau}{\kappa} \right).
\end{aligned}$$

The theorem then follows from the following claim.

Claim 21. *Let $C \geq 1$. Let g_t be a H-bounded positive sequence such that*

$$g_t \leq \left(1 - \frac{\sigma_t}{C} \right) g_{t-1} + \sigma_t^2 D + \sigma_t E.$$

Then choosing $\sigma_t = \min\{1, \frac{2C}{t}\}$ implies $g_t \leq \frac{2C^2 \max\{2D, H\}}{t} + CE$.

□

F.3 Smoothness of value function (Lemma 16)

Proof of Lemma 16. Consider any two policies π, π'. Using the Performance Difference Lemma (Lemma 3.2 in [2], e.g.) and Equation **??**, we have

$$\begin{aligned}
&|V^{\pi'} - V^\pi - \nabla V^\pi (\pi' - \pi)| \\
&\quad = \frac{1}{1-\gamma} \Big| \mathbb{E}_{s \sim d^{\pi'}} \left[Q^\pi(\cdot|s)^\top (\pi'(\cdot|s) - \pi(\cdot|s)) \right] \\
&\qquad - \mathbb{E}_{s \sim d^\pi} \left[Q^\pi(\cdot|s)^\top (\pi'(\cdot|s) - \pi(\cdot|s)) \right] \Big| \\
&\quad \leq \frac{1}{(1-\gamma)^2} \|d^{\pi'} - d^\pi\|_1 \|\pi' - \pi\|_{\infty,1}.
\end{aligned}$$

The last inequality uses the fact that $\max_{s,a} Q^\pi(s,a) \leq \frac{1}{1-\gamma}$. It suffices to show $\|d^{\pi'} - d^\pi\|_1 \leq \frac{\gamma}{1-\gamma} \|\pi' - \pi\|_{\infty,1}$. To establish this, consider the Markov operator $P^\pi(s'|s) = \sum_{a \in A} P(s'|s,a) \pi(a|s)$ induced by a policy π on MDP M. For any distribution d supported on S, we have

$$\begin{aligned}
&\|(P^{\pi'} - P^\pi) d\|_1 \\
&\quad = \sum_{s'} \left| \sum_{s,a} P(s'|s,a) d(s) (\pi'(a|s) - \pi(a|s) \right| \\
&\quad \leq \sum_{s'} P(s'|s,a) \|d\|_1 \|\pi' - \pi\|_{\infty,1} \\
&\quad \leq \|\pi' - \pi\|_{\infty,1}.
\end{aligned}$$

Using sub-additivity of the l_1 norm and applying the above observation t times, we have for any t

$$\|((P^{\pi'})^t - (P^\pi)^t) d\|_1 \leq t \|\pi' - \pi\|_{\infty,1}.$$

Finally, observe that

$$\begin{aligned}\|d^{\pi'} - d^{\pi}\|_1 &\leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\|((P^{\pi'})^t - (P^{\pi})^t)d_0\|_1 \\ &\leq \|\pi' - \pi\|_{\infty,1}(1-\gamma)\sum_{t=0}^{\infty}t\gamma^t \\ &= \frac{\gamma}{1-\gamma}\|\pi' - \pi\|_{\infty,1}.\end{aligned}$$

□

F.4 Gradient domination (Lemma 18 and Lemma 19)

Proof of Lemma 18. Invoking Lemma 4.1 from [2] with $\mu = d_0$, we have

$$\begin{aligned}V^* - V^{\pi} &\leq \left\|\frac{d^{\pi^*}}{d^{\pi}}\right\|_{\infty}\max_{\pi_0}(\nabla V^{\pi})^{\top}(\pi_0 - \pi) \\ &\leq C_{\infty}(\max_{\pi_0}(\nabla V^{\pi})^{\top}\pi_0 - \max_{\pi'\in\Pi}(\nabla V^{\pi})^{\top}\pi' \\ &\quad + \max_{\pi'\in\Pi}(\nabla V^{\pi})^{\top}(\pi' - \pi)).\end{aligned}$$

Finally, with the aid of Equation **??**, observe that

$$\begin{aligned}&\max_{\pi_0}(\nabla V^{\pi})^{\top}\pi_0 - \max_{\pi'\in\Pi}(\nabla V^{\pi})^{\top}\pi' \\ &\quad= \min_{\pi'\in\Pi}\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}}\left[\max_a Q^{\pi}(s,a) - Q^{\pi}(\cdot|s)^{\top}\pi'\right] \\ &\quad\leq \frac{1}{1-\gamma}\mathcal{E}(\Pi,\mathbb{\Pi}).\end{aligned}$$

□

Proof of Lemma 19. Invoking Lemma 4.1 from [2] with $\mu = \nu$, we have

$$\begin{aligned}&V^* - V^{\pi} \\ &\quad\leq \frac{1}{1-\gamma}\left\|\frac{d^{\pi^*}}{\nu}\right\|_{\infty}\max_{\pi_0}(\nabla V_{\nu}^{\pi})^{\top}(\pi_0 - \pi) \\ &\quad\leq \frac{1}{1-\gamma}D_{\infty}(\max_{\pi_0}(\nabla V_{\nu}^{\pi})^{\top}\pi_0 - \max_{\pi'\in\Pi}(\nabla V_{\nu}^{\pi})^{\top}\pi' \\ &\qquad+ \max_{\pi'\in\Pi}(\nabla V_{\nu}^{\pi})^{\top}(\pi' - \pi)).\end{aligned}$$

Again, with the aid of Equation **??**, observe that

$$\begin{aligned}&\max_{\pi_0}(\nabla V_{\nu}^{\pi})^{\top}\pi_0 - \max_{\pi'\in\Pi}(\nabla V_{\nu}^{\pi})^{\top}\pi' \\ &\quad= \min_{\pi'\in\Pi}\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\nu}^{\pi}}\left[\max_a Q^{\pi}(s,a) - Q^{\pi}(\cdot|s)^{\top}\pi'\right] \\ &\quad\leq \frac{1}{1-\gamma}\mathcal{E}_{\nu}(\Pi,\mathbb{\Pi}).\end{aligned}$$

□

F.5 Supervised linear optimization guarantees

Proof of Claim 10. The internal boosting subroutine of Algorithm 1, that is presented in Algorithm 5, is an instantiation of Algorithm 3 from [18], specializing the decision set to be Δ_A. To note the equivalence, note that in [18] the algorithm is stated assuming that the center-of-mass of the decision

set is at the origin (after a coordinate transform); correspondingly, the update rule in Algorithm 1 can be written as

$$(\tilde{\pi}_n - \pi) = (1 - \eta_{2,n})(\tilde{\pi}_{n-1} - \pi) + \frac{\eta_{2,n}}{\alpha}(\mathcal{A}_{t,n} - \pi).$$

For any state s, $\pi(\cdot|s) = \frac{1}{A}\mathbf{1}_{|A|}$ corresponds to the center-of-mass of Δ_A. Finally, note that maximizing $f^\top x$ over $x \in \mathcal{K}$ is equivalent to minimizing $(-f)^\top x$ over the same domain. Therefore, we can apply previous result on boosting for statistical learning from [18] (Theorem 13). Note that $\widehat{Q^\pi}(s,\cdot)$ produced by Algorithm 3 satisfies $\|\widehat{Q^\pi}(s,\cdot)\| = \frac{|A|}{1-\gamma}$. Let $\mathcal{D}_t$ be the distribution induced by the trajectory sampler in round t. This yields the bound in the claim. □

Proof of Claim 17. Lemma 11 allows us to restate the guarantees from Claim 10 in terms of linear optimization over functional gradients. The conclusion thus follows immediately by combining Lemma 11 and Theorem 10. □

F.6 Online linear optimization guarantees (Claim 20)

Proof of Claim 20. In a similar vein to the proof of Claim 17, here we state the a result on boosting for online convex optimization (OCO) from [18] (Theorem 6), the counterpart of Theorem **??** for the online weak learning case.

Theorem 22. *Let $\beta = \sqrt{\frac{1}{\alpha N}}$, and $\eta_{2,n} = \min\{\frac{2}{n}, 1\}$. Then, for any t, $\Gamma[\tilde{\pi}_{t,m,N}]$ produced by Algorithm 5 satisfies*

$$\max_{\pi\in\Pi} \sum_{m=1}^{M} \left[\hat{Q}_{t,m}^\top \pi(s_{t,m})\right] - \sum_{m=1}^{M} \left[\hat{Q}_{t,m}^\top \Gamma[\tilde{\pi}_{m,N}](s_{t,m})\right] \;\le\; \frac{2|A|}{(1-\gamma)\alpha}\left(\frac{2M}{\sqrt{N}} + R_{\mathcal{W}}(M)\right).$$

Next we invoke online-to-batch conversions. Note that in Algorithm 5, $(s_{t,m}, \hat{Q}_{t,m})$ for any fixed t is sampled i.i.d. from the same distribution. Therefore, we can apply online-to-batch results, i.e. Theorem 9.5 in [16], on Theorem 22 to get

$$\max_{\pi\in\Pi} \mathbb{E}_{(s,Q)\sim\mathcal{D}_t}\left[Q^\top \pi(s)\right] - \mathbb{E}_{(s,Q)\sim\mathcal{D}_t}\left[Q^\top \pi'_t(s)\right] \;\le\; \frac{2|A|}{(1-\gamma)\alpha}\left(\frac{2}{\sqrt{N}} + \frac{R_{\mathcal{W}}(M)}{M} + \sqrt{\frac{16\log\delta^{-1}}{M}}\right).$$

We finally invoke Lemma 11. □

F.7 Remaining proofs (Claim 21)

Proof of Claim 21. Let $T^* = \arg\max_t\{t : t \le 2C\}$. For any $t \le T^*$, we have $\sigma_t = 1$ and $g_t \le H \le \frac{2C^2 H}{t}$. For $t \ge T^*$, we proceed by induction. The base case ($t = T^*$) is true by the previous display. Now, assume $g_{t-1} \le \frac{2C^2\max\{2D,H\}}{t-1} + CE$ for some $t > T^*$.

$$\begin{aligned}
g_t \;&\le\; \left(1 - \frac{2}{t}\right)\left(\frac{2C^2\max\{2D,H\}}{t-1} + CE\right) \\
&\quad + \frac{4C^2 D}{t^2} + \frac{2CE}{t} \\
&\le\; CE + 2C^2\max\{2D,H\}\left(\frac{1}{t-1}\left(1 - \frac{2}{t}\right) + \frac{1}{t^2}\right) \\
&= CE + 2C^2\max\{2D,H\}\frac{t^2 - 2t + t - 1}{t^2(t-1)} \\
&\le\; CE + 2C^2\max\{2D,H\}\frac{t(t-1)}{t^2(t-1)}.
\end{aligned}$$

□