Convergence of Value Aggregation for Imitation Learning

Ching-An Cheng
Institute for Robotics and Intelligent Machines
Georgia Institute of Technology
Atlanta, GA 30332
cacheng@gatech.edu

Byron Boots
Institute for Robotics and Intelligent Machines
Georgia Institute of Technology
Atlanta, GA 30332
bboots@cc.gatech.edu

Abstract

Value aggregation is a general framework for solving imitation learning problems. Based on the idea of data aggregation, it generates a policy sequence by iteratively interleaving policy optimization and evaluation in an online learning setting. While the existence of a good policy in the policy sequence can be guaranteed non-asymptotically, little is known about the convergence of the sequence or the performance of the last policy. In this paper, we debunk the common belief that value aggregation always produces a convergent policy sequence. Moreover, we identify a critical stability condition for convergence and provide a tight non-asymptotic bound on the performance of the last policy. These new theoretical insights let us stabilize problems with regularization, which removes the inconvenient process of identifying the best policy in the policy sequence in stochastic problems.

1 Introduction

Reinforcement learning (RL) is a general framework for solving sequential decision problems (Sutton and Barto, 1998). Using policy gradient methods, it has demonstrated impressive results in GO (Silver et al., 2016) and video-game playing (Mnih et al., 2013). However, due its generality, it can be difficult to learn a policy sample-efficiently or to characterize the performance of the found policy, which is critical in applications that involves real-world costs, such as robotics (Pan et al., 2017). To better exploit the domain knowledge about a problem, one popular approach is imitation learning (IL) (Pomerleau, 1989). In this framework, instead of learning a policy from scratch, it leverages a black-box policy π^*, called the *expert*, from which the learner can query demonstrations. The goal of IL is to identify a policy π such that its performance is similar to or better than π^*.

A recent approach to IL is based on the idea of data aggregation and online learning (Ross et al., 2011; Sun et al., 2017). The main idea is as follows: the algorithm starts with an empty dataset and an initial policy; in the nth iteration, the algorithm uses the current policy π_n to gather new training data into the current dataset and then a supervised learning problem is solved on the updated dataset to compute the next policy π_{n+1}. By interleaving the optimization and the data collection processes in an online fashion, it overcomes the covariate shift problem in the traditional batch IL (Ross et al., 2011).

This family of algorithms can be realized under the general framework of value aggregation (Ross and Bagnell, 2014), which has gained increasing attention due to its non-asymptotic performance guarantee. After N iterations, a good policy π *exists* in the generated policy sequence $\{\pi_n\}_{n=1}^N$ with performance $J(\pi) \leq J(\pi^*) + T\epsilon + \tilde{O}(\frac{1}{N})$, where J is the performance index, ϵ is describes error due lack of expressiveness of the policy class, and T is the horizon of the problem. While this result seems strong at the first glance, its guarantee concerns only the existence of a good policy and therefore is not ideal for stochastic problems. In other words, in order to find the best policy in $\{\pi_n\}_{n=1}^N$ without

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

incurring large statistical errors, a sufficient amount of data has to be acquired in each iteration, or all policies have to be memorized for a final evaluation with another large dataset (Ross et al., 2011).

This inconvenience incentivizes practitioners to just return the last policy π_N (Laskey et al., 2017), and, anecdotally, the last policy π_N has been reported to have good empirical performance (Ross et al., 2013; Pan et al., 2017). Supporting this heuristic is the insight that the last policy π_N is trained with *all* observations and therefore *ideally* should perform the best. Indeed, such idealism works when all the data are sampled i.i.d., as in the traditional batch learning problems (Vapnik, 1998). However, because here new data is collected using the updated policy in each iteration, whether such belief applies depends on the convergence of the distributions generated by the policy sequence.

While Ross and Bagnell (2014) alluded that "...the distribution of visited states converges over the iterations of learning.", we show this is *not* always true—the convergence is rather problem-dependent. In this paper, we identify a critical stability constant θ that determines the convergence of the policy sequence. We show that there is a simple example in which the policy sequence diverges when $\theta > 1$, and we prove that the sequence always converges when $\theta < 1$. Moreover, we provide a tight non-asymptotic bound on the performance of the last policy π_N in both deterministic and stochastic problems.

In Section 2 and 3, we first define our problem of interest and provide a concise introduction to value aggregation. In Section 4, we give the simple counter-example that motivates our main analysis in Section 5, in which we provide conditions for convergence and performance guarantees. Additionally, we provide ways to stabilize the problem by regularization in Section 6 and discuss potential implications and applications of our analysis in Section 7.

2 Problem Setup

We consider solving a discrete-time RL problem. Let $\mathbb{S}$ be the state space and $\mathbb{A}$ be the action space of an agent. Let Π be the class of policies and let T be the length of the planning horizon. In this paper, we restrict ourselves to finite-horizon, continuous-valued problems and deterministic policies.[1] The objective of the agent is to search for a policy $\pi \in \Pi$ to minimize an accumulated cost $J(\pi)$:

$$\min_{\pi\in\Pi} J(\pi) := \min_{\pi\in\Pi} \mathbb{E}_{\rho_\pi}\left[\sum_{t=0}^{T-1} c_t(s_t, a_t)\right] \tag{1}$$

in which c_t is the instantaneous cost at time t, and ρ_π denotes the trajectory distribution of $(s_t, a_t) \in \mathbb{S} \times \mathbb{A}$, for $t = 1, \dots, T$, under policy $a_t = \pi(s_t)$ given a prior distribution $p_0(s_0)$.

For notation: we denote $Q_{\pi|t}(s, a)$ as the Q-function at time t under policy π and $V_{\pi|t}(s) = \mathbb{E}_{a\sim\pi}[Q_{\pi|t}(s, a)]$ as the associated value function. In addition, we introduce some shorthand: we denote $d_{\pi|t}(s)$ as the state distribution at time t generated by running the policy π for the first t steps, and define a joint distribution $d_\pi(s, t) = d_{\pi|t}(s)U(t)$, where $U(t)$ is the uniform distribution over the set $\{0, \dots, T-1\}$. Due to space limitations, we will often omit explicit dependencies on random variables in expectations, e.g. we will write $\min_{\pi\in\Pi} \mathbb{E}_{d_\pi}\mathbb{E}_\pi [c_t]$ to denote $\min_{\pi\in\Pi} \mathbb{E}_{s,t\sim d_\pi}\mathbb{E}_{a\sim\pi} [c_t(s, a)]$, which, by definition of d_π, can be shown to be equivalent to the RL problem in (1).

3 Value Aggregation

Solving general RL problems is challenging. In this paper, we focus on a particular scenario, in which the agent, or the *learner*, has access to an *expert* policy π^* from which the learner can query demonstrations. Here we embrace a general notion of expert. While it is often preferred that the expert is nearly optimal in (1), the expert here can be *any* policy, e.g. the agent's initial policy. Note, additionally, that the RL problem considered here is not necessarily directly related to a real-world application; it can be a surrogate problem which arises in solving the true problem.

The goal of IL is to find a policy π that outperforms or behaves similarly to the expert π^* in the sense that $J(\pi) \leq J(\pi^*) + O(T)$. That is, we treat IL as performing a robust, approximate policy iteration

[1] A similar derivation can be applied to problems with a discounted infinite-horizon, discrete-valued spaces, and stochastic policies.

from π^*: ideally IL should lead to a policy that outperforms the expert, but it at least returns a policy that performs similarly to the expert.

AGGREVATE (Aggregate Value to Imitate) is an IL algorithm proposed by Ross and Bagnell (2014) based on the idea of online learning (Hazan et al., 2016). Here we give a compact derivation and discuss its important features in preparation for the analysis in Section 5. To this end, we introduce the performance difference lemma due to Kakade and Langford (2002), which will be used as the foundation to derive AGGREVATE.

Lemma 1. *(Kakade and Langford, 2002) Let π and π' be two policies and $A_{\pi'|t}(s,a) = Q_{\pi'|t}(s,a) - V_{\pi'|t}(s)$ be the (dis)advantage function at time t with respect to running π'. Then it holds that*

$$J(\pi) = J(\pi') + T\mathbb{E}_{s,t\sim d_\pi}\mathbb{E}_{a\sim\pi}[A_{\pi'|t}(s,a)]. \tag{2}$$

3.1 Motivation

The main idea of AGGREVATE is to minimize the performance difference between the learner's policy and the expert policy, which by Lemma 1 is given as $\frac{1}{T}\left(J(\pi) - J(\pi^*)\right) = \mathbb{E}_{d_\pi}\mathbb{E}_\pi[A_{\pi^*|t}(s,a)]$. AGGREVATE can be viewed as solving an RL problem with $A_{\pi^*|t}(s,a)$ as the instantaneous cost at time t:

$$\min_{\pi\in\Pi} \mathbb{E}_{d_\pi}\mathbb{E}_\pi\left[A_{\pi^*|t}\right]. \tag{3}$$

Although the transformation from (1) to (3) seems trivial, it unveils some critical properties. Most importantly, the range of the problem in (3) is normalized. For example, regardless of the original definition of c_t, if $\Pi \ni \pi^*$, there exists at least a policy $\pi \in \Pi$ such that (3) is non-positive (i.e. $J(\pi) \leq J(\pi^*)$). As now the problem (3) is relative, it becomes possible to place a qualitative assumption to bound the performance in (3) in terms of some measure of expressiveness of the policy class Π.

We formalize this idea into Assumption 1, which is one of the core assumptions implicitly imposed by Ross and Bagnell (2014).[2] To simplify the notation, we define a function F such that for any two policies π, π'

$$F(\pi',\pi) := \mathbb{E}_{d_{\pi'}}\mathbb{E}_\pi\left[A_{\pi^*|t}\right] \tag{4}$$

This function captures the main structure in (3). By separating the roles of π' (which controls the state distribution) and π (which controls the reaction/prediction), the performance of a policy class Π relative to an expert π^* can be characterized with the approximation error in a supervised learning problem.

Assumption 1. Given a policy π^*, the policy class Π satisfies that for arbitrary sequence of policies $\{\pi_n \in \Pi\}_{n=1}^N$, there exists a small constant ϵ_{Π,π^*} such that

$$\min_{\pi\in\Pi} \frac{1}{N} f_{1:N}(\pi) \leq \epsilon_{\Pi,\pi^*}, \tag{5}$$

where $f_n(\pi) := F(\pi_n, \pi)$ and $f_{1:n}(\pi) = \sum_{n=1}^N f_n(\pi)$.

This assumption says that there exists at least a policy $\pi \in \Pi$ which is as good as π^* in the sense that π can predict π^* well in a cost-sensitive supervised learning problem, with small error ϵ_{Π,π^*}, under the average distribution generated by arbitrary sequence $\{\pi_n \in \Pi\}_{n=1}^N$.

Following this assumption, AGGREVATE exploits another critical structural property of the problem.

Assumption 2. $\forall \pi' \in \Pi$, $F(\pi', \pi)$ is a strongly convex function in π.

While Ross and Bagnell (2014) did not explicitly discuss under which condition Assumption 2 holds, here we point out some examples.We further note that AGGREVATE has demonstrated impressive empirical success even when Assumption 2 cannot be verified (Sun et al., 2017; Pan et al., 2017).

Proposition 1. *Suppose Π consists of deterministic linear policies (i.e. $a = \phi(s)^T x$ for some feature map $\phi(s)$ and weight x) and $\forall s \in \mathbb{S}$, $c_t(s,\cdot)$ is strongly convex. Assumption 2 holds under any of the following: 1) $V_{\pi^*|t}(s)$ is constant over $\mathbb{S}$ (in this case $A_{\pi^*|t}(s,a)$ is equivalent to $c_t(s,a)$ up to a constant in a) 2) The problem is continuous-time and the dynamics are affine in action.*

[2] The assumption is implicitly made when Ross and Bagnell (2014) assume the existence of ϵ_{class} in Theorem 2.1 on page 4.

3.2 Algorithm and Performance

Given Assumption 2, AggreVaTe treats $f_n(\cdot)$ as the per-round cost in an online learning problem and updates the policy sequence as follows: Let π_1 be an initial policy. In the nth iteration of AggreVaTe, the policy is updated by[3]

$$\pi_{n+1} = \arg\min_{\pi\in\Pi} f_{1:n}(\pi). \tag{6}$$

After N iterations, the best policy in the sequence $\{\pi_n\}_{n=1}^N$ is returned, i.e. $\pi = \hat{\pi}_N$, where

$$\hat{\pi}_N := \arg\min_{\pi\in\{\pi_n\}_{n=1}^N} J(\pi). \tag{7}$$

As the update rule (6) (aka Follow-the-Leader) has sublinear regret, it can be shown that (cf. Section 5.1) $J(\hat{\pi}_N) \leq J(\pi^*) + T\left(\epsilon_{\text{class}} + \epsilon_{\text{regret}}\right)$, in which $\epsilon_{\text{regret}} = \tilde{O}(\frac{1}{N})$ is the average regret, and $\epsilon_{\text{class}} := \min_{\pi\in\Pi} \frac{1}{N}\sum_{n=1}^N \mathbb{E}_{d_{\pi_n}}\left[\mathbb{E}_\pi[Q_{\pi^*|t}] - \mathbb{E}_{\pi^*}[Q_{\pi^*|t}]\right]$ compares the best policy in the policy class Π and the expert policy π^*. The term ϵ_{class} can be negative if there exists a policy in Π that is better than π^* under the average distribution, $\frac{1}{N}\sum_{n=1}^N d_{\pi_n}$, generated by AggreVaTe. By Assumption 1, $\epsilon_{\text{class}} \leq \epsilon_{\Pi,\pi^*}$; we know ϵ_{class} at least should be small.

The performance bound above satisfies the requirement of IL that $J(\hat{\pi}_N) \leq J(\pi^*) + O(T)$. Especially because ϵ_{class} can be non-positive, AggreVaTe can be viewed as robustly performing one approximate policy iteration from π^*. One notable special case of AggreVaTe is DAgger (Ross et al., 2011). DAgger tackles the problem of solving an unknown RL problem by imitating a desired policy π^*. The reduction to AggreVaTe can be seen by setting $c_t(s,a) = \mathbb{E}_{a^*\sim\pi^*}[\|a - a_t^*\|]$ in (1). In this case, π^* is optimal for this specific choice of cost and therefore $V_{\pi^*|t}(s) = 0$. By Proposition 1, $A_{\pi^*|t}(s,a) = c_t(s,a)$ and ϵ_{class} reduces to $\min_{\pi\in\Pi}\frac{1}{N}\sum_{n=1}^N \mathbb{E}_{d_{\pi_n}}\mathbb{E}_\pi[c_t] \geq 0$, which is related to the expressiveness of the policy class Π.

4 Guarantee on the Last Policy?

The performance bound in Section 3 implicitly assumes that the problem is either deterministic or that infinite samples are available in each iteration. For stochastic problems, $f_{1:n}$ can be approximated by finite samples or by function approximators (Ross and Bagnell, 2014). Suppose m samples are collected in each iteration to approximate f_n. An additional error in $O(\frac{1}{\sqrt{mN}})$ will be added to the performance of $\hat{\pi}_N$. However, in practice, another constant statistical error[4] in $O(\frac{1}{m})$ is introduced when one attempts to identify $\hat{\pi}_N$ from the sequence $\{\pi_n\}_{n=1}^N$.

This practical issue motivates us to ask whether a similar guarantee applies to the last policy π_N so that the selection process to find $\hat{\pi}_N$ can be removed. In fact, the last policy π_n has been reported to have good performance empirically (Ross et al., 2013; Pan et al., 2017). It becomes interesting to know what can one say about π_N. It turns out that running AggreVaTe does not always yield a policy sequence $\{\pi_n\}$ with reasonable performance, as given in the example below.

A Motivating Example Consider a two-stage deterministic optimal control problem:

$$\min_{\pi\in\Pi} J(\pi) = \min_{\pi\in\Pi} c_1(s_1,a_1) + c_2(s_2,a_2) \tag{8}$$

where the transition and costs are given as

$$s_1 = 0, \quad s_2 = \theta(s_1 + a_1),$$
$$c_1(s_1,a_1) = 0, \quad c_2(s_2,a_2) = (s_2 - a_2)^2.$$

[3] We adopt a different notation from Ross and Bagnell (2014), in which the per-round cost $\mathbb{E}_{d_{\pi_n}}\mathbb{E}_\pi\left[Q_{\pi^*|t}\right]$ was used. Note these two terms are equivalent up to an additive constant, as the optimization here is over π with π_n fixed.

[4] The original analysis in the stochastic case by Ross and Bagnell (2014) only guarantees the existence of a good policy in the sequence. The $O(\frac{1}{m})$ error is due to identifying the best policy (Lee et al., 1998) (as the function is strongly convex) and the $O(\frac{1}{\sqrt{mN}})$ error is the generalization error (Cesa-Bianchi et al., 2004).

Since the problem is deterministic, we consider a policy class Π consisting of open-loop stationary deterministic policies, i.e. $a_1 = a_2 = x$ for some x (for convenience π and x will be used interchangeably). It can be easily seen that Π contains a globally optimal policy, namely $x = 0$. We perform AGGREVATE with a feedback expert policy $a_t^* = s_t$ and some initial policy $|x_1| > 0$. While it is a custom to initialize $x_1 = \arg\min_{x \in \mathcal{X}} F(x^*, x)$ (which in this case would ideally return $x_1 = 0$), setting $|x_1| > 0$ simulates the effect of finite numerical precision.

We consider two cases ($\theta > 1$ or $\theta < 1$) to understand the behavior of AGGREVATE. First, suppose $\theta > 1$. Without loss generality, take $\theta = 10$ and $x_1 = 1$. We can see running AGGREVATE will generate a divergent sequence $x_2 = 10, x_3 = 55, x_4 = 220\dots$ (in this case AGGREVATE would return x_1 as the best policy). Since $J(x) = (\theta - 1)^2 x^2$, the performance $\{J(x_n)\}$ is an increasing sequence. Therefore, we see even in this simple case, which can be trivially solved by gradient descent in $O(\frac{1}{n})$, using AGGREVATE results in a sequence of policies with degrading performance, though the policy class Π includes a globally optimal policy. Now suppose on the contrary $\theta < 1$. We can see that $\{x_n\}$ asymptotically converges to $x^* = 0$.

This example illustrates several important properties of AGGREVATE. It showcases that whether AGGREVATE can generate a reasonable policy sequence depends on some intrinsic property of the problem (i.e. the value of θ). In addition, it shows that ϵ_{Π,π^*} can be large while Π contains an optimal policy. This suggests that Assumption 1 is too strong.

5 Theoretical Analysis

Motivated by the example in Section 4, we investigate the convergence of the policy sequence generated by AGGREVATE in general problems. We assume the policy class Π consists of policies parametrized by some parameter $x \in \mathcal{X}$, in which $\mathcal{X}$ is a convex set in a normed space with norm $\|\cdot\|$ (and $\|\cdot\|_*$ as its dual norm). With abuse of notation, we abstract the RL problem in (3) as

$$\min_{x \in \mathcal{X}} F(x, x) \tag{9}$$

where we overload the notation $F(\pi', \pi)$ defined in (4) as $F(\pi', \pi) = F(y, x)$ when $\pi, \pi' \in \Pi$ are parametrized by $x, y \in \mathcal{X}$, respectively. Similarly, we will write $f_n(x) = F(x_n, x)$ for short. In this new notation, AGGREVATE's update rule in (6) can be simply written as $x_{n+1} = \arg\min_{x \in \mathcal{X}} f_{1:n}(x)$.

Here we will focus on the bound on $F(x, x)$, because, for π parameterized by x, this result can be directly translated to a bound on $J(\pi)$: by definition of F in (4) and Lemma 1, $J(\pi) = J(\pi^*) + TF(\pi, \pi)$. For simplicity, we will assume for now F is deterministic; the convergence in stochastic problems will be discussed at the end of the section.

5.1 Classical Result

For completeness, we restate the structural assumptions made by AGGREVATE in terms of $\mathcal{X}$ and present the known convergence of AGGREVATE (Ross and Bagnell, 2014). The proof is given in Appendix.

Assumption 3. Let ∇_2 denote the derivative with respect to the second argument.

1. F is uniformly α-strongly convex in the second argument: $\forall x, y, z \in \mathcal{X}$, $F(z, x) \geq F(z, y) + \langle \nabla_2 F(z, y), x - y \rangle + \frac{\alpha}{2}\|x - y\|^2$.
2. F is uniformly G_2-Lipschitz continuous in the second argument: $\forall x, y, z \in \mathcal{X}$, $|F(z, x) - F(z, y)| \leq G_2 \|x - y\|$.

Assumption 4. $\forall \{x_n \in \mathcal{X}\}_{n=1}^N$, there exists a small constant ϵ_{Π,π^*} such that $\min_{x \in \mathcal{X}} \frac{1}{N} f_{1:N}(x) \leq \epsilon_{\Pi,\pi^*}$.

Theorem 1. *Under Assumption 3 and 4,* AGGREVATE *generates a sequence such that, for all* $N \geq 1$,

$$F(\hat{x}_N, \hat{x}_N) \leq \frac{1}{N} \sum_{n=1}^{N} f_n(x_n) \leq \epsilon_{\Pi,\pi^*} + \frac{G_2^2}{2\alpha} \frac{\ln(N) + 1}{N}$$

5.2 New Structural Assumptions

AGGREVATE can be viewed as an attempt to solve the optimization problem in (9) without any information (not even continuity) regarding how $F(x, x)$ changes with perturbations in the first argument. Because making even a local improvement for general Lipschitz continuous problem is known to be NP-hard (Nesterov, 2013), the performance of AGGREVATE is mainly due to Assumption 4, which implies the existence of some good policy. Therefore, to analyze the performance of the last iterate x_N, we need additional structure on F.

Here we introduce a continuity assumption.

Assumption 5. $\nabla_2 F$ is uniformly β-Lipschitz continuous in the first argument: $\forall x, y, z \in \mathcal{X}$ $\|\nabla_2 F(x, z) - \nabla_2 F(y, z)\|_* \leq \beta \|x - y\|$.

Because the first argument of F in (4) defines the change of state distribution, Assumption 5 basically requires that the expectation over d_π changes continuously with respect to π, which is satisfied in most RL problems. Intuitively, this quantifies the difficulty of a problem in terms of how sensitive the state distribution is to policy changes.

In addition, we relax Assumption 4. As shown in Section 4, Assumption 4 is sometimes too strong, because it might not be satisfied even when Π contains a globally optimal policy. In the analysis of convergence, we instead rely on a necessary condition of Assumption 4 (i.e. $\tilde{\epsilon}_{\Pi,\pi^*} \leq \epsilon_{\Pi,\pi^*}$), which is satisfied by the example in Section 4.

Assumption 6. Let π be a policy parametrized by x. There exists a small constant $\tilde{\epsilon}_{\pi,\pi^*}$ such that $\forall x \in \mathcal{X}$, $\min_{y \in \mathcal{X}} F(x, y) \leq \tilde{\epsilon}_{\Pi,\pi^*}$.

5.3 Guarantee on the Last Policy

In our analysis, we define a stability constant $\theta = \frac{\beta}{\alpha}$. One can verify that this definition agrees with the θ in the example in Section 4. This stability constant will play a crucial role in determining the convergence of $\{x_n\}$, similar to the spectral norm of the Jacobian matrix in discrete-time dynamical systems (Antsaklis and Michel, 2007). We have already shown above that if $\theta > 1$ there is a problem such that AGGREVATE generates a divergent sequence $\{x_n\}$ with degrading performance over iterations. We now show that if $\theta < 1$, then $\lim_{n \to \infty} F(x_n, x_n) \leq \tilde{\epsilon}_{\Pi,\pi^*}$ and moreover $\{x_n\}$ is convergent. The proof in given in Appendix.

Theorem 2. *Suppose Assumption 3, 5, and 6 are satisfied. Let $\theta = \frac{\beta}{\alpha}$. Then for all $N \geq 1$ it holds*

$$F(x_N, x_N) \leq \tilde{\epsilon}_{\Pi,\pi^*} + \frac{\left(\theta e^{1-\theta} G_2\right)^2}{2\alpha} N^{2(\theta-1)}$$

and $\|x_N - \bar{x}_N\| = \frac{G_2 e^{1-\theta}}{\alpha} N^{\theta-1}$, where $\bar{x}_N = \frac{1}{N} x_{1:N}$. In particular, if $\theta < 1$, then $\{x_n\}_{n=1}^\infty$ is convergent

Theorem 2 implies that the stability and convergence of AGGREVATE depends solely on the problem properties. If the state distribution d_π is sensitive to minor changes of policy, running AGGREVATE would fail to provide any guarantee on the last policy. Moreover, Theorem 2 also characterizes the performance of the average policy $\bar{x}_N$ when $\theta < 1$, .

The upper bound in Theorem 2 is tight as indicated in the next theorem. Note a lower bound on $F(x_N, x_N)$ leads directly to a lower bound on $J(\pi_N)$, for π_N is parametrized by x_N.

Theorem 3. *There is a problem such that running* AGGREVATE *for N iterations results in $F(x_N, x_N) \geq \tilde{\epsilon}_{\Pi,\pi^*} + \Omega(N^{2(\theta-1)})$. In particular, if $\theta > 1$, there is a problem in which the policy sequence and performance sequence diverge.*

5.4 Stochastic Problems

We analyze the convergence of AGGREVATE in stochastic problems using finite-sample approximation: Define $f(x; s) = \mathbb{E}_\pi[A_{\pi^*|t}]$ (namely, $f_n(x) = \mathbb{E}_{d_{\pi_n}}[f(x; s)]$, for a policy π is parametrized by x). Instead of using $f_n(\cdot)$ as the per-round cost in the nth iteration, we take its finite samples approximation $g_n(\cdot) = \sum_{k=1}^{m_n} f(\cdot; s_{n,k})$, where m_n is the number of independent samples collected in the nth iteration under distribution d_{π_n}. That is, the update rule in (6) in stochastic setting is modified to $\pi_{n+1} = \arg\min_{\pi\in\Pi} g_{1:n}(\pi)$. Then we have the following result. The proof is based on the concentration of vector-valued martingales (Hayes, 2005), which is technical and therefore omitted.

Theorem 4. *In addition to Assumption 5 and 6, assume $f(x; s)$ is α-strongly convex in x and $\|f(x; s)\|_* < G_2$ almost surely. Let $\theta = \frac{\beta}{\alpha}$ and suppose $m_n = m_0 n^r$ for some $r \geq 0$. For all $N > 0$, with probability at least $1 - \delta$,*

$$F(x_N, x_N) \leq \tilde{\epsilon}_{\Pi,\pi^*} + \tilde{O}\left(\frac{\theta^2}{c}\frac{\ln(1/\delta) + C_{\mathcal{X}}/n}{n^{\min\{r,2,2-2\theta\}}}\right) + \tilde{O}\left(\frac{\ln(1/\delta) + C_{\mathcal{X}}}{cn^{\min\{2,1+r\}}}\right)$$

where $c = \frac{\alpha}{G_2^2 m_0}$ and $C_{\mathcal{X}}$ is a constant[5] *of the complexity of* Π.

The bound in Theorem 4 has a weak dependency on $C_{\mathcal{X}}$ and the major stochastic error is due to $\|\nabla g_n(x_n) - \nabla f_n(x_n)\|_*$, which is bounded by $O(\frac{1}{\sqrt{m_n}})$, as reflected through the dependency on r. Therefore, the growth of sample size m_n over iterations determines the main behavior of AGGREVATE in stochastic problems. For $r = 0$, compared with Theorem 2, Theorem 4 has an additional constant error in $\tilde{O}(\frac{1}{m_0})$, which is comparable to the stochastic error in selecting the best policy in the classical approach. For $r > 0$, by slightly taking more samples over iterations (e.g. $r = 2 - 2\theta$), we see the convergence rate can get closer to $\tilde{O}(N^{2-2\theta})$ as in the ideal case given by Theorem 2. However, it cannot be better than $\tilde{O}(\frac{1}{N})$. Therefore, for stochastic problems, a stability constant $\theta < 1/2$ and a growing rate $r > 1$ does not contribute to faster convergence as opposed to the deterministic case in Theorem 2. Note while our analysis here is based finite-sample approximation. A similar technique can also be applied to the scenario in which only samples of function value $f_n(x_n; s)$ are available and another online regression problem is perform to learn $f_n(\cdot)$ as in the case considered by Ross and Bagnell (2014)

6 Regularization for Stability

We have shown that whether AGGREVATE generates a convergent policy sequence and a last policy with the desired performance depends on the stability constant θ. Here we show that by adding regularization to the problem we can make the problem stable. For simplicity, here we consider deterministic problems or stochastic problems with infinite samples.

6.1 Mixing Policies

We first consider the idea of using mixing policies to collect samples, which was originally proposed as a heuristic in (Ross et al., 2011). It works as follows: in the nth iteration of AGGREVATE, instead of using $F(\pi_n, \cdot)$ as the per-round cost, it uses $\hat{F}(\pi_n, \cdot)$ which is defined by

$$\hat{F}(\pi_n, \pi) = \mathbb{E}_{d_{\tilde{\pi}_n}} \mathbb{E}_{\pi}[A_{\pi^*|t}] \tag{10}$$

The state distribution $d_{\tilde{\pi}_n}(s)$ is generated by running π^* with probability q and π_n with probability $1 - q$ at each time step. Originally, Ross et al. (2011) proposes to set q to decay exponentially over the iterations of AGGREVATE. (The proofs are given in Appendix).

Here we show that the usage of mixing policies has the effect of stabilizing the problem.

Lemma 2. *Let $\|p_1 - p_2\|_1$ denote the total variational distance between distribution p_1 and p_2. Assume*[6] *for any policy π, π' parameterized by x, y it satisfies $\frac{2G_2}{T}\sum_{t=0}^{T-1} \|d_{\pi|t} - d_{\pi'|t}\|_1 \leq \beta\|x - y\|$ and assume $\|\nabla_x \mathbb{E}_\pi[A_{\pi^*|t}](s)\|_* < G_2$.Then $\nabla_2 F$ is uniformly $(1 - q^T)\beta$-Lipschitz continuous in the second argument.*

By Lemma 2, if $\theta > 1$, then choosing $q > (1 - \frac{1}{\theta})^{1/T}$ ensures the stability constant of $\hat{F}$ to be $\hat{\theta} < 1$. However, stabilizing the problem in this way incurs a constant cost as shown in Corollary 1.

Corollary 1. *Suppose $\mathbb{E}_\pi[A_{\pi^*|t}] < M$ for all π. Define $\Delta_N = \frac{(\hat{\theta} e^{1-\hat{\theta}} G_2)^2}{2\alpha} N^{2(\hat{\theta}-1)}$. Then under the assumptions Lemma 2 and Assumption 3.1, running* AGGREVATE *with $\tilde{F}$ in* (10) *and mixing rate q gives $F(x_N, x_N) \leq \Delta_N + \tilde{\epsilon}_{\Pi,\pi^*} + 2M\min(1, Tq)$*

[5] The constant $C_{\mathcal{X}}$ can be thought as $\ln|\mathcal{X}|$, where $|\mathcal{X}|$ measures the size of $\mathcal{X}$ in e.g. Rademacher complexity or covering number (Mohri et al., 2012). For example, $\ln|\mathcal{X}|$ is linear in $\dim \mathcal{X}$.

[6] These two are sufficient to Assumption 3.2 and 5.

6.2 Weighted Regularization

Here we consider another scheme for stabilizing the problem. Suppose F satisfies Assumption 3 and 5. For some $\lambda > 0$, define

$$\tilde{F}(x, x) = F(x, x) + \lambda R(x) \tag{11}$$

in which $R(x)$ is an α-strongly convex regularization term such that $R(x) \geq 0$, $\forall x \in \mathcal{X}$ and $\min_{y \in \mathcal{X}} F(x, y) + \lambda R(y) = (1 + \lambda) O(\tilde{\epsilon}_{\Pi, \pi^*})$. For example, R can be $F(\pi^*, \cdot)$ when π^* is (close) to optimal (e.g. in the case of DAGGER), or $R(x) = \mathbb{E}_{s,t \sim d_{\pi^*}} \mathbb{E}_{a \sim \pi} \mathbb{E}_{a^* \sim \pi^*} [d(a, a^*)]$, where π is a policy parametrized by x and $d(\cdot, \cdot)$ is some metric of space $\mathbb{A}$ (i.e. it uses the distance between π and π^* as regularization).

It can be easily seen that $\tilde{F}$ is uniformly $(1 + \lambda)\alpha$-strongly convex in the second argument and $\nabla_2 \tilde{F}$ is uniformly β-continuous in the second argument. That is, if we choose $\lambda > \theta - 1$, then the stability constant $\tilde{\theta}$ of $\tilde{F}$ satisfies $\tilde{\theta} < 1$.

Corollary 2. *Define* $\Delta_N = \frac{(\tilde{\theta} e^{1-\tilde{\theta}} G_2)^2}{2\alpha} N^{2(\tilde{\theta}-1)}$. *Running* AGGREVATE *with* $\tilde{F}$ *in* (11) *as the per-round cost has performance satisfies: for all* $N > 0$,

$$F(x_N, x_N) \leq (1 + \lambda) \left(O(\tilde{\epsilon}_{\Pi, \pi^*}) + \Delta_N \right)$$

Proof. Because $F(x_N, x_N) = \tilde{F}(x_N, x_N) - \lambda R(x_N)$, the inequality can be proved by applying Theorem 2 to $\tilde{F}(x_N, x_N)$. ■

By Corollary 2, using AGGREVATE to solve a weighted regularized problem in (11) would generate a convergent sequence for λ large enough. Unlike using a mixing policy, here the performance guarantee on the last policy is only worsened by a multiplicative constant on $\tilde{\epsilon}_{\Pi, \pi^*}$, which can be made small by choosing a larger policy class.

The result in Corollary 2 can be strengthened particularly when $R(x) = \mathbb{E}_{s,t \sim d_{\pi^*}} \mathbb{E}_{a \sim \pi} \mathbb{E}_{a^* \sim \pi^*} [d(a, a^*)]$ is used. In this case, it can be shown that $CR(x) \geq F(x, x)$ for some $C > 0$ (usually $C > 1$) (Pan et al., 2017). That is, $F(x, x) + \lambda R(x) \geq (1 + \lambda / C) F(x, x)$. Thus, the multiplicative constant in Corollary 2 can be reduced from $1 + \lambda$ to $\frac{1+\lambda}{1+\lambda/C}$. It implies that simply by adding a portion of demonstrations gathered under the expert's distribution so that the leaner can anchor itself to the expert while minimizing $F(x, x)$, one does not have to find the best policy in the sequence $\{\pi_n\}_{n=1}^N$ as in (7), but just return the last policy π_N.

7 DISCUSSION

We contribute a new analysis of value aggregation, unveiling several interesting theoretical insights. Under a weaker assumption than the classical result, we prove that the convergence of the last policy depends solely on a problem's structural property and provide a tight non-asymptotic bound on its performance in both deterministic and stochastic problems. In addition, using the new theoretical results, we show that the stability of the last policy can be reinforced by additional regularization with minor performance loss. This suggests that under proper conditions a practitioner can just run AGGREVATE and then take the last policy, without performing an additional statistical test to find the best policy required by the classical analysis. In addition, as our results of the last policy are based on the perturbation of gradients, we believe this provide a potential explanation to why AGGREVATE has shown empirical success in non-convex problems with neural-network policies.

While our original aim is to understand the performance of the last policy x_N, we achieve a number of extra outcomes for free. First, the theoretical results of x_N can directly translate to that of the mean policy $\bar{x}_N$ as suggested by Theorem 2. We note this property continues to hold in stochastic problems. Furthermore, our analysis given as Theorem 4 can be viewed as a generalization of the analysis of Empirical Risk Minimization (ERM) to non-i.i.d. scenarios, where the distribution depends on the decision variable. For optimizing a strongly convex objective function with i.i.d. samples, it has been shown by Shalev-Shwartz et al. (2009) that x_N exhibits a fast convergence to the optimal performance in $O(\frac{1}{N})$. By specializing our general result in Theorem 4 with $\theta, r = 0$ to recover the classical i.i.d. setting, we arrives at a bound on the performance of x_N in $\tilde{O}(\frac{1}{N})$, which matches the best known result up to a log factor. However, Theorem 4 is proved by a completely different technique using the martingale concentration of the gradient sequence.

References

Antsaklis, P. J. and Michel, A. N. (2007). *A linear systems primer*, volume 1. Birkhäuser Boston.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

Hayes, T. P. (2005). A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*.

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, volume 2, pages 267–274.

Laskey, M., Chuck, C., Lee, J., Mahler, J., Krishnan, S., Jamieson, K., Dragan, A., and Goldberg, K. (2017). Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations. In *IEEE International Conference on Robotics and Automation*, pages 358–365. IEEE.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1998). The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980.

McMahan, H. B. (2014). A survey of algorithms and analysis for adaptive online learning. *arXiv preprint arXiv:1403.3465*.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Pan, Y., Cheng, C.-A., Saigol, K., Lee, K., Yan, X., Theodorou, E., and Boots, B. (2017). Agile off-road autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*.

Pomerleau, D. A. (1989). Alvinn: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, pages 305–313.

Rakhlin, A. and Sridharan, K. (2013). Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019.

Ross, S. and Bagnell, J. A. (2014). Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*.

Ross, S., Gordon, G. J., and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 627–635.

Ross, S., Melik-Barkhudarov, N., Shankar, K. S., Wendel, A., Dey, D., Bagnell, J. A., and Hebert, M. (2013). Learning monocular reactive uav control in cluttered natural environments. In *IEEE International Conference onRobotics and Automation*, pages 1765–1772. IEEE.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. In *Conference on Learning Theory*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Sun, W., Venkatraman, A., Gordon, G. J., Boots, B., and Bagnell, J. A. (2017). Deeply aggrevated: Differentiable imitation learning for sequential prediction. *arXiv preprint arXiv:1703.01030*.

Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*, volume 135. MIT Press Cambridge.

Vapnik, V. N. (1998). *Statistical learning theory*, volume 1. Wiley New York.

A Proof of Theorem 1

The proof is based on a basic perturbation lemma in convex analysis (Lemma 3), which for example can be found in (McMahan, 2014), and a lemma for online learning (Lemma 4).

Lemma 3. *Let $\phi_1 : \mathbb{R}^d \mapsto \mathbb{R} \bigcup \{\infty\}$ be a convex function such that $x_1 = \arg\min_x \phi_t(x)$ exits. Let ψ be a function such that $\phi_2(x) = \phi_1(x) + \psi(x)$ is α-strongly convex with respect to $\| \cdot \|$. Let $x_2 = \arg\min_x \phi_2(x)$. Then, for any $g \in \partial\psi(x_1)$, we have*

$$\|x_1 - x_2\| \leq \frac{1}{\alpha}\|g\|_*$$

and for any x'

$$\phi_2(x_1) - \phi_2(x') \leq \frac{1}{2\alpha}\|g\|_*^2$$

When ϕ_1 and ψ are quadratics (with ψ possibly linear) the above holds with equality.

Lemma 4. *Let $l_t(x)$ be a sequence of functions. Denote $l_{1:t}(x) = \sum_{\tau=1}^{t} l_\tau(x)$. and let*

$$x_t^* = \arg\min_{x \in K} l_{1:t}(x)$$

Then for any sequence $\{x_1, \dots, x_T\}$, $\tau \geq 1$, and any $x^* \in K$, it holds

$$\begin{aligned}\sum_{t=\tau}^{T} l_t(x_t) \leq l_{1:T}(x_T^*) - l_{1:\tau-1}(x_{\tau-1}^*) \\ + \sum_{t=\tau}^{T} l_{1:t}(x_t) - l_{1:t}(x_t^*)\end{aligned}$$

Proof. Introduce a slack loss function $l_0(\cdot) = 0$ and define $x_0^* = 0$ for index convenience. This does not change the optimum, since $l_{0:t}(x) = l_{1:t}(x)$.

$$\begin{aligned}\sum_{t=\tau}^{T} l_t(x_t) &= \sum_{t=\tau}^{T} l_{0:t}(x_t) - l_{0:t-1}(x_t) \\ &\leq \sum_{t=\tau}^{T} l_{0:t}(x_t) - l_{0:t-1}(x_{t-1}^*) \\ &= l_{0:T}(x_T^*) - l_{0:\tau-1}(x_{\tau-1}^*) \\ &\quad + \sum_{t=\tau}^{T} l_{0:t}(x_t) - l_{0:t}(x_t^*)\end{aligned}$$

■

To prove Theorem 1, we first note that by definition of $\hat{x}_N$, it satisfies $F(\hat{x}_N, \hat{x}_N) \leq \frac{1}{N}\sum_{n=1}^{N} f_n(x_n)$. To bound the average performance, we use Lemma 4 and write

$$\sum_{n=1}^{N} f_n(x_n) \leq f_{1:N}(x_{N+1}) + \sum_{n=1}^{N} f_{1:n}(x_n) - f_{1:n}(x_{n+1})$$

since $x_n = \arg\min_{x \in \mathcal{X}} f_{1:n-1}(x)$. Then because $f_{1:k}$ is $k\alpha$-strongly convex, by Lemma 3,

$$\sum_{n=1}^{N} f_n(x_n) \leq f_{1:N}(x_n^*) + \sum_{n=1}^{N} \frac{\|\nabla f_n(x_n)\|_*^2}{2\alpha n}.$$

Finally, dividing the upper-bound by n and using the facts that $\sum_{k=1}^{n} \frac{1}{k} \leq \ln(n) + 1$ and $\min a_i \leq \frac{1}{n}\sum a_i$ for any scalar sequence $\{a_n\}$, we have the desired result.

B Proof of Theorem 2

Now we give the proof of Theorem 2. Without using the first-order information of F in the first argument, we construct our analysis based on the convergence of an intermediate quantity, which indicates how fast the sequence concentrates toward its last element:

$$S_n := \frac{\sum_{k=1}^{n-1} \|x_n - x_k\|}{n-1} \tag{12}$$

which is defined $n \geq 2$ and $S_2 = \|x_2 - x_1\|$.

First, we use Assumption 5 to strengthen the bound $\|x_{n+1} - x_n\| = O(\frac{1}{n})$ used in Theorem 1 by techniques from online learning with prediction (Rakhlin and Sridharan, 2013).

Lemma 5. *Under Assumption 3, 5, running* AGGREVATE *gives, for $n \geq 2$, $\|x_{n+1} - x_n\| \leq \frac{\theta S_n}{n}$.*

Proof. First, because $f_{1:n}(x)$ is $n\alpha$-strongly convex,

$$\begin{aligned} \frac{n\alpha}{2}\|x_{n+1} - x_n\|^2 &\leq f_{1:n}(x_n) - f_{1:n}(x_{n+1}) \\ &\leq \langle \nabla f_{1:n}(x_n), x_n - x_{n+1}\rangle - \frac{\alpha n}{2}\|x_n - x_{n+1}\|^2. \end{aligned}$$

Let $\bar{f}_n = \frac{1}{n} f_{1:n}$. The above inequality implies

$$\begin{aligned} n\alpha\|x_{n+1} - x_n\|^2 &\leq \langle \nabla f_n(x_n), x_n - x_{n+1}\rangle \\ &\leq \langle \nabla f_n(x_n) - \nabla \bar{f}_{n-1}(x_n), x_n - x_{n+1}\rangle \\ &\leq \|\nabla f_n(x_n) - \nabla \bar{f}_{n-1}(x_n)\|\|x_n - x_{n+1}\| \\ &\leq \beta S_n \|x_n - x_{n+1}\| \end{aligned}$$

where the second inequality is due to $x_n = \arg\min_{x\in\mathcal{X}} f_{1:n-1}(x)$ and the last inequality is due to Assumption 5. Thus, $\|x_n - x_{n+1}\| \leq \frac{\beta S_n}{\alpha n}$. ■

Using the refined bound provided by Lemma 5, we can bound the progress of S_n.

Proposition 2. *Under the assumptions in Lemma 5, for $n \geq 2$, $S_n \leq e^{1-\theta} n^{\theta-1} S_2$ and $S_2 = \|x_2 - x_1\| \leq \frac{G_2}{\alpha}$.*

Proof. The bound on $S_2 = \|x_2 - x_1\|$ is due to that $x_2 = \arg\min_{x\in\mathcal{X}} f_1(x)$ and that f_1 is α-strongly convex and G_2-Lipschitz continuous.

To bound S_n, first we bound S_{n+1} in terms of S_n by

$$\begin{aligned} S_{n+1} &\leq \left(1 - \frac{1}{n}\right) S_n + \|x_{n+1} - x_n\| \\ &\leq \left(1 - \frac{1}{n} + \frac{\theta}{n}\right) S_n = \left(1 - \frac{1-\theta}{n}\right) S_n \end{aligned}$$

in which the first in equality is due to triangular inequality (i.e. $\|x_k - x_{n+1}\| \leq \|x_k - x_n\| + \|x_n - x_{n+1}\|$) and the second inequality is due to Lemma 5. Let $P_n = \ln S_n$. Then we can bound $P_n - P_2 \leq \sum_{k=2}^{n-1} \ln\left(1 - \frac{1-\theta}{k}\right) \leq \sum_{k=2}^{n-1} -\frac{1-\theta}{k} \leq -(1-\theta)(\ln n - 1)$, where we use the facts that $\ln(1+x) \leq x$, $\sum_{k=1}^{n} \frac{1}{k} \geq \ln(n+1)$. This implies $S_n = \exp(P_n) \leq e^{1-\theta} n^{\theta-1} S_2$. ■

More generally, define $S_{m:n} = \frac{\sum_{k=m}^{n-1} \|x_n - x_k\|}{n-m}$ (i.e. $S_n = S_{1:n}$). Using Proposition 2, we give a bound on $S_{m:n}$. We see that the convergence of $S_{m:n}$ depends mostly on n not m. (The proof is given in Appendix.)

Corollary 3. *Under the assumptions in Lemma 5, for $n > m$, $S_{m:n} \leq O(\frac{\theta}{(n-m)m^{2-\theta}} + \frac{1}{n^{1-\theta}})$.*

Now we are ready prove Theorem 2 by using the concentration of S_n in Proposition 2.

Proof of Theorem 2. First, we prove the bound on $F(x_N, x_N)$. Let $x_n^* := \arg\min_{x \in \mathcal{X}} f_n(x)$ and let $\bar{f}_n = \frac{1}{n} f_{1:n}$. Then by α-strongly convexity of f_n,

$$\begin{aligned}
& f_n(x_n) - \min_{x \in \mathcal{X}} f_n(x) \\
& \leq \langle \nabla f_n(x_n), x_n - x_n^* \rangle - \frac{\alpha}{2} \| x_n - x_n^* \|^2 \\
& \leq \langle \nabla f_n(x_n) - \bar{f}_{n-1}(x_n), x_n - x_n^* \rangle - \frac{\alpha}{2} \| x_n - x_n^* \|^2 \\
& \leq \| \nabla f_n(x_n) - \bar{f}_{n-1}(x_n) \|_* \| x_n - x_n^* \| - \frac{\alpha}{2} \| x_n - x_n^* \|^2 \\
& \leq \frac{\| \nabla f_n(x_n) - \bar{f}_{n-1}(x_n) \|_*^2}{2\alpha} \leq \frac{\beta^2}{2\alpha} S_n^2
\end{aligned}$$

where the second inequality uses the fact that $x_n = \arg\min_{x \in \mathcal{X}} \bar{f}_{n-1}(x)$, the second to the last inequality takes the maximum over $\|x_n - x_n^*\|$, and the last inequality uses Assumption 5. To bound $F(x_N, x_N)$, we use Proposition 2 and Assumption 6:

$$\begin{aligned}
f_n(x_n) & \leq \min_{x \in \mathcal{X}} f_n(x) + \frac{\beta^2}{2\alpha} S_n^2 \\
& \leq \tilde{\epsilon}_{\Pi, \pi^*} + \frac{\beta^2}{2\alpha} \left(e^{1-\theta} n^{\theta - 1} \frac{G_2}{\alpha} \right)^2
\end{aligned}$$

Rearranging the terms gives the bound in Theorem 2, and that $\|x_n - \bar{x}_n\| \leq S_n$ gives the second result.

Now we show the convergence of $\{x_n\}$ under the condition $\theta < 1$. It is sufficient to show that $\lim_{n \to \infty} \sum_{k=1}^n \|x_k - x_{k+1}\| < \infty$. To see this, we apply Lemma 5 and Proposition 2: for $\theta < 1$, $\sum_{k=1}^n \|x_k - x_{k+1}\| \leq \|x_1 - x_2\| + \sum_{k=2}^n \frac{\theta}{k} S_k \leq c_1 + c_2 \sum_{k=2}^n \frac{\theta}{k} \frac{S_2}{k^{1-\theta}} < \infty$, where $c_1, c_2 \in O(1)$. ■

C Proof of Lemma 2

Define $\delta_{\pi|t}$ such that $d_{\pi|t;q}(s) = (1 - q^t)\delta_{\pi|t}(s) + q^t d_{\pi^*}(s)$, and define $g_{z|t}(s) = \nabla_z E_{\pi_z}[Q_{\pi^*|t}](s)$, where π_z is a policy parametrized by arbitrary $z \in \mathcal{X}$. By assumption, $\|g_{z|t}\|_* < G_2$. Let π, π' be two policies parameterized by $x, y \in \mathcal{X}$, respectively. Then

$$\begin{aligned}
& \|\nabla_2 \hat{F}(x, z) - \nabla_2 \hat{F}(y, z)\|_* \\
& = \|\mathbb{E}_{d_{\tilde{\pi}}}[g_{z|t}] - \mathbb{E}_{d_{\tilde{\pi}'}}[g_{z|t}]\|_* \\
& = \| \frac{1}{T} \sum_{t=0}^{T-1} (1 - q^t) (\mathbb{E}_{\delta_{\pi|t;q}}[g_{z|t}] - \mathbb{E}_{\delta_{\pi'|t;q}}[g_{z|t}]) \|_* \\
& \leq (1 - q^T) \frac{1}{T} \sum_{t=0}^{T-1} \|\mathbb{E}_{\delta_{\pi|t;q}}[g_{z|t}] - \mathbb{E}_{\delta_{\pi'|t;q}}[g_{z|t}]\|_* \\
& \leq (1 - q^T) \frac{2G_2}{T} \sum_{t=0}^{T-1} \|\delta_{\pi|t;q} - \delta_{\pi'|t;q}\|_1 \\
& \leq (1 - q^T) \frac{2G_2}{T} \sum_{t=0}^{T-1} \|d_{\pi|t} - d_{\pi'|t}\|_1 \\
& \leq (1 - q^T) \beta \|x - y\|
\end{aligned}$$

in which the second to the last inequality is because the divergence between $d_{\pi|t}$ and $d_{\pi'|t}$ is the largest among all state distributions generated by the mixing policies.

D Proof of Corollary 1

The proof is similar to Lemma 2 and the proof of (Ross et al., 2011, Theorem 4.1).