A Notation

Table 2: Summary of the symbols used in the paper

Symbol	Definition
N	the total number of rounds in online learning
$J(\pi)$	the average accumulated cost, $\mathbb{E}_{d_\pi}\mathbb{E}_\pi[c_t]$ of RL in (1)
d_π	the generalized stationary state distribution
$D(q\|\|p)$	the difference between distributions p and q
π^*	the expert policy
Π	the hypothesis class of policies
π_n	the policy run in the environment at the nth online learning iteration
$\hat{\mathcal{F}}$	the hypothesis class of models (elements denoted as $\hat{F}$)
$\hat{F}_n$	the model used at the $n-1$ iteration to predict the future gradient of the nth iteration
ϵ_Π^w	the policy class complexity (Definition 4.1)
$\epsilon_{\hat{\mathcal{F}}}^w$	the model class complexity (Definition 4.1)
$F(\pi', \pi)$	the bivariate function $E_{d'_\pi}[D(\pi^*\|\|\pi)]$ in (5)
$f_n(\pi)$	$F(\pi_n, \pi)$ in (6)
$\tilde{f}_n(\pi)$	an unbiased estimate of $f_n(\pi)$
$h_n(\hat{F})$	an upper bound of $\|\nabla_2 F(\pi_n, \pi_n) - \nabla_2 \hat{F}(\pi_n, \pi_n)\|_*^2$
$\tilde{h}_n(\hat{F})$	an unbiased estimate of $h_n(\hat{F})$
μ_f	the modulus of strongly convexity of $\tilde{f}_n$ (Assumption 4.2)
G_f	an upper bound of $\|\nabla \tilde{f}_n\|_*$ (Assumption 4.2)
G	an upper bound of $\|\nabla f_n\|_*$ (Theorem 2.1)
μ_h	modulus of strongly convexity of $\tilde{h}_n$ (Assumption 4.2)
G_h	an upper bound of $\|\nabla \tilde{h}_n\|_*$ (Assumption 4.2)
L	the Lipschitz constant such that $\|\nabla_2 \hat{F}(\pi, \pi) - \nabla_2 \hat{F}(\pi', \pi')\|_* \le L\|\pi - \pi'\|$ (Assumption 4.1)
$\mathcal{R}(p)$	the expected weighted average regret, $\mathbb{E}\left[\frac{\text{regret}^w(\Pi)}{w_{1:N}}\right]$ in (10)
regret^w	the weighted regret, defined in Lemma 3.1
$\{w_n\}$	the sequence of weights used to define regret^w; we set $w_n = n^p$

B Imitation Learning Objective Function and Choice of Distance

Here we provide a short introduction to the objective function of IL in (2). The idea of IL is based on the Performance Difference Lemma, whose proof can be found, e.g. in [13].

Lemma B.1 (Performance Difference Lemma)**.** *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 with respect to running π'. Then it holds that*

$$J(\pi) = J(\pi') + \mathbb{E}_{d_\pi}\mathbb{E}_\pi[A_{\pi',t}]. \tag{B.1}$$

Using Lemma B.1, we can relate the performance of the learner's policy and the expert policy as

$$\begin{aligned} J(\pi) &= J(\pi^*) + \mathbb{E}_{d_\pi}\mathbb{E}_\pi[A_{\pi^*,t}] \\ &= J(\pi^*) + \mathbb{E}_{d_\pi}\left[(\mathbb{E}_\pi - \mathbb{E}_{\pi^*})[Q_{\pi^*,t}]\right] \end{aligned}$$

where the last equality uses the definition of $A_{\pi',t}$ and that $V_{\pi,t} = \mathbb{E}_\pi[Q_{\pi,t}]$. Therefore, if the inequality below holds

$$\mathbb{E}_{a\sim\pi_s}[Q_{\pi^*,t}(s,a)] - \mathbb{E}_{a^*\sim\pi_s^*}[Q_{\pi^*,t}(s,a^*)] \le C_{\pi^*} D(\pi_s^*||\pi_s), \qquad \forall t \in \mathbb{N}, s \in \mathbb{S}, \pi \in \Pi$$

then minimizing (2) would minimize the performance difference between the policies as in (3)

$$J(\pi) - J(\pi^*) \le C_{\pi^*}\mathbb{E}_{d_\pi}[D(\pi^*||\pi)].$$

Intuitively, we can set $D(\pi^*||\pi) = \mathbb{E}_\pi[A_{\pi^*,t}]$ and (3) becomes an equality with $C_{\pi^*} = 1$. This corresponds to the objective function used in AggreVaTe by Ross and Bagnell [3]. However, this choice requires $A_{\pi^*,t}$ to be given as a function or to be estimated online, which may be inconvenient or complicated in some settings.

Therefore, D is usually used to construct a strict upper bound in (3). The choice of D and C_{π^*} is usually derived from some statistical distances, and it depends on the topology of the action space $\mathbb{A}$ and the policy class Π. For discrete action spaces, D can be selected as a convex upper bound of the total variational distance between π and π^* and C_{π^*} is a bound on the range of $Q_{\pi^*,t}$ (e.g., a hinge loss used by [2]). For continuous action spaces, D can be selected as an upper bound of the Wasserstein distance between π and π^* and C_{π^*} is the Lipschitz constant of $Q_{\pi^*,t}$ with respect to action [12]. More generally, for stochastic policies, we can simply set D to Kullback-Leibler (KL) divergence (e.g. by [7]), because it upper bounds both total variational distance and Wasserstein distance. The direction of KL divergence, i.e. $D(\pi_s^*||\pi_s) = \mathrm{KL}[\pi_s||\pi_s^*]$ or $D(\pi_s^*||\pi_s) = \mathrm{KL}[\pi_s^*||\pi_s]$, can be chosen based on the characteristics of the expert policy. For example, if the log probability of the expert policy (e.g. a Gaussian policy) is available, $\mathrm{KL}[\pi_s||\pi_s^*]$ can be used. If the expert policy is only accessible through stochastic queries, then $\mathrm{KL}[\pi_s^*||\pi_s]$ is the only feasible option.

C Missing Proofs

C.1 Proof of Section 3.1

Lemma 3.1. *For arbitrary sequences $\{\pi_n \in \Pi\}_{n=1}^N$ and $\{w_n > 0\}_{n=1}^N$, it holds that*

$$\mathbb{E}\left[\sum_{n=1}^N \frac{w_n J(\pi_n)}{w_{1:N}}\right] \leq J(\pi^*) + C_{\pi^*}\left(\epsilon_\Pi^w + \mathbb{E}\left[\frac{\mathrm{regret}^w(\Pi)}{w_{1:N}}\right]\right)$$

where $\tilde{f}_n$ is an unbiased estimate of f_n, $\mathrm{regret}^w(\Pi) \coloneqq \max_{\pi\in\Pi}\sum_{n=1}^N w_n\tilde{f}_n(\pi_n) - w_n\tilde{f}_n(\pi)$, ϵ_Π^w is given in Definition 4.1, and the expectation is due to sampling $\tilde{f}_n$.

Proof of Lemma 3.1. By inequality in (3) and definition of f_n,

$$\mathbb{E}\left[\sum_{n=1}^N w_n(J(\pi_n) - J(\pi^*))\right] \leq C_{\pi^*}\mathbb{E}\left[\sum_{n=1}^N w_n f_n(\pi_n)\right] = C_{\pi^*}\mathbb{E}\left[\sum_{n=1}^N w_n\tilde{f}_n(\pi_n)\right],$$

where the last equality is due to π_n is non-anticipating. This implies that

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^N w_n J(\pi_n)\right] &\leq w_{1:N}J(\pi^*) + C_{\pi^*}\mathbb{E}\left[\sum_{n=1}^N w_n\tilde{f}_n(\pi_n)\right] \\ &= w_{1:N}J(\pi^*) + C_{\pi^*}\mathbb{E}\left[\min_{\pi\in\Pi}\sum_{n=1}^N w_n\tilde{f}_n(\pi) + \mathrm{regret}^w(\Pi)\right]\end{aligned}$$

The statement is obtained by dividing both sides by $w_{1:N}$ and by the definition of $\epsilon_{\hat{\mathcal{F}}}^w$. ■

C.2 Proof of Section 4.2

Theorem 4.1. *For* MoBIL-VI *with $p > 1$, $R(p) \leq C_p\left(\frac{pG_h^2}{2(p-1)\mu_h}\frac{1}{N^2} + \frac{\epsilon_{\hat{\mathcal{F}}}^w}{pN}\right)$, where $C_p = \frac{(p+1)^2e^{p/N}}{2\mu_f}$.*

Proof. We prove a more general version of Theorem 4.1 below. ■

Theorem C.1. *For* MoBIL-VI*,*

$$\mathcal{R}(p) \leq \begin{cases} \frac{G_h^2}{4\mu_f\mu_h}\frac{p(p+1)^2e^{\frac{p}{N}}}{p-1}\frac{1}{N^2} + \frac{1}{2\mu_f}\frac{(p+1)^2e^{\frac{p}{N}}}{p}\frac{1}{N}\epsilon_{\hat{\mathcal{F}}}^w, & \text{for } p > 1 \\ \frac{G_h^2}{\mu_f\mu_h}\frac{\ln(N+1)}{N^2} + \frac{2}{\mu_f}\frac{1}{N}\epsilon_{\hat{\mathcal{F}}}^w, & \text{for } p = 1 \\ \frac{G_h^2}{4\mu_f\mu_h}(p+1)^2\frac{O(1)}{N^{p+1}} + \frac{1}{2\mu_f}\frac{(p+1)^2e^{\frac{p}{N}}}{p}\frac{1}{N^2}\epsilon_{\hat{\mathcal{F}}}^w, & \text{for } 0 < p < 1 \\ \frac{G_h^2}{2\mu_f\mu_h}\frac{1}{N} + \frac{1}{2\mu_f}\frac{\ln N+1}{N}\epsilon_{\hat{\mathcal{F}}}^w, & \text{for } p = 0 \end{cases}$$

Proof. The solution π_{n+1} of the VI problem (8) satisfies the optimality condition of

$$\pi_{n+1} = \arg\min_{\pi\in\Pi} \sum_{m=1}^{n} w_m f_m(\pi_n) + w_{n+1}\hat{F}_{n+1}(\pi_{n+1}, \pi).$$

Therefore, we can derive the bound of $\mathcal{R}(p)$[12] as

$$\begin{aligned}
\mathcal{R}(p) &= \frac{\text{regret}^w(\Pi)}{w_{1:N}} \\
&\le \frac{p+1}{2\mu_f w_{1:N}} \sum_{n=1}^{N} n^{p-1} \|\nabla_2 F(\pi_n, \pi_n) - \nabla_2 \hat{F}_n(\pi_n, \pi_n)\|_*^2 && \text{(Lemma H.5)} \\
&\le \frac{p+1}{2\mu_f w_{1:N}} \sum_{n=1}^{N} n^{p-1} h_n(\pi_n) && \text{(Property of } h_n\text{)} \qquad \text{(C.1)}
\end{aligned}$$

Next, we treat $n^{p-1}h_n$ as the per-round cost for an online learning problem, and utilize Lemma H.6 to upper bound the accumulated cost. In particular, we set w_n in Lemma H.6 to n^{p-1} and l_n to h_n. Finally, $w_{1:N} = \sum_{n=1}^{N} n^p$ can be lower bounded using Lemma H.1. Hence, for $p > 1$, we have

$$\begin{aligned}
\mathcal{R}(p) &\le \frac{p+1}{2\mu_f} \frac{p+1}{N^{p+1}} \left(\frac{G_h^2}{2\mu_h} \frac{p}{p-1} (N+1)^{p-1} + \frac{1}{p}(N+1)^p \epsilon_{\hat{\mathcal{F}}}^w \right) \\
&= \frac{G_h^2}{4\mu_f\mu_h} \frac{p(p+1)^2}{p-1} \left(\frac{N+1}{N}\right)^{p-1} \frac{1}{N^2} + \frac{1}{2\mu_f} \frac{(p+1)^2}{p} \left(\frac{N+1}{N}\right)^p \frac{1}{N} \epsilon_{\hat{\mathcal{F}}}^w \\
&\le \frac{G_h^2}{4\mu_f\mu_h} \frac{p(p+1)^2 e^{\frac{p}{N}}}{p-1} \frac{1}{N^2} + \frac{1}{2\mu_f} \frac{(p+1)^2 e^{\frac{p}{N}}}{p} \frac{1}{N} \epsilon_{\hat{\mathcal{F}}}^w,
\end{aligned}$$

where in the last inequality we utilize the fact that $1 + x \le e^x, \forall x \in \mathbb{R}$. Cases other than $p > 1$ follow from straightforward algebraic simplification.

■

Proposition 4.1. *For* MoBIL-VI *with* $p = 0$, $\mathcal{R}(0) \le \frac{G_f^2}{2\mu_f\mu_h} \frac{1}{N} + \frac{\epsilon_{\hat{\mathcal{F}}}^w}{2\mu_f} \frac{\ln N + 1}{N}$.

Proof. Proved in Theorem C.1 by setting $p = 0$. ■

C.3 Proof of Section 4.3

Lemma 4.1 (Stronger FTL Lemma)**.** *Let* $x_n^\star \in \arg\min_{x\in\mathcal{X}} l_{1:n}(x)$. *For any sequence of decisions* $\{x_n\}$ *and losses* $\{l_n\}$, $\text{regret}(\mathcal{X}) = \sum_{n=1}^{N} l_{1:n}(x_n) - l_{1:n}(x_n^\star) - \Delta_n$, *where* $\Delta_{n+1} := l_{1:n}(x_{n+1}) - l_{1:n}(x_n^\star) \ge 0$.

Proof. The proof is based on observing $l_n = l_{1:n} - l_{1:n-1}$ and $l_{1:N}$ as a telescoping sum:

$$\begin{aligned}
\text{regret}(\mathcal{X}) &= \sum_{n=1}^{N} l_n(x_n) - l_{1:N}(x_N^\star) \\
&= \sum_{n=1}^{N} \left(l_{1:n}(x_n) - l_{1:n-1}(x_n)\right) - \sum_{n=1}^{N} \left(l_{1:n}(x_n^\star) - l_{1:n-1}(x_{n-1}^\star)\right) \\
&= \sum_{n=1}^{N} \left(l_{1:n}(x_n) - l_{1:n}(x_n^\star) - \Delta_n\right),
\end{aligned}$$

where for notation simplicity we define $l_{1:0} \equiv 0$. ■

Lemma 4.2. $\text{regret}_{\text{path}}^w(\Pi) \le \frac{p+1}{2\alpha\mu_f} \sum_{n=1}^{N} n^{p-1} \|g_n - \hat{g}_n\|_*^2 - \frac{\alpha\mu_f}{2(p+1)} \sum_{n=1}^{N} (n-1)^{p+1} \|\pi_n - \hat{\pi}_n\|^2$.

[12] The expectation of $\mathcal{R}(p)$ is not required here because MoBIL-VI assumes the problem is deterministic.

Proof. We utilize our new Lemma 4.1. First, we bound $\sum_{n=1}^{N} l_{1:n}(\pi_n) - l_{1:n}(\pi_n^\star)$, where $\pi_n^\star = \arg\min_{\pi\in\Pi} l_{1:n}(\pi)$. We achieve this by Lemma H.4. Let $l_n = w_n \bar{f}_n = w_n(\langle g_n, \pi\rangle + r_n(\pi))$. To use Lemma H.4, we note that because r_n is centered at π_n, π_{n+1} satisfies

$$\begin{aligned}\pi_{n+1} &= \arg\min_{\pi\in\Pi} \sum_{m=1}^{n} w_m \bar{f}(\pi) + w_{n+1}\langle \hat{g}_{n+1}, \pi\rangle \\ &= \arg\min_{\pi\in\Pi} \sum_{m=1}^{n} \underbrace{w_m \bar{f}(\pi)}_{l_n(\pi)} + \underbrace{w_{n+1}\langle \hat{g}_{n+1}, \pi\rangle + w_{n+1} r_{n+1}(\pi_{n+1})}_{v_{n+1}(\pi)}\end{aligned}$$

Because by definition l_n is $w_n\alpha\mu_f$-strongly convex, it follows from Lemma H.4 and Lemma H.1 that

$$\sum_{n=1}^{N} l_{1:n}(\pi_n) - l_{1:n}(\pi_n^\star) \le \frac{1}{\alpha\mu_f}\sum_{n=1}^{N}\frac{w_n^2}{w_{1:n}}\|\hat{g}_n - g_n\|_*^2 \le \frac{p+1}{2\alpha\mu_f}\sum_{n=1}^{N} n^{p-1}\|g_n - \hat{g}_n\|_*^2.$$

Next, we bound Δ_{n+1} as follows

$$\begin{aligned}\Delta_{n+1} &= l_{1:n}(\pi_{n+1}) - l_{1:n}(\pi_n^\star) \\ &\ge \langle \nabla l_{1:n}(\pi_n^\star), \pi_{n+1} - \pi_n^\star\rangle + \frac{\alpha\mu_f w_{1:n}}{2}\|\pi_{n+1} - \pi_n^\star\|^2 && \text{(Strong convexity)} \\ &\ge \frac{\alpha\mu_f w_{1:n}}{2}\|\pi_{n+1} - \pi_n^\star\|^2 && \text{(Optimality condition of } \pi_n^*) \\ &= \frac{\alpha\mu_f w_{1:n}}{2}\|\pi_{n+1} - \hat{\pi}_{n+1}\|^2 && \text{(Definition of } \hat{\pi}_{n+1}) \\ &\ge \frac{\alpha\mu_f n^{p+1}}{2(p+1)}\|\pi_{n+1} - \hat{\pi}_{n+1}\|^2. && \text{(Definition of } w_n \text{ and Lemma H.1)}\end{aligned}$$

Combining these results proves the bound. ■

Lemma 4.3. $\mathbb{E}[\|g_n - \hat{g}_n\|_*^2] \le 4(\sigma_g^2 + \sigma_{\hat{g}}^2 + L^2\mathbb{E}[\|\pi_n - \hat{\pi}_n\|^2] + \mathbb{E}[\tilde{h}_n(\hat{F}_n)])$.

Proof. By Lemma H.3, we have

$$\begin{aligned}\mathbb{E}\big[\|g_n - \hat{g}_n\|_*^2\big] \le 4\Big(&\mathbb{E}\big[\|g_n - \nabla_2 F(\pi_n,\pi_n)\|_*^2\big] + \mathbb{E}\big[\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\pi_n,\pi_n)\|_*^2\big] + \\ &\mathbb{E}\big[\|\nabla_2\hat{F}_n(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\hat{\pi}_n,\hat{\pi}_n)\|_*^2\big] + \mathbb{E}\big[\|\nabla_2\hat{F}_n(\hat{\pi}_n,\hat{\pi}_n) - \hat{g}_n\|_*^2\big]\Big).\end{aligned}$$

Because the random quantities are generated in order $\ldots, \pi_n, g_n, \hat{F}_{n+1}, \hat{\pi}_{n+1}, \hat{g}_{n+1}, \pi_{n+1}, g_{n+1}\ldots$, by the variance assumption (Assumption 4.3), the first and fourth terms can be bounded by

$$\begin{aligned}\mathbb{E}\big[\|g_n - \nabla_2 F(\pi_n,\pi_n)\|_*^2\big] &= \mathbb{E}_{\pi_n}\left[\mathbb{E}_{g_n}[\|g_n - \nabla_2 F(\pi_n,\pi_n)\|_*^2|\pi_n]\right] \le \sigma_g^2, \\ \mathbb{E}\big[\|\nabla_2\hat{F}_n(\hat{\pi}_n,\hat{\pi}_n) - \hat{g}_n\|_*^2\big] &= \mathbb{E}_{\hat{F}_n,\hat{\pi}_n}\left[\mathbb{E}_{\hat{g}_n}[\|\nabla_2\hat{F}_n(\hat{\pi}_n,\hat{\pi}_n) - \hat{g}_n\|_*^2|\hat{\pi}_n,\hat{F}_n]\right] \le \sigma_{\hat{g}}^2.\end{aligned}$$

And, for the second term, we have

$$\mathbb{E}\big[\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\pi_n,\pi_n)\|_*^2\big] \le \mathbb{E}\big[h_n(\hat{F}_n)\big] = \mathbb{E}\big[\tilde{h}_n(\hat{F}_n)\big]$$

Furthermore, due to the Lipschitz assumption of $\nabla_2\hat{F}_{n+1}$ (Assumption 4.1), the third term is bounded by

$$\mathbb{E}\big[\|\nabla_2\hat{F}_n(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\hat{\pi}_n,\hat{\pi}_n)\|_*^2\big] \le L^2\mathbb{E}\big[\|\pi_n - \hat{\pi}_n\|^2\big].$$

Combing the bounds above, we conclude the lemma.

■

Theorem 4.2. *For* MoBIL-Prox *with $p > 1$ and $\alpha_n = \alpha \in (0, 1]$, it satisfies*

$$\mathcal{R}(p) \le \frac{(p+1)^2 e^{\frac{p}{N}}}{\alpha\mu_f}\left(\frac{G_h^2}{\mu_h}\frac{p}{p-1}\frac{1}{N^2} + \frac{2}{p}\frac{\sigma_g^2 + \sigma_{\hat{g}}^2 + \epsilon_{\hat{F}}^w}{N}\right) + \frac{(p+1)\nu_p}{N^{p+1}},$$

where $\nu_p = O(1)$ and $n_{\mathrm{ceil}} = \lceil \frac{2e^{\frac{1}{2}}(p+1)LG_f}{\alpha\mu_f} \rceil$.

Proof. We prove a more general version of Theorem 4.1 below. ■

Theorem C.2. *For* MoBIL-Prox*,*

$$\mathcal{R}(p) \le \frac{4}{\alpha}\mathcal{R}_{\text{MoBIL-VI}}(p) + \epsilon_\Pi^w + \sigma(p)\left(\sigma_g^2 + \sigma_{\hat{g}}^2\right) + \frac{(p+1)\nu_p}{N^{p+1}},$$
$$\sigma(p) \le \begin{cases} \frac{2}{\alpha\mu_f}\frac{(p+1)^2 e^{\frac{p}{N}}}{p}\frac{1}{N}, & \text{if } p > 0 \\ \frac{2}{\alpha\mu_f}\frac{\ln N + 1}{N}, & \text{if } p = 0 \end{cases}$$
$$\nu(p) = 2e\left(\frac{(p+1)LG_f}{\alpha\mu_f}\right)^2 \sum_{n=2}^{n_{\text{ceil}}} n^{2p-2} - \frac{eG_f^2}{2}\sum_{n=2}^{n_{\text{ceil}}}(n-1)^{p+1}n^{p-1} = O(1), \quad n_{\text{ceil}} = \lceil \frac{2e^{\frac{1}{2}}(p+1)LG_f}{\alpha\mu_f} \rceil$$

where $\mathcal{R}_{\text{MoBIL-VI}}(p)$ is the upper bound of the average regret $\mathcal{R}(p)$ in Theorem C.1, and the expectation is due to sampling $\tilde{f}_n$ and $\tilde{h}_n$.

Proof. Recall $\mathcal{R}(p) = \mathbb{E}[\frac{\text{regret}^w(\Pi)}{w_{1:N}}]$, where

$$\text{regret}^w(\Pi) = \sum_{n=1}^N w_n \tilde{f}_n(\pi_n) - \min_{\pi\in\Pi}\sum_{n=1}^N w_n \tilde{f}_n(\pi).$$

Define $\bar{f}_n(\pi) := \langle g_n, \pi\rangle + r_n(\pi)$. Since $\tilde{f}_n$ is μ_f-strongly convex, r_n is $\alpha\mu_f$-strongly convex, and $r(\pi_n) = 0$, $\bar{f}_n$ satisfies

$$\tilde{f}_n(\pi_n) - \tilde{f}_n(\pi) \le \bar{f}_n(\pi_n) - \bar{f}_n(\pi), \quad \forall \pi \in \Pi.$$

which implies $\mathcal{R}(p) \le \mathbb{E}[\frac{\text{regret}^w_{\text{path}}(\Pi)}{w_{1:N}}]$, where

$$\text{regret}^w_{\text{path}}(\Pi) \coloneqq \sum_{n=1}^N w_n \bar{f}_n(\pi_n) - \min_{\pi\in\Pi}\sum_{n=1}^N w_n \bar{f}_n(\pi)$$

is regret of an online learning problem with per-round cost $w_n\bar{f}_n$.

Lemma 4.2 upper bounds $\text{regret}^w_{\text{path}}(\Pi)$ by using Stronger FTL lemma (Lemma 4.1). Since the second term in Lemma 4.2 is negative, which is in our favor, we just need to upper bound the expectation of the first item. Using triangular inequality, we proceed to bound $\mathbb{E}\big[\|g_n - \hat{g}_n\|_*^2\big]$, which measures how well we are able to predict the next per-round cost using the model.

By substituting the result of Lemma 4.3 into Lemma 4.2, we see

$$\mathbb{E}\left[\text{regret}^w_{\text{path}}(\Pi)\right] \le \mathbb{E}\Big[\sum_{n=1}^N \rho_n \|\pi_n - \hat{\pi}_n\|^2\Big] + \left(\frac{2(p+1)}{\alpha\mu_f}\sum_{n=1}^N n^{p-1}\right)\left(\sigma_g^2 + \sigma_{\hat{g}}^2\right) + \frac{2(p+1)}{\alpha\mu_f}\mathbb{E}\Big[\sum_{n=1}^N n^{p-1}\tilde{h}_n(\hat{F}_n)\Big] \tag{C.2}$$

where $\rho_n = \frac{2(p+1)L^2}{\alpha\mu_f}n^{p-1} - \frac{\alpha\mu_f}{2(p+1)}(n-1)^{p+1}$. When n is large enough, $\rho_n \le 0$, and hence the first term of (C.2) is $O(1)$. To be more precise, $\rho_n \le 0$ if

$$\frac{2(p+1)L^2}{\alpha\mu_f}n^{p-1} \le \frac{\alpha\mu_f}{2(p+1)}(n-1)^{p+1}$$
$$\iff (n-1)^2 \ge \left(\frac{2(p+1)LG_f}{\alpha\mu_f}\right)^2\left(\frac{n}{n-1}\right)^{p-1}$$
$$\impliedby (n-1)^2 \ge \left(\frac{2(p+1)LG_f}{\alpha\mu_f}\right)^2 e^{\frac{p-1}{n-1}}$$

$$\begin{aligned}
&\Longleftarrow (n-1)^2 \geq \left(\frac{2(p+1)LG_f}{\alpha\mu_f}\right)^2 e && \text{(Assume } n \geq p\text{)}\\
&\Longleftarrow n \geq \frac{2e^{\frac{1}{2}}(p+1)LG_f}{\alpha\mu_f} + 1
\end{aligned}$$

Therefore, we just need to bound the first $n_{\text{ceil}} = \lceil \frac{2e^{\frac{1}{2}}(p+1)LG_f}{\alpha\mu_f} \rceil$ terms of $\rho_n\|\pi_n - \hat{\pi}_n\|^2$. Here we use a basic fact of convex analysis in order to bound $\|\pi_n - \hat{\pi}_n\|^2$

Lemma C.1. *Let $\mathcal{X}$ be a compact and convex set and let f, g be convex functions. Suppose $f+g$ is μ-strongly convex. Let $x_1 \in \arg\min_{x\in\mathcal{X}} f(x)$ and $x_2 = \arg\min_{x\in\mathcal{X}} (f(x)+g(x))$. Then $\|x_1 - x_2\| \leq \frac{\|\nabla g(x_1)\|_*}{\mu}$.*

Proof of Lemma C.1. Let $h = f + g$. Because h is μ-strongly convex and $x_2 = \arg\min_{x\in\mathcal{X}} h(x)$

$$\begin{aligned}
\frac{\mu}{2}\|x_1 - x_2\|^2 \leq h(x_1) - h(x_2) &\leq \langle \nabla h(x_1), x_1 - x_2\rangle - \frac{\mu}{2}\|x_1 - x_2\|^2\\
&\leq \langle \nabla g(x_1), x_1 - x_2\rangle - \frac{\mu}{2}\|x_1 - x_2\|^2
\end{aligned}$$

This implies $\mu\|x_1 - x_2\|^2 \leq \langle \nabla g(x_1), x_1 - x_2\rangle \leq \|\nabla g(x_1)\|_*\|x_1 - x_2\|$. Dividing both sides by $\|x_1 - x_2\|$ concludes the lemma. ■

Utilizing Lemma C.1 and the definitions of π_n and $\hat{\pi}_n$, we have, for $n \geq 2$,

$$\begin{aligned}
\|\pi_n - \hat{\pi}_n\|^2 &\leq \frac{1}{\alpha\mu_f w_{1:n-1}}\|w_n\hat{g}_n\|_*^2\\
&\leq \frac{(p+1)G_f^2}{\alpha\mu_f}\frac{n^{2p}}{(n-1)^{p+1}} && \text{(Bounded } \hat{g}_n \text{ and Lemma H.1)}\\
&\leq \frac{(p+1)e^{\frac{p+1}{n-1}}G_f^2}{\alpha\mu_f}n^{p-1} && (1 + x \leq e^x)\\
&\leq \frac{e(p+1)G_f^2}{\alpha\mu_f}n^{p-1} && \text{(Assume } n \geq p+2\text{).}
\end{aligned}$$

and therefore, after assuming initialization $\pi_1 = \hat{\pi}_1$, we have the bound

$$\sum_{n=2}^{n_{\text{ceil}}} \rho_n\|\pi_n - \hat{\pi}_n\|^2 \leq 2e\left(\frac{(p+1)LG_f}{\alpha\mu_f}\right)^2 \sum_{n=2}^{n_{\text{ceil}}} n^{2p-2} - \frac{eG_f^2}{2}\sum_{n=2}^{n_{\text{ceil}}}(n-1)^{p+1}n^{p-1} \tag{C.3}$$

of which more delicate upper bound can be derived from Lemma H.1. For the third term of (C.2), we can tie it back to the bound of $\mathcal{R}(p)$ of MoBIL-VI, which we denote $\mathcal{R}_{\text{MoBIL-VI}}(p)$. More concretely, recall that for MoBIL-VI in (C.1), we have

$$\mathcal{R}(p) \leq \frac{p+1}{2\mu_f w_{1:N}}\sum_{n=1}^{N} n^{p-1}h_n(\pi_n),$$

and we derived the upper bound ($\mathcal{R}_{\text{MoBIL-VI}}(p)$) for the RHS term. By observing that the third term of (C.2) after averaging is

$$\begin{aligned}
\frac{2(p+1)}{\alpha\mu_f w_{1:N}}\mathbb{E}\Big[\sum_{n=1}^{N} n^{p-1}\tilde{h}_n(\hat{F}_n)\Big] &= \mathbb{E}\Big[\frac{4}{\alpha}\left(\frac{p+1}{2\mu_f w_{1:N}}\sum_{n=1}^{N} n^{p-1}\tilde{h}_n(\hat{F}_n)\right)\Big]\\
&\leq \frac{4}{\alpha}\mathbb{E}\Big[\mathcal{R}_{\text{MoBIL-VI}}(p)\Big]\\
&= \frac{4}{\alpha}\mathcal{R}_{\text{MoBIL-VI}}(p).
\end{aligned} \tag{C.4}$$

Dividing (C.2) by $w_{1:N}$, and plugging in (C.3), (C.4), we see

$$\begin{aligned}\mathcal{R}(p) &\leq \mathbb{E}[\text{regret}^w_{\text{path}}(\Pi)/w_{1:N}] \\ &\leq \frac{4}{\alpha}\mathcal{R}_{\text{MoBIL-VI}}(p) + \frac{1}{w_{1:N}}\left(\nu_p + \left(\frac{2(p+1)}{\alpha\mu_f}\sum_{n=1}^{N} n^{p-1}\right)\left(\sigma_g^2 + \sigma_{\hat{g}}^2\right)\right)\end{aligned}$$

where $\nu_p = 2e\left(\frac{(p+1)LG_f}{\alpha\mu_f}\right)^2 \sum_{n=2}^{n_{\text{ceil}}} n^{2p-2} - \frac{eG_f^2}{2}\sum_{n=2}^{n_{\text{ceil}}}(n-1)^{p+1}n^{p-1}$, $n_{\text{ceil}} = \lceil \frac{2e^{\frac{1}{2}}(p+1)LG_f}{\alpha\mu_f} \rceil$.

Finally, we consider the case $p > 1$ as stated in Theorem 4.2

$$\begin{aligned}\mathcal{R}(p) &\leq \frac{4}{\alpha}\left(\frac{G_h^2}{4\mu_f\mu_h}\frac{p(p+1)^2e^{\frac{p}{N}}}{p-1}\frac{1}{N^2} + \frac{1}{2\mu_f}\frac{(p+1)^2e^{\frac{p}{N}}}{p}\frac{1}{N}\epsilon^w_{\hat{\mathcal{F}}}\right) + \frac{p+1}{N^{p+1}}\left(\nu_p + \left(\frac{2(p+1)}{\alpha\mu_f}\frac{n^p}{p}\right)\left(\sigma_g^2 + \sigma_{\hat{g}}^2\right)\right) \\ &\leq \frac{(p+1)^2e^{\frac{p}{N}}}{\alpha\mu_f}\left(\frac{G_h^2}{\mu_h}\frac{p}{p-1}\frac{1}{N^2} + \frac{2}{p}\frac{\sigma_g^2 + \sigma_{\hat{g}}^2 + \epsilon^w_{\hat{\mathcal{F}}}}{N}\right) + \frac{(p+1)\nu_p}{N^{p+1}},\end{aligned}$$

where $\nu_p = 2e\left(\frac{(p+1)LG_f}{\alpha\mu_f}\right)^2\left(\frac{(n_{\text{ceil}}+1)^{2p-1}}{2p-1} - 1\right) - \frac{eG_f^2}{2}\frac{(n_{\text{ceil}}-1)^{2p+1}}{2p+1}$, $n_{\text{ceil}} = \lceil \frac{2e^{\frac{1}{2}}(p+1)LG_f}{\alpha\mu_f} \rceil$. ■

D Model Learning through Learning Dynamics Models

So far we have stated model learning rather abstractly, which only requires $h_n(\hat{F})$ to be an upper bound of $\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2\hat{F}(\pi_n,\pi_n)\|_*^2$. Now we give a particular example of h_n and $\tilde{h}_n$ when the predictive model is constructed as a simulator with online learned dynamics models. Specifically, we consider learning a transition model $M \in \mathcal{M}$ online that induces a bivariate function $\hat{F}$, where $\mathcal{M}$ is the class of transition models. Let D_{KL} denote the KL divergence and let $d^M_{\pi_n}$ be the generalized stationary distribution (cf. (1)) generated by running policy π_n under transition model M. We define, for $M_n \in \mathcal{M}$, $\hat{F}_n(\pi',\pi) := \mathbb{E}_{d^{M_n}_{\pi'}}[D(\pi^*||\pi)]$. We show the error of $\hat{F}_n$ can be bounded by the KL-divergence error of M_n.

Lemma D.1. *Assume $\nabla D(\pi^*||\cdot)$ is L_D-Lipschitz continuous with respect to $\|\cdot\|_*$. It holds that $\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\pi_n,\pi_n)\|_*^2 \leq 2^{-1}(L_D\text{Diam}(\mathbb{S}))^2 D_{KL}(d_{\pi_n}||d^{M_n}_{\pi_n})$.*

Directly minimizing the marginal KL-divergence $D_{KL}(d_{\pi_n}, d^{M_n}_{\pi_n})$ is a nonconvex problem and requires backpropagation through time. To make the problem simpler, we further upper bound it in terms of the KL divergence between the true and the modeled *transition probabilities*.

To make the problem concrete, here we consider T-horizon RL problems.

Proposition D.1. *For a T-horizon problem with dynamics P, let M_n be the modeled dynamics. Then $\exists C > 0$ s.t $\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\pi_n,\pi_n)\|_*^2 \leq \frac{C}{T}\sum_{t=0}^{T-1}(T-t)\mathbb{E}_{d_{\pi_n,t}}\mathbb{E}_\pi\left[D_{KL}(P||M_n)\right]$.*

Therefore, we can simply takes h_n as the upper bound in Proposition D.1, and $\tilde{h}$ as its empirical approximation by sampling state-action transition triples through running policy π_n in the real environment. This construction agrees with the causal relationship assumed in the Section 3.2.1.

D.1 Proofs

Lemma D.1. *Assume $\nabla D(\pi^*||\cdot)$ is L_D-Lipschitz continuous with respect to $\|\cdot\|_*$. It holds that $\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2\hat{F}_n(\pi_n,\pi_n)\|_*^2 \leq 2^{-1}(L_D\text{Diam}(\mathbb{S}))^2 D_{KL}(d_{\pi_n}||d^{M_n}_{\pi_n})$.*

Proof. First, we use the definition of dual norm

$$\|\nabla_2\hat{F}(\pi_n,\pi_n) - \nabla_2 F(\pi_n,\pi_n)\|_* = \max_{\delta:\|\delta\|\leq 1}\left(\mathbb{E}_{d_{\pi_n}} - \mathbb{E}_{d^{M_n}_{\pi_n}}\right)\left[\langle\delta, \nabla D(\pi^*||\pi_n)\rangle\right] \tag{D.1}$$

and then we show that $\langle\delta, \nabla D(\pi^*||\pi_n)\rangle$ is L_D-Lipschitz continuous: for $\pi,\pi' \in \Pi$,

$$\langle\delta, \nabla D(\pi^*||\pi) - \nabla D(\pi^*||\pi')\rangle \leq \|\delta\|\|\nabla D(\pi^*||\pi) - \nabla D(\pi^*||\pi')\|_* \leq L_D\|\pi - \pi'\|$$

Note in the above equations ∇ is with respect to $D(\pi^*||\cdot)$.

Next we bound the right hand side of (D.1) using Wasserstein distance D_W, which is defined as follows [20]: for two probability distributions p and q defined on a metric space $D_W(p,q) \coloneqq \sup_{f:\text{Lip}(f(\cdot))\leq 1} \mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{x\sim q}[f(x)]$.

Using the property that $\langle \delta, \nabla D(\pi^*||\pi_n)\rangle$ is L_D-Lipschitz continuous, we can derive

$$\|\nabla_2 \hat{F}(\pi_n,\pi_n) - \nabla_2 F(\pi_n,\pi_n)\|_* \leq L_D D_W(d_{\pi_n}, \hat{d}_{\pi_n}) \leq \frac{L_D \text{Diam}(\mathbb{S})}{\sqrt{2}} \sqrt{D_{KL}(d_{\pi_n}||\hat{d}^n_{\pi_n})}$$

in which the last inequality is due to the relationship between D_{KL} and D_W [20]. ■

Proposition D.1. *For a T-horizon problem with dynamics P, let M_n be the modeled dynamics. Then $\exists C > 0$ s.t $\|\nabla_2 F(\pi_n,\pi_n) - \nabla_2 \hat{F}_n(\pi_n,\pi_n)\|_*^2 \leq \frac{C}{T}\sum_{t=0}^{T-1}(T-t)\mathbb{E}_{d_{\pi_n,t}}\mathbb{E}_\pi\left[D_{KL}(P||M_n)\right]$.*

Proof. Let $\rho_{\pi,t}$ be the state-action trajectory up to time t generated by running policy π, and let $\hat{\rho}_{\pi,t}$ be that of the dynamics model. To prove the result, we use a simple fact:

Lemma D.2. *Let p and q be two distributions.*

$$KL[p(x,y)||q(x,y)] = KL[p(x)||q(x)] + \mathbb{E}_{p(x)} KL[p(y|x)||q(y|x)]$$

Then the rest follows from Lemma D.1 and the following inequality.

$$\begin{aligned} D_{KL}(d_{\pi_n}||\hat{d}_{\pi_n}) &\leq \frac{1}{T}\sum_{t=0}^{T-1} D_{KL}(\rho_{\pi_n,t}||\hat{\rho}_{\pi_n,t}) \\ &= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{\rho_{\pi_n,t}}\left[\sum_{\tau=0}^{t-1} \ln \frac{p_M(s_{\tau+1}|s_\tau,a_\tau)}{p_{\hat{M}}(s_{\tau+1}|s_\tau,a_\tau)}\right] \\ &= \frac{1}{T}\sum_{t=0}^{T-1}(T-t)\mathbb{E}_{d_{\pi,t}}\mathbb{E}_\pi\left[D_{KL}(p_M||p_{\hat{M}})\right] \end{aligned}$$

■

E Relaxation of Strong Convexity Assumption

The strong convexity assumption (Assumption 4.2) can be relaxed to just convexity. We focus on studying the effect of $\tilde{f}_n$ and/or $\tilde{h}_n$ being just convex on $\mathcal{R}(p)$ in Theorem 2.1 and Theorem 4.2 in big-O notation. Suggested by Lemma 3.2, when strong convexity is not assumed, additional regularization has to be added in order to keep the stabilization terms $l_{1:n}(x_n) - l_{1:n}(x_n^\star)$ small.

Lemma E.1 (FTRL with prediction)**.** *Let l_n be convex with bounded gradient and let $\mathcal{X}$ be a compact set. In round n, let regularization r_n be μ_n-strongly convex for some $\mu_n \geq 0$ such that $r_n(x_n) = 0$ and $x_n \in \arg\min_{\mathcal{X}} r_n(x)$, and let v_{n+1} be a (non)convex function such that $\sum_{m=1}^n w_m(l_n + r_n) + w_{n+1}v_{n+1}$ is convex. Suppose that learner plays Follow-The-Regularized-Leader (FTRL) with prediction, i.e. $x_{n+1} = \arg\min_{x\in\mathcal{X}} \sum_{m=1}^n (w_m(l_n+r_n) + w_{n+1}v_{n+1})(x)$, and suppose that $\sum_{m=1}^n w_m\mu_m = \Omega(n^k) > 0$ and $\sum_{m=1}^n w_m r_n(x) \leq O(n^k)$ for all $x\in\mathcal{X}$ and some $k \geq 0$. Then, for $w_n = n^p$,*

$$\text{regret}^w(\mathcal{X}) = O(N^k) + \sum_{n=1}^N O\left(n^{2p-k}\right)\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2$$

Proof. The regret of the online learning problem with *convex* per-round cost $w_n l_n$ can be bounded by the regret of the online learning problem with *strongly convex* per-round cost $w_n(l_n + r_n)$ as follows. Let $x_n^\star \in \arg\min_{x\in\mathcal{X}} \sum_{n=1}^N w_n l_n(x)$.

$$\begin{aligned} \text{regret}^w(\mathcal{X}) &= \sum_{n=1}^N w_n l_n(x_n) - \min_{x\in\mathcal{X}} \sum_{n=1}^N w_n l_n(x) \\ &= \sum_{n=1}^N w_n(l_n(x_n) + r_n(x_n)) - \sum_{n=1}^N w_n(l_n(x_n^\star) + r_n(x_n^\star)) + \sum_{n=1}^N w_n r_n(x_n^\star) \end{aligned}$$

$$\leq \left(\sum_{n=1}^{N} w_n \left(l_n(x_n) + r_n(x_n) \right) - \min_{x \in \mathcal{X}} \sum_{n=1}^{N} w_n \left(l_n(x) + r_n(x) \right) \right) + O(N^k).$$

Since the first term is the regret of the online learning problem with *strongly convex* per-round cost $w_n (l_n + r_n)$, and $x_{n+1} = \arg\min_{\mathcal{X}} \left(\sum_{m=1}^{n} w_m (l_n + r_n) + w_{n+1} v_{n+1} \right)$, we can bound the first term via Lemma H.5 by setting $w_n = n^p$ and $\sum_{m=1}^{n} w_m \mu_m = O(n^k)$. ■

The lemma below is a corollary of Lemma E.1.

Lemma E.2 (FTRL)**.** *Under the same condition in Lemma E.1, suppose that learner plays FTRL, i.e.* $x_{n+1} = \arg\min_{\mathcal{X}} \sum_{m=1}^{n} w_m (l_n + r_n)$*. Then, for* $w_n = n^p$ *with* $p > -\frac{1}{2}$*, choose* $\{r_n\}$ *such that* $\sum_{m=1}^{n} w_m \mu_m = \Omega(n^{p+1/2}) > 0$ *and it achieves* $\text{regret}^w(\mathcal{X}) = O(N^{p+\frac{1}{2}})$ *and* $\frac{\text{regret}^w(\mathcal{X})}{w_{1:N}} = O(N^{-1/2})$.

Proof. Let $\sum_{m=1}^{n} w_m \mu_m = \Theta(n^k) > 0$ for some $k \geq 0$. First, if $2p - k > -1$, then we have

$$\begin{aligned}
\text{regret}(\mathcal{X}) &\leq O(N^k) + \sum_{n=1}^{N} O\left(n^{2p-k}\right) \|\nabla l_n(x_n)\|_*^2 && \text{(Lemma E.1)} \\
&\leq O(N^k) + \sum_{n=1}^{N} O\left(n^{2p-k}\right) && (l_n \text{ has bounded gradient}) \\
&\leq O(N^k) + O\left(N^{2p-k+1}\right) && \text{(Lemma H.1)}
\end{aligned}$$

In order to have the best rate, we balance the two terms $O(N^k)$ and $O\left(N^{2p-k+1}\right)$

$$k = 2p - k + 1 \Longrightarrow k = p + \frac{1}{2},$$

That is, $p > -\frac{1}{2}$, because $2p - (p + \frac{1}{2}) > -1$. This setting achieves regret in $O(N^{p+\frac{1}{2}})$. Because $w_{1:N} = O(N^{p+1})$, the average regret is in $O(N^{-\frac{1}{2}})$. ■

With these lemmas, we are ready to derive the upper bounds of $\mathcal{R}(p)$ when either $\tilde{f}_n$ or $\tilde{h}_n$ is just convex, with some minor modification of Algorithm 1. For example, when $\tilde{f}_n$ is only convex, r_n will not be $\alpha\mu_f$ strongly; instead we will concern the strongly convexity of $\sum_{m=1}^{n} w_n r_n$. Similarly, if $\tilde{h}_n$ is only convex, the model cannot be updated by FTL as in line 5 of Algorithm 1; instead it has to be updated by FTRL.

In the following, we will derive the rate for MoBIL-VI (i.e. $\tilde{f}_n = f_n$ and $\tilde{h} = h$) and assume $\epsilon_{\tilde{\mathcal{F}}}^w = 0$ for simplicity. The same rate applies to the MoBIL-Prox when there is no noise. To see this, for example, if $\tilde{f}_n$ is only convex, we can treat r_n as an additional regularization and we can see

$$\mathcal{R}(p) = \mathbb{E}\Big[\frac{\text{regret}^w(\Pi)}{w_{1:N}}\Big] \leq \frac{1}{w_{1:N}} \mathbb{E}\Big[\underbrace{\sum_{n=1}^{N} w_n \bar{f}_n(\pi_n) - \min_{\pi \in \Pi} \sum_{n=1}^{N} w_n \bar{f}_n(\pi)}_{\text{regret}^w_{\text{path}}(\Pi)} + \sum_{n=1}^{N} w_n r_n(\pi_N^\star) \Big]$$

where $\pi_N^\star = \arg\min_{\pi \in \Pi} \sum_{n=1}^{N} \tilde{f}_n(\pi)$. As in the proof of Theorem 4.2, $\text{regret}^w_{\text{path}}$ is decomposed into several terms: the $\tilde{h}_n$ part in conjunction with $\sum_{n=1}^{N} w_n r_n(\pi_N^\star)$ constitute the same $\mathcal{R}(p)$ part for MoBIL-VI, while other terms in $\text{regret}^w_{\text{path}}$ are kept the same.

Strongly convex $\tilde{f}_n$ and convex $\tilde{h}_n$ Here we assume $p > \frac{1}{2}$. Under this condition, we have

$$\begin{aligned}
\text{regret}^w(\Pi) &= \sum_{n=1}^{N} O(n^{p-1}) \tilde{h}_n(\hat{F}_n) && \text{(Lemma H.5)} \\
&= O\left(N^{p-\frac{1}{2}}\right) && \text{(Lemma E.2)}
\end{aligned}$$

Because $w_{1:N} = \Omega(N^{p+1})$, the average regret $\mathcal{R}(p) = O(N^{-3/2})$.

Convex $\tilde{f}_n$ and strongly convex $\tilde{h}_n$ Here we assume $p > 0$. Suppose $r_{1:n}$ is $\Theta(n^k)$-strongly convex and $2p - k > 0$. Under this condition, we have

$$\begin{aligned}\text{regret}^w(\Pi) &= O(N^k) + \sum_{n=1}^{N} O\left(n^{2p-k}\right) \tilde{h}_n(\hat{F}_{n+1}) && \text{(Lemma E.1)}\\ &= O\left(N^k\right) + O\left(N^{2p-k}\right). && \text{(Lemma H.6)}\end{aligned}$$

We balance the two terms and arrive at

$$k = 2p - k \Longrightarrow k = p,$$

which satisfies the condition $2p - k > 0$, if $p > 0$. Because $w_{1:N} = \Omega(N^{p+1})$, the average regret $\mathcal{R}(p) = O(N^{-1})$.

Convex $\tilde{f}_n$ and convex $\tilde{h}_n$ Here we assume $p \geq 0$. Suppose $r_{1:n}$ is $\Theta(n^k)$-strongly convex and $2p - k > -\frac{1}{2}$. Under this condition, we have

$$\begin{aligned}\text{regret}^w(\Pi) &= O(N^k) + \sum_{n=1}^{N} O\left(n^{2p-k}\right) \tilde{h}_n(\hat{F}_{n+1}) && \text{(Lemma E.1)}\\ &= O\left(N^k\right) + O\left(N^{2p-k+\frac{1}{2}}\right) && \text{(Lemma E.1)}\end{aligned}$$

We balance the two terms and see

$$k = 2p - k + \frac{1}{2} \Longrightarrow k = p + \frac{1}{4},$$

which satisfies the condition $2p-k > -\frac{1}{2}$, if $p \geq 0$. Because $w_{1:N} = \Omega(N^{p+1})$, the average regret $\mathcal{R}(p) = O(N^{-3/4})$.

Convex f_n without model Setting $p = 0$ in Lemma E.2, we have $\text{regret}(\Pi) = O(N^{\frac{1}{2}})$.

Therefore, the average regret becomes $O(N^{-\frac{1}{2}})$.

Stochastic problems The above rates assume that there is no noise in the gradient and the model is realizable. If the general case, it should be selected $k = p + 1$ for strongly convex $\tilde{f}_n$ and $k = p + \frac{1}{2}$ for convex $\tilde{f}_n$. The convergence rate will become $O(\frac{\epsilon_{\mathcal{F}} + \sigma_g^2 + \sigma_{\hat{g}^2}}{N})$ and $O(\frac{\epsilon_{\mathcal{F}} + \sigma_g^2 + \sigma_{\hat{g}^2}}{\sqrt{N}})$, respectively.

F Connection with Stochastic Mirror-Prox

In this section, we discuss how MoBIL-Prox generalizes stochastic Mirror-Prox by Juditsky et al. [11], Nemirovski [18] and how the new Stronger FTL Lemma 4.1 provides more constructive and flexible directions to design new algorithms.

F.1 Variational Inequality Problems

Mirror-Prox [18] was first proposed to solve VI problems with monotone operators, which is a unified framework of "convex-like" problems, including convex optimization, convex-concave saddle-point problems, convex multi-player games, and equilibrium problems, etc (see [14] for a tutorial). Here we give the definition of VI problems and review some of its basic properties.

Definition F.1. Let $\mathcal{X}$ be a convex subset in an Euclidean space $\mathcal{E}$ and let $F : \mathcal{X} \to \mathcal{E}$ be an operator, the *VI problem*, denoted as $\text{VI}(\mathcal{X}, F)$, is to find a vector $x^* \in \mathcal{X}$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0, \qquad \forall x \in \mathcal{X}.$$

The set of solutions to this problem is denoted as $\text{SOL}(\mathcal{X}, F)$

It can be shown that, when $\mathcal{X}$ is also compact, then $\mathrm{VI}(\mathcal{X}, F)$ admits at least one solution [14]. For example, if $F(x) = \nabla f(x)$ for some function f, then solving $\mathrm{VI}(\mathcal{X}, F)$ is equivalent to finding stationary points.

VI problems are, in general, more difficult than optimization. To make the problem more structured, we will consider the problems equipped with some *general convex* structure, which we define below. When $F(x) = \nabla f(x)$ for some convex function f, the below definitions agree with their convex counterparts.

Definition F.2. An operator $F : \mathcal{X} \to \mathcal{E}$ is called

1. *pseudo-monotone* on $\mathcal{X}$ if for all $x, y \in \mathcal{X}$,
$$\langle F(y), x - y\rangle \geq 0 \implies \langle F(x), x - y\rangle \geq 0$$
2. *monotone* on $\mathcal{X}$ if for all $x, y \in \mathcal{X}$,
$$\langle F(x) - F(y), x - y\rangle \geq 0$$
3. *strictly monotone* on $\mathcal{X}$ if for all $x, y \in \mathcal{X}$,
$$\langle F(x) - F(y), x - y\rangle > 0$$
4. *μ-strongly monotone* on $\mathcal{X}$ if for all $x, y \in \mathcal{X}$,
$$\langle F(x) - F(y), x - y\rangle \geq \mu\|x - y\|^2$$

A VI problem is a special case of general equilibrium problems [21]. Therefore, for a VI problem, we can also define its dual VI problem.

Definition F.3. Given a VI problem $\mathrm{VI}(\mathcal{X}, F)$, the *dual VI problem*, denoted as $\mathrm{DVI}(\mathcal{X}, F)$, is to find a vector $x_D^* \in \mathcal{X}$ such that
$$\langle F(x), x - x_D^*\rangle \geq 0, \qquad \forall x \in \mathcal{X}.$$
The set of solutions to this problem is denoted as $\mathrm{DSOL}(\mathcal{X}, F)$.

The solution sets of the primal and the dual VI problems are connected as given in next proposition, whose proof e.g. can be found in [22].

Proposition F.1.

1. *If F is pseudo-monotone, then* $\mathrm{SOL}(\mathcal{X}, F) \subseteq \mathrm{DSOL}(\mathcal{X}, F)$.
2. *If F is continuous, then* $\mathrm{DSOL}(\mathcal{X}, F) \subseteq \mathrm{SOL}(\mathcal{X}, F)$.

However, unlike primal VI problems, a dual VI problem does not always have a solution even if $\mathcal{X}$ is compact. To guarantee the existence of solution to $\mathrm{DSOL}(\mathcal{X}, F)$ it needs stronger structure, such as pseudo-monotonicity as shown in Proposition F.1. Like solving primal VI problems is related to finding *local* stationary points in optimization, solving dual VI problems is related to finding *global* optima when $F(x) = \nabla f(x)$ for some function f [23].

F.2 Stochastic Mirror-Prox

Stochastic Mirror-Prox solves a monotone VI problem by indirectly finding a solution to its dual VI problem using stochastic first-order oracles. This is feasible because of Proposition F.1. The way it works is as follows: given an initial condition $x_1 \in \mathcal{X}$, it initializes $\hat{x}_1 = x_1$; at iteration n, it receives unbiased estimates g_n and $\hat{g}_n$ satisfying $\mathbb{E}[g_n] = F(x_n)$ and $\mathbb{E}[\hat{g}_n] = F(\hat{x}_n)$ and then performs updates
$$\begin{aligned} x_{n+1} &= \mathrm{Prox}_{\hat{x}_n}(\gamma_n \hat{g}_n) \\ \hat{x}_{n+1} &= \mathrm{Prox}_{\hat{x}_n}(\gamma_n g_{n+1}) \end{aligned} \tag{F.1}$$

where $\gamma_n > 0$ is the step size, and the proximal operator Prox is defined as

$$\text{Prox}_y(g) = \arg\min_{x \in \mathcal{X}} \langle g, x \rangle + B_\omega(x||y)$$

and $B_\omega(x||y) = \omega(x) - \omega(y) - \langle \nabla\omega(y), x - y \rangle$ is the Bregman divergence with respect to an α-strongly convex function ω. At the end, stochastic Mirror-Prox outputs

$$\bar{x}_N = \frac{\sum_{n=1}^N \gamma_n x_n}{\gamma_{1:n}}$$

as the final decision.

For stochastic Mirror-Prox, the accuracy of an candidate solution x is based on the error

$$\text{ERR}(x) \coloneqq \max_{y \in \mathcal{X}} \langle F(y), x - y \rangle .$$

This choice of error follows from the optimality criterion of the dual VI problem in Definition F.3. That is, $\text{ERR}(x) \leq 0$ if and only if $x \in \text{DSOL}(\mathcal{X}, F)$. From Proposition F.1, we know that if the problem is pseudo-monotone, a dual solution is also a primal solution. Furthermore, we can show an approximate dual solution is also an approximate primal solution.

Let $\Omega^2 = \max_{x,y \in \mathcal{X}} B_\omega(x||y)$. Now we recap the main theorem of [11].[13]

Theorem F.1. [11] *Let F be monotone. Assume F is L-Lipschitz continuous, i.e.*

$$\|F(x) - F(y)\|_* \leq L\|x - y\| \qquad \forall x, y \in \mathcal{X}$$

and for all n, the sampled vectors are unbiased and have bounded variance, i.e.

$$\begin{aligned} \mathbb{E}[g_n] = F(x_n), \qquad & \mathbb{E}[\hat{g}_n] = F(\hat{x}_n) \\ \mathbb{E}[\|g_n - F(x_n)\|_*^2] \leq \sigma^2, \qquad & \mathbb{E}[\|\hat{g}_n - F(\hat{x}_n)\|_*^2] \leq \sigma^2 \end{aligned}$$

Then for $\gamma_n = \gamma$ with $0 < \gamma_n \leq \frac{\alpha}{\sqrt{3}L}$, it satisfies that

$$\mathbb{E}[\text{ERR}(\bar{x}_N)] \leq \frac{2\alpha\Omega^2}{N\gamma} + \frac{7\gamma\sigma^2}{\alpha}$$

In particular, if $\gamma = \min\{\frac{\alpha}{\sqrt{3}L}, \alpha\Omega\sqrt{\frac{2}{7N\sigma^2}}\}$, then

$$\mathbb{E}[\text{ERR}(\bar{x}_N)] \leq \max\left\{\frac{7}{2}\frac{\Omega^2 L}{\alpha}\frac{1}{N}, \Omega\sqrt{\frac{14\sigma^2}{N}}\right\}$$

If the problem is deterministic, the original bound of Nemirovski [18] is as follows.

Theorem F.2. [18] *Under the same assumption in Theorem F.1, suppose the problem is deterministic. For $\gamma \leq \frac{\alpha}{\sqrt{2}L}$,*

$$\text{ERR}(\bar{x}_N) \leq \sqrt{2}\frac{\Omega^2 L}{\alpha}\frac{1}{N}$$

Unlike the uniform scheme above, a recent analysis by Ho-Nguyen and Kilinc-Karzan [24] also provides a performance bound the weighted average version of Mirror-Prox when the problem is deterministic.

Theorem F.3. [24] *Under the same assumption in Theorem F.1, suppose the problem is deterministic. Let $\{w_n \geq 0\}$ be a sequence of weights and let the step size to be $\gamma_n = \frac{\alpha}{L}\frac{w_{1:n}}{\max_m w_m}$.*

$$\text{ERR}(\bar{x}_N) \leq \frac{\Omega^2 L}{\alpha}\frac{\max_n w_n}{w_{1:N}}$$

Theorem F.3 (with $w_n = w$) tightens Theorem F.1 and Theorem F.2 by a constant factor.

[13] Here simplify the condition they made by assuming F is Lipschitz continuous and g_n and $\hat{g}_n$ are unbiased.

F.3 Connection with MoBIL-Prox

To relate stochastic Mirror-Prox and MoBIL-Prox, we first rename the variables in (F.1) by setting $\hat{x}_{n+1} \coloneqq \hat{x}_n$ and $\gamma_{n+1} \coloneqq \gamma_n$

$$\begin{aligned} x_{n+1} &= \mathrm{Prox}_{\hat{x}_n}(\gamma_n \hat{g}_n) \\ \hat{x}_{n+1} &= \mathrm{Prox}_{\hat{x}_n}(\gamma_n g_{n+1}) \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} x_{n+1} &= \mathrm{Prox}_{\hat{x}_{n+1}}(\gamma_{n+1} \hat{g}_{n+1}) \\ \hat{x}_{n+2} &= \mathrm{Prox}_{\hat{x}_{n+1}}(\gamma_{n+1} g_{n+1}) \end{aligned}$$

and then reverse the order of updates and write them as

$$\begin{aligned} \hat{x}_{n+1} &= P_{\hat{x}_n}(\gamma_n g_n) \\ x_{n+1} &= P_{\hat{x}_{n+1}}(\gamma_{n+1} \hat{g}_{n+1}) \end{aligned} \tag{F.2}$$

Now we will show that the update in (F.2) is a special case of (9), which we recall below

$$\begin{aligned} \hat{\pi}_{n+1} &= \arg\min_{\pi \in \Pi} \sum_{m=1}^{n} w_m \big(\langle g_m, \pi \rangle + r_m(\pi) \big), \\ \pi_{n+1} &= \arg\min_{\pi \in \Pi} \sum_{m=1}^{n} w_m \big(\langle g_m, \pi \rangle + r_m(\pi) \big) + w_{n+1} \langle \hat{g}_{n+1}, \pi \rangle , \end{aligned} \tag{9}$$

That is, we will show that $x_n = \pi_n$ and $\hat{x} = \hat{\pi}_n$ under certain setting.

Proposition F.2. *Suppose $w_n = \gamma_n$, $\hat{F}_n = F$, $r_1(\pi) = B_\omega(\pi || \pi_1)$ and $r_n = 0$ for $n > 1$. If $\Pi = \mathcal{X}$ is unconstrained, then $x_n = \pi_n$ and $\hat{x}_n = \hat{\pi}_n$ as defined in* (F.2) *and* (9)*.*

Proof. We prove the assertion by induction. For $n = 1$, it is trivial, since $\pi_1 = \hat{\pi}_1 = x_1 = \hat{x}_1$. Suppose it is true for n. We show it also holds for $n + 1$.

We first show $\hat{x}_{n+1} = \hat{\pi}_{n+1}$. By the optimality condition of $\hat{\pi}_{n+1}$, it holds that

$$\begin{aligned} 0 &= \sum_{m=1}^{n} w_m g_m + \nabla\omega(\hat{\pi}_{n+1}) - \nabla\omega(\pi_1) \\ &= (w_n g_n + \nabla\omega(\hat{\pi}_{n+1}) - \nabla\omega(\hat{\pi}_n)) + \left(\sum_{m=1}^{n-1} w_m g_m + \nabla\omega(\hat{\pi}_n) - \nabla\omega(\pi_1) \right) \\ &= w_n g_n + \nabla\omega(\hat{\pi}_{n+1}) - \nabla\omega(\hat{\pi}_n) \end{aligned}$$

where the last equality is by the optimality condition of $\hat{\pi}_n$. This is exactly the optimality condition of $\hat{x}_{n+1}$ given in (F.2), as $\hat{x}_n = \hat{\pi}_n$ by induction hypothesis and $w_n = \gamma_n$. Finally, because Prox is single-valued, it implies $\hat{x}_{n+1} = \hat{\pi}_{n+1}$.

Next we show that $\pi_{n+1} = x_{n+1}$. By optimality condition of π_{n+1}, it holds that

$$\begin{aligned} 0 &= w_{n+1} \hat{g}_{n+1} + \sum_{m=1}^{n} w_m g_m + \nabla\omega(\pi_{n+1}) - \nabla\omega(\pi_1) \\ &= (w_{n+1} \hat{g}_{n+1} + \nabla\omega(\pi_{n+1}) - \nabla\omega(\hat{\pi}_{n+1})) + \left(\sum_{m=1}^{n} w_m g_m + \nabla\omega(\hat{\pi}_{n+1}) - \nabla\omega(\pi_1) \right) \\ &= w_{n+1} \hat{g}_{n+1} + \nabla\omega(\pi_{n+1}) - \nabla\omega(\hat{\pi}_{n+1}) \end{aligned}$$

This is the optimality condition also for x_{n+1}, since we have shown that $\hat{\pi}_{n+1} = \hat{x}_{n+1}$. The rest of the argument follows similarly as above. ■

In other words, stochastic Mirror-Prox is a special case of MoBIL-Prox, when $\hat{F}_n = F$ (i.e. the update of π_n also queries the environment not the simulator) and the regularization is constant. The condition that $\mathcal{X}$ and Π are unconstrained is necessary to establish the exact equivalence between Prox-based updates and FTL-based

updates. This is a known property in the previous studies on the equivalence between lazy mirror descent and FTRL [16]. Therefore, when $\hat{F}_n = F$, we can view MoBIL-Prox as a lazy version of Mirror-Prox. It has been empirical observed the FT(R)L version sometimes empirically perform better than the Prox version [16].

With the connection established by Proposition F.2, we can use a minor modification of the strategy used in Theorem 4.2 to prove the performance of MoBIL-Prox when solving VI problems. To show the simplicity of the FTL-style proof compared with the algebraic proof of Juditsky et al. [11], below we will prove from scratch but only using the new Stronger FTL Lemma (Lemma 4.1).

To do so, we introduce a lemma to relate expected regret and $\mathrm{ERR}(\bar{x}_N)$.

Lemma F.1. *Let F be a monotone operator. For any $\{x_n \in \mathcal{X}\}_{n=1}^N$ and $\{w_n \geq 0\}$,*

$$\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \leq \mathbb{E}\left[\max_{x\in\mathcal{X}} \frac{1}{w_{1:N}} \sum_{n=1}^N w_n \left\langle F(x_n), x_n - x\right\rangle\right]$$

where $\bar{x}_N = \frac{\sum_{n=1}^N w_n x_n}{w_{1:n}}$.

Proof. Let $x^\star \in \arg\max_{x\in\mathcal{X}} \langle F(x), \bar{x}_N - x\rangle$. By monotonicity, for all x_n, $\langle F(x^\star), x_n - x^\star\rangle \leq \langle F(x_n), x_n - x^\star\rangle$. and therefore

$$\begin{aligned}\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] &= \mathbb{E}\left[\frac{1}{w_{1:N}} \sum_{n=1}^N w_n \left\langle F(x^\star), x_n - x^\star\right\rangle\right] \\ &\leq \mathbb{E}\left[\frac{1}{w_{1:N}} \sum_{n=1}^N w_n \left\langle F(x_n), x_n - x^\star\right\rangle\right] \leq \mathbb{E}\left[\max_{x\in\mathcal{X}} \frac{1}{w_{1:N}} \sum_{n=1}^N w_n \left\langle F(x_n), x_n - x\right\rangle\right]\end{aligned}$$

■

Theorem F.4. *Under the same assumption as in Theorem F.1. Suppose $w_n = n^p$ and $r_n(x) = \beta_n B_\omega(x||x_n)$, where β_n is selected such that $\sum_{n=1}^N w_n\beta_n = \frac{1}{\eta} n^k$ for some $k \geq 0$ and $\eta > 0$. If $k > p$, then*

$$\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \leq \frac{1}{w_{1:N}}\left(\frac{\alpha\Omega^2}{\eta} N^k + \frac{3\sigma^2\eta}{\alpha}\sum_{n=1}^N n^{2p-k}\right) + \frac{O(1)}{w_{1:N}}$$

Proof. To simplify the notation, define $l_n(x) = w_n(\langle F(x_n), x\rangle + r_n(x))$ and let

$$\begin{aligned}\mathrm{regret}^w(\mathcal{X}) &= \sum_{n=1}^N w_n \left\langle F(x_n), x_n\right\rangle - \min_{x\in\mathcal{X}} \sum_{n=1}^N w_n \left\langle F(x_n), x\right\rangle \\ \mathcal{R}^w(\mathcal{X}) &= \sum_{n=1}^N l_n(x_n) - \min_{x\in\mathcal{X}} \sum_{n=1}^N l_n(x)\end{aligned}$$

By this definition, it holds that

$$\mathrm{regret}^w(\mathcal{X}) \leq \mathcal{R}^w(\mathcal{X}) + \max_{x\in\mathcal{X}} \sum_{n=1}^N w_n r_n(x)$$

In the following, we bound the two terms in the upper bound above. First, by applying Stronger FTL Lemma (Lemma 4.1) with l_n and we can show that

$$\begin{aligned}\mathcal{R}^w(\mathcal{X}) &\leq \sum_{n=1}^N l_{1:n}(x_n) - l_{1:n}(x_n^\star) - \Delta_n \\ &\leq \sum_{n=1}^N \frac{\eta}{2\alpha} n^{2p-k} \|g_n - \hat{g}_n\|_*^2 - \frac{\alpha(n-1)^{k-1}}{2\eta}\|x_n - \hat{x}_n\|^2\end{aligned}$$

where $x_n^\star \coloneqq \arg\max_{x\in\mathcal{X}} l_{1:n}(x)$. Because by Lemma H.3 and Lipschitz continuity of F, it holds

$$\|g_n - \hat{g}_n\|_*^2 \le 3(L^2\|x_n - \hat{x}_n\|^2 + 2\sigma^2) \tag{F.3}$$

Therefore, we can bound

$$\mathcal{R}^w(\mathcal{X}) \le \sum_{n=1}^N \left(\frac{3}{2}\frac{L^2\eta}{\alpha}n^{2p-k} - \frac{\alpha}{2\eta}(n-1)^k\right)\|x_n - \hat{x}_n\|^2 + \frac{3\sigma^2\eta}{\alpha}\sum_{n=1}^N n^{2p-k} \tag{F.4}$$

If $k > p$, then the first term above is $O(1)$ independent of N. On the other hand,

$$\max_{x\in\mathcal{X}} \sum_{n=1}^N w_n r_n(x) \le \frac{\alpha\Omega^2}{\eta}N^k \tag{F.5}$$

Combining the two bounds and Lemma F.1, i.e. $\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \le \mathbb{E}\left[\frac{\mathrm{regret}^w(\mathcal{X})}{w_{1:N}}\right]$ concludes the proof. ■

Deterministic Problems For deterministic problems, we specialize the proof Theorem F.4 gives. We set $k = p = 0$, $x_1 = \arg\min_{x\in\mathcal{X}} \omega(x)$, which removes the 2 factor in (F.5), and modify 3 to 1 in (F.3) (because the problem is deterministic). By recovering the constant in the proof, we can show that

$$\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \le \frac{1}{N}\left(\frac{\alpha\Omega^2}{\eta} + \sum_{n=1}^N \left(\frac{1}{2}\frac{L^2\eta}{\alpha} - \frac{\alpha}{2\eta}\right)\|x_n - \hat{x}_n\|^2\right)$$

Suppose . We choose η to make the second term non-positive, i.e.

$$\frac{1}{2}\frac{L^2\eta}{\alpha} - \frac{\alpha}{2\eta} \le 0 \iff \eta \le \frac{\alpha}{L}$$

and the error bound becomes

$$\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \le \frac{L\Omega^2}{N}$$

This bound and the condition on η matches that in [24].

Stochastic Problems For stochastic problems, we use the condition specified in Theorem F.4. Suppose $2p - k > -1$. To balance the second term in (F.4) and (F.5), we choose

$$2p - k + 1 = k \implies k = p + \frac{1}{2}$$

To satisfy the hypothesis $2p - k > -1$, it requires $p > -\frac{1}{2}$. Note with this choice, it satisfies the condition $k > p$ required in Theorem F.4. Therefore, the overall bound becomes

$$\begin{aligned}
\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] &\le \frac{1}{w_{1:N}}\left(\frac{\alpha\Omega^2}{\eta}N^{p+\frac{1}{2}} + \frac{3\sigma^2\eta}{\alpha}\sum_{n=1}^N n^{p-\frac{1}{2}}\right) + \frac{O(1)}{w_{1:N}} \\
&\le \frac{p+1}{N^{p+1}}\left(\frac{\alpha\Omega^2}{\eta} + \frac{3\eta\sigma^2}{\alpha(p+\frac{1}{2})}\right)(N+1)^{p+\frac{1}{2}} + \frac{O(1)}{N^{p+1}} \\
&\le e^{\frac{p+1/2}{N}}(p+1)\left(\frac{\alpha\Omega^2}{\eta} + \frac{3\eta\sigma^2}{\alpha(p+\frac{1}{2})}\right)N^{-\frac{1}{2}} + \frac{O(1)}{N^{p+1}}
\end{aligned}$$

where we use Lemma H.1 and $(\frac{N+1}{N})^{p+1/2} \le e^{\frac{p+1/2}{N}}$. If we set η such that

$$\frac{\alpha\Omega^2}{\eta} = \frac{3\eta\sigma^2}{\alpha(p+\frac{1}{2})} \implies \eta = \alpha\frac{\Omega}{\sigma}\sqrt{\frac{p+\frac{1}{2}}{3}}$$

Then

$$\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \leq 2e^{\frac{p+1/2}{N}}(p+1)\Omega\sigma\sqrt{\frac{3}{p+\frac{1}{2}}}N^{-\frac{1}{2}} + \frac{O(1)}{N^{p+1}} \tag{F.6}$$

For example, if $p = 0$, then

$$\mathbb{E}[\mathrm{ERR}(\bar{x}_N)] \leq \frac{O(1)}{N} + \frac{2\sqrt{6}\sigma\Omega e^{\frac{p+1/2}{N}}}{\sqrt{N}}$$

which matches the bound in by Juditsky et al. [11] with a slightly worse constant. We leave a complete study of tuning p as future work.

F.4 Comparison of stochastic Mirror-Prox and MoBIL-Prox in Imitation Learning

The major difference between stochastic Mirror-Prox and MoBIL-Prox is whether the gradient from the environment is used to also update the decision π_{n+1}. It is used in the Mirror-Prox, whereas MoBIL-Prox uses the estimation from simulation. Therefore, for N iterations, MoBIL-Prox requires only N interactions, whereas Mirror-Prox requires $2N$ interactions.

The price MoBIL-Prox pays extra when using the estimated gradient is that a secondary online learning problem has to be solved. This shows up in the term, for example of strongly convex problems,

$$\frac{(p+1)G_h^2}{2\mu_h}\frac{1}{N^2} + \frac{\epsilon_{\hat{\mathcal{F}}}^w + \sigma_g^2 + \sigma_{\hat{g}}^2}{N}$$

in Theorem 4.2. If both gradients are from the environment, then $\epsilon_{\hat{\mathcal{F}}}^w = 0$ and $\sigma_{\hat{g}}^2 = \sigma_g^2$. Therefore, if we ignore the $O(\frac{1}{N^2})$ term, using an estimated gradient to update π_{n+1} is preferred, if it requires less interactions to get to the magnitude of error, i.e.

$$2 \times 2\sigma_g^2 \geq \epsilon_{\hat{\mathcal{F}}}^w + \sigma_{\hat{g}}^2 + \sigma_g^2$$

in which the multiplier of 2 on the left-hand side is due to MoBIL-Prox only requires one interaction per iterations, whereas stochastic Mirror-Prox requires two.

Because σ_g^2 is usually large in real-world RL problems and $\sigma_{\hat{g}}^2$ can be made close to zero easily (by running more simulations), if our model class is reasonably expressive, then MoBIL-Prox is preferable. Essentially, this is because MoBIL-Prox can roughly cut the noise of gradient estimates by half.

The preference over MoBIL-Prox would be more significant for convex problems, because the error decays slower over iterations (e.g. $\frac{1}{\sqrt{N}}$) and therefore more iterations are required by the stochastic Mirror-Prox approach to counter balance the slowness due to using noisy gradient estimator.

G Experimental Details

G.1 Tasks

Two robot control tasks (Cartpole and Reacher3D) powered by the DART physics engine [19] were used as the task environments.

Cartpole The Cart-Pole Balancing task is a classic control problem, of which the goal is to keep the pole balanced in an upright posture with force only applied to the cart. The state and action spaces are both continuous, with dimension 4 and 1, respectively. The state includes the horizontal position and velocity of the cart, and the angle and angular velocity of the pole. The time-horizon of this task is 1000 steps. There is a small uniformly random perturbation injected to initial state, and the transition is deterministic. The agent receives +1 reward for every time step it stays in a predefined region, and a rollout terminates when the agent steps outside the region.

Reacher3D In this task, a 5-DOF (degrees-of-freedom) manipulator is controlled to reach a random target position in a 3D space. The reward is the sum of the negative distance to the target point from the finger tip and a control magnitude penalty. The actions correspond to the torques applied to the 5 joints. The time-horizon of this task is 500 steps. At the beginning of each rollout, the target point to reach is reset to a random location.

G.2 Algorithms

Policies We employed Gaussian policies in our experiments, i.e. for any state $s \in \mathbb{S}$, π_s is Gaussian distributed. The mean of π_s was modeled by either a linear function or a neural network that has 2 hidden layers of size 32 and tanh activation functions. The covariance matrix of π_s was restricted to be diagonal and independent of state. The expert policies in the IL experiments share the same architecture as the corresponding learners (e.g. a linear learner is paired with a linear expert) and were trained using actor-critic-based policy gradients.

Imitation learning loss With regard to the IL loss, we set $D(\pi_s^*||\pi_s)$ in (2) to be the KL-divergence between the two Gaussian distributions: $D(\pi_s^*||\pi_s) = \mathrm{KL}[\pi_s||\pi_s^*]$. (We observed that using $\mathrm{KL}[\pi_s||\pi_s^*]$ converges noticeably faster than using $\mathrm{KL}[\pi_s^*||\pi_s]$).

Implementation details of MoBIL-Prox The regularization of MoBIL-Prox was set to $r_n(\pi) = \frac{\mu_f \alpha_n}{2}\|\pi - \pi_n\|^2$ such that $\sum w_n \alpha_n \mu_f = (1 + cn^{p+1/2})/\eta_n$, where $c = 0.1$ and η_n was adaptive to the norm of the prediction error. Specifically, we used $\eta_n = \eta\lambda_n$: $\eta > 0$ and λ_n is a moving-average estimator of the norm of $e_n = g_n - \hat{g}_n$ defined as

$$\bar{\lambda}_n = \beta\bar{\lambda}_{n-1} + (1-\beta)\|e_n\|_2$$
$$\lambda_n = \bar{\lambda}_n/(1-\beta^n)$$

where β was chosen to be 0.999. This parameterization is motivated by the form of the optimal step size of MoBIL-Prox in Theorem 4.2, and by the need of having adaptive step sizes so different algorithms are more comparable. The model-free setting was implemented by setting $\hat{g}_n = 0$ in MoBIL-Prox, and the same adaptation rule above was used (which in this case effectively adjusts the learning rate based on $\|g_n\|$). In the experiments, η was selected to be 0.1 and 0.01 for $p = 0$ and $p = 2$, respectively, so the areas under the effective learning rate $\eta_n w^p/(1 + cn^{p+1/2})$ for $p = 0$ and $p = 2$ are close, making MoBIL-Prox perform similarly in these two settings.

In addition to the update rule of MoBIL-Prox, a running normalizer, which estimates the upper and the lower bounds of the state space, was used to center the state before it was fed to the policies.

Dynamics model learning The dynamics model used in the experiments is deterministic (the true model is deterministic too). It is represented by a neural network with 2 hidden layers of size 64 and tanh activation functions. Given a batch of transition triples $\{(s_{t_k}, a_{t_k}, s_{t_k+1})\}_{k=1}^K$ collected by running π_n under the true dynamics in each round, we set the per-round cost for model learning as $\frac{1}{K}\sum_{k=1}^K \|s_{t_{k+1}} - M(s_{t_k}, a_{t_k})\|_2^2$, where M is the neural network dynamics model. It can be shown that this loss is an upper bound of $\|\nabla_2 F(\pi_n, \pi_n) - \nabla_2 \hat{F}_n(\pi_n, \pi_n)\|_*^2$ by applying a similar proof as in Appendix D. The minimization was achieved through gradient descent using ADAM [25] with a fixed number of iterations (2048) and fixed-sized mini-batches (128). The step size of ADAM was set to 0.001.

H Useful Lemmas

This section summarizes some useful properties of polynomial partial sum, sequence in Banach space, and variants of FTL in online learning. These results will be useful to the proofs in Appendix C.

H.1 Polynomial Partial Sum

Lemma H.1. *This lemma provides estimates of* $\sum_{n=1}^{N} n^p$.

1. *For* $p > 0$, $\frac{N^{p+1}}{p+1} = \int_0^N x^p dx \le \sum_{n=1}^N n^p \le \int_1^{N+1} x^p dx \le \frac{(N+1)^{p+1}}{p+1}$.
2. *For* $p = 0$, $\sum_{n=1}^N n^p = N$.
3. *For* $-1 < p < 0$,
$$\frac{(N+1)^{p+1}-1}{p+1} = \int_1^{N+1} x^p dx \le \sum_{n=1}^N n^p \le 1 + \int_1^N x^p dx = \frac{N^{p+1}+p}{p+1} \le \frac{(N+1)^{p+1}}{p+1}.$$
4. *For* $p = -1$, $\ln(N+1) \le \sum_{n=1}^N n^p \le \ln N + 1$.
5. *For* $p < -1$, $\sum_{n=1}^N n^p \le \frac{N^{p+1}+p}{p+1} = O(1)$. *For* $p = -2$, $\sum_{n=1}^N n^p \le \frac{N^{-1}-2}{-2+1} \le 2$.

Lemma H.2. *For* $p \ge -1$, $N \in \mathbb{N}$,
$$S(p) = \sum_{n=1}^N \frac{n^{2p}}{\sum_{m=1}^n m^p} \le \begin{cases} \frac{p+1}{p}(N+1)^p, & \text{for } p > 0 \\ \ln(N+1), & \text{for } p = 0 \\ O(1), & \text{for } -1 < p < 0 \\ 2, & \text{for } p = -1 \end{cases}.$$

Proof. If $p \ge 0$, by Lemma H.1,
$$S(p) = (p+1)\sum_{n=1}^N n^{p-1} \le \begin{cases} \frac{p+1}{p}(N+1)^p, & \text{for } p > 0 \\ \ln(N+1), & \text{for } p = 0 \end{cases}.$$
If $-1 < p < 0$, by Lemma H.1, $S(p) \le (p+1)\sum_{n=1}^N \frac{n^{2p}}{(n+1)^{p+1}-1}$. Let $a_n = \frac{n^{2p}}{(n+1)^{p+1}-1}$, and $b_n = n^{p-1}$. Since $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$ and by Lemma H.1 $\sum_{n=0}^\infty b_n$ converges, thus $\sum_{n=0}^\infty a_n$ converges too. Finally, if $p = -1$, by Lemma H.1, $S(-1) \le \sum_{n=1}^N \frac{1}{n^2 \ln(n+1)} \le \sum_{n=1}^N \frac{1}{n^2} \le 2$. ∎

H.2 Sequence in Banach Space

Lemma H.3. *Let* $\{a = x_0, x_1, \cdots, x_N = b\}$ *be a sequence in a Banach space with norm* $\|\cdot\|$. *Then for any* $N \in \mathbb{N}_+$, $\|a-b\|^2 \le N\sum_{n=1}^N \|x_{n-1} - x_n\|^2$.

Proof. First we note that by triangular inequality it satisfies that $\|a-b\| \le \sum_{n=1}^N \|x_{n-1} - x_n\|$. Then we use the basic fact that $2ab \le a^2 + b^2$ in the second inequality below and prove the result.
$$\begin{aligned}
\|a-b\|^2 &\le \sum_{n=1}^N \|x_{n-1} - x_n\|^2 + \sum_{n=1}^N \sum_{m=1;m\neq n}^N \|x_{n-1} - x_n\|\|x_{m-1} - x_m\| \\
&\le \sum_{n=1}^N \|x_{n-1} - x_n\|^2 + \sum_{n=1}^N \sum_{m=1;m\neq n}^N \frac{1}{2}\left(\|x_{n-1} - x_n\|^2 + \|x_{m-1} - x_m\|^2\right) \\
&= \sum_{n=1}^N \|x_{n-1} - x_n\|^2 + \frac{N-1}{2}\sum_{n=1}^N \|x_{n-1} - x_n\|^2 + \frac{1}{2}\sum_{n=1}^N \sum_{m=1;m\neq n}^N \|x_{m-1} - x_m\|^2
\end{aligned}$$

$$= \sum_{n=1}^{N} \|x_{n-1} - x_n\|^2 + (N-1) \sum_{n=1}^{N} \|x_{n-1} - x_n\|^2$$
$$= N \sum_{n=1}^{N} \|x_{n-1} - x_n\|^2$$ ■

H.3 Basic Regret Bounds of Online Learning

For the paper to be self-contained, we summarize some fundamental results of regret bound when the learner in an online problem updates the decisions by variants of FTL. Here we consider a general setup and therefore use a slightly different notation from the one used in the main paper for policy optimization.

Online Learning Setup Consider an online convex optimization problem. Let $\mathcal{X}$ be a compact decision set in a normed space with norm $\|\cdot\|$. In round n, the learner plays $x_n \in \mathcal{X}$ receives a convex loss $l_n : \mathcal{X} \to \mathbb{R}$ satisfying $\|\nabla l_n(x_n)\|_* \leq G$, and then make a new decision $x_{n+1} \in \mathcal{X}$. The *regret* is defined as

$$\text{regret}(\mathcal{X}) = \sum_{n=1}^{N} l_n(x_n) - \min_{x \in \mathcal{X}} \sum_{n=1}^{N} l_n(x)$$

More generally, let $\{w_n \in \mathbb{R}_+\}_{n=1}^N$ be a sequence of weights. The *weighted regret* is defined as

$$\text{regret}^w(\mathcal{X}) = \sum_{n=1}^{N} w_n l_n(x_n) - \min_{x \in \mathcal{X}} \sum_{n=1}^{N} w_n l_n(x)$$

In addition, we define a constant $\epsilon_{\mathcal{X}}^w$ (which can depend on $\{l_n\}_{n=1}^N$) such that

$$\epsilon_{\mathcal{X}}^w \geq \min_{x \in \mathcal{X}} \frac{\sum_{n=1}^{N} w_n l_n(x)}{w_{1:N}}.$$

In the following, we prove some basic properties of FTL with prediction. At the end, we show the result of FTL as a special case. These results are based on the Strong FTL Lemma (Lemma 3.2), which can also be proven by Stronger FTL Lemma (Lemma 4.1).

Lemma 3.2 (Strong FTL Lemma [16])**.** *For any sequence of decisions $\{x_n \in \mathcal{X}\}$ and loss functions $\{l_n\}$, $\text{regret}(\mathcal{X}) \leq \sum_{n=1}^{N} l_{1:n}(x_n) - l_{1:n}(x_n^\star)$, where $x_n^\star \in \arg\min_{x \in \mathcal{X}} l_{1:n}(x)$, where $\mathcal{X}$ is the decision set.*

To use Lemma 3.2, we first show an intermediate bound.

Lemma H.4. *In round n, let $l_{1:n}$ be $\mu_{1:n}$-strongly convex for some $\mu_{1:n} > 0$, and let v_{n+1} be a (non)convex function such that $l_{1:n} + v_{n+1}$ is convex. Suppose the learner plays FTL with prediction, i.e. $x_{n+1} \in \arg\min_{x \in \mathcal{X}} (l_{1:n} + v_{n+1})(x)$. Then it holds*

$$\sum_{n=1}^{N} (l_{1:n}(x_n) - l_{1:n}(x_n^\star)) \leq \sum_{n=1}^{N} \frac{1}{2\mu_{1:n}} \|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2$$

where $x_n^\star = \arg\min_{\mathcal{X}} \sum_{n=1}^{N} l_n(x)$.

Proof. For any $x \in \mathcal{X}$, since $l_{1:n}$ is $\mu_{1:n}$ strongly convex, we have

$$l_{1:n}(x_n) - l_{1:n}(x) \leq \langle \nabla l_{1:n}(x_n), x_n - x \rangle - \frac{\mu_{1:n}}{2} \|x_n - x\|^2. \tag{H.1}$$

And by the hypothesis $x_n = \arg\min_{x \in \mathcal{X}} (l_{1:n-1} + v_n)(x)$, it holds that

$$\langle -\nabla l_{1:n-1}(x_n) - \nabla v_n(x_n), x_n - x \rangle \geq 0. \tag{H.2}$$

Adding (H.1) and (H.2) yields

$$\begin{aligned} l_{1:n}(x_n) - l_{1:n}(x) &\le \langle \nabla l_n(x_n) - \nabla v_n(x_n), x_n - x\rangle - \frac{\mu_{1:n}}{2}\|x_n - x\|^2 \\ &\le \max_d \langle \nabla l_n(x_n) - \nabla v_n(x_n), d\rangle - \frac{\mu_{1:n}}{2}\|d\|^2 \\ &= \frac{1}{2\mu_{1:n}}\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2, \end{aligned}$$

where the last equality is due to a property of dual norm (e.g. Exercise 3.27 of [26]). Substituting $x_n^\star$ for x and taking the summation over n prove the lemma. ■

Using Lemma 3.2 and Lemma H.4, we can prove the regret bound of FTL with prediction.

Lemma H.5 (FTL with prediction)**.** *Let l_n be a μ_n-strongly convex for some $\mu_n \ge 0$. In round n, let v_{n+1} be a (non)convex function such that $\sum_{m=1}^n w_m l_m + w_{m+1}v_{n+1}$ is convex. Suppose the learner plays FTL with prediction, i.e. $x_{n+1} \in \arg\min_{x\in\mathcal{X}} \sum_{m=1}^n (w_m l_m + w_{m+1}v_{n+1})(x)$ and suppose that $\sum_{m=1}^n w_m\mu_m > 0$. Then*

$$\mathrm{regret}^w(\mathcal{X}) \le \sum_{n=1}^N \frac{w_n^2\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2}{2\sum_{m=1}^n w_m\mu_m}$$

In particular, if $\mu_n = \mu$, $w_n = n^p$, $p \ge 0$, $\mathrm{regret}^w(\mathcal{X}) \le \frac{p+1}{2\mu}\sum_{n=1}^N n^{p-1}\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2$.

Proof. By Lemma 3.2 and Lemma H.4, we see

$$\mathrm{regret}^w(\mathcal{X}) \le \sum_{n=1}^N (l_{1:n}(x_n) - l_{1:n}(x_n^\star)) \le \sum_{n=1}^N \frac{w_n^2\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2}{2\sum_{m=1}^n w_m\mu_m}.$$

If $\mu_n = \mu$, $w_n = n^p$, and $p \ge 0$, then it follows from Lemma H.1

$$\mathrm{regret}^w(\mathcal{X}) \le \frac{1}{2\mu}\sum_{n=1}^N \frac{n^{2p}}{\frac{n^{p+1}}{p+1}}\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2 = \frac{p+1}{2\mu}\sum_{n=1}^N n^{p-1}\|\nabla l_n(x_n) - \nabla v_n(x_n)\|_*^2.$$

■

The next lemma about the regret of FTL is a corollary of Lemma H.5.

Lemma H.6 (FTL)**.** *Let l_n be μ-strongly convex for some $\mu > 0$. Suppose the learner play FTL, i.e. $x_n = \arg\min_{x\in\mathcal{X}} \sum_{m=1}^n w_m l_m(x)$. Then $\mathrm{regret}^w(\mathcal{X}) \le \frac{G^2}{2\mu}\sum_{n=1}^N \frac{w_n^2}{w_{1:n}}$. In particular, if $w_n = n^p$, then*

$$\sum_{n=1}^N w_n l_n(x_n) \le \begin{cases} \frac{G^2}{2\mu}\frac{p+1}{p}(N+1)^p + \frac{1}{p+1}(N+1)^{p+1}\epsilon_{\mathcal{X}}^w, & \text{for } p > 0 \\ \frac{G^2}{2\mu}\ln(N+1) + N\epsilon_{\mathcal{X}}^w, & \text{for } p = 0 \\ \frac{G^2}{2\mu}O(1) + \frac{1}{p+1}(N+1)^{p+1}\epsilon_{\mathcal{X}}^w, & \text{for } -1 < p < 0 \\ \frac{G^2}{\mu} + (\ln N + 1)\epsilon_{\mathcal{X}}^w, & \text{for } p = -1 \end{cases}$$

Proof. By definition of $\mathrm{regret}^w(\mathcal{X})$, the absolute cost satisfies $\sum_{n=1}^N w_n l_n(x_n) = \mathrm{regret}^w(\mathcal{X}) + \min_{x\in\mathcal{X}}\sum_{n=1}^N w_n l_n(x)$. We bound the two terms separately. For $\mathrm{regret}^w(\mathcal{X})$, set $v_n = 0$ in Lemma H.5 and we have

$$\begin{aligned} \mathrm{regret}^w(\mathcal{X}) &\le \frac{G^2}{2\mu}\sum_{n=1}^N \frac{w_n^2}{w_{1:n}} && \text{(Lemma H.5 and gradient bound)} \\ &= \frac{G^2}{2\mu}\sum_{n=1}^N \frac{n^{2p}}{\sum_{m=1}^n m^p} && \text{(Special case } w_n = n^p\text{)}, \end{aligned}$$

in which $\sum_{n=1}^N \frac{n^{2p}}{\sum_{m=1}^n m^p}$ is exactly what H.2 bounds. On the other hand, the definition of $\epsilon_{\mathcal{X}}^w$ implies that $\min_{x\in\mathcal{X}}\sum_{n=1}^N w_n l_n(x) \le w_{1:N}\epsilon_{\mathcal{X}}^w = \sum_{n=1}^N n^p\epsilon_{\mathcal{X}}^w$, where $\sum_{n=1}^N n^p$ is bounded by Lemma H.1. Combining these two bounds, we conclude the lemma. ■