Review of the manuscript:

Following the Leader and Fast Rates in Linear Prediction: Curved Constraint Sets and Other Regularities

Contribution

The list of the main contributions of the paper is as follows:

- Showing that FTL achieves logarithmic regret under if the minimal curvature of the constraint set is lower bounded by $\lambda_0 > 0$, and the cumulative loss is bounded away from 0, $\|\Theta_t\|_2 > 0$ for all t (Theorem 4). This is the main result of the paper, which is also equipped with a matching lower bound (Theorem 6), which shows that FTL is essentially optimal in this scenario.
- Showing that FTL gets constant regret if the constraint set is a polytope. While the fact that the regret of FTL can be upper bounded by the number of leader changes is well known in the expert setting, it is nice to see that it easily generalizes to arbitrary polytopes (Theorem 7). I think that Theorem 4 and 7 together generalize all known results on FTL in the stochastic setup (the authors show that in the stochastic setting, assumptions of Theorem 4 and 7 are met with high probability).
- Showing equivalence between strongly convex sets and sets with curvature. This result is important on its own.
- Adaptive algorithms for linear game. While $(\mathcal{A}, \mathcal{B})$-prod is a rather straightforward application of known result, the other two algorithms are more interesting. Follow the Shrunken Leader (FTSL) is very similar to the minimax algorithm already considered by Abernethy et al. (2008), and I found it quite surprising that a small modification of the minimax algorithm leads to an algorithm which is adaptive and gets logarithmic regret as long as $\|\Theta_t\|_2 > 0$. Even more surprising is that a simple FTRL strategy with essentially standard tuning of the learning rate is adaptive, in the sense that it achieves fast rate $O(\log n)$ in the stochastic setting as long as $|\mu\|_2 > 0$ (Theorem 11).

The contribution is in my opinion strong and significant, and I think the paper is definitely worth accepting for publication in JMRL. However, minor revision is necessary.

Comments

I think it would be worth doing the paper a bit more self-contained when it comes to differential geometry background. I was not able to follow some of the arguments in the proof due to my lack of knowledge in this field.

Your result on fast rate of FTRL (Theorem 11) reminds me of a result by Koolen et al.: *Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning*, NIPS 2016, precisely their Lemma 5 (from `https://arxiv.org/abs/1605.06439`), which concerns predicting with hinge loss on the unit ball (unit ball constraints imply that hinge is never active and the loss is actually linear). They provide an algorithm (MetaGrad) which achieves fast rates under the same assumptions (i.i.d. + "$|\mu\|_2 > 0$"), but their algorithm is actually more complicated – MetaGrad is a second-order algorithm. You achieve fast rate using a simple FTRL first-order algorithm, with (I think) even a better dependence on the dimension d (there is no dependence in your case)! I would encourage you to compare these two results and check whether you also get fast rate in the hinge loss case.

Regarding simulations: I am not sure how to interpret the results in light of Theorem 11, where you show that FTRL is also adaptive and gets logarithmic regret for $L > 0$. On the other hand, for $L = 0.1$ in the experiments (Figure 3), FTRL behaves much worse than FTL, despite its adaptiveness. Is it only due to a worse constant in front of the $\log(n)$ bound for FTRL, or the rate of FTRL is actually worse? I understand Theorem 11 does not directly apply here due to ellipsoid constraints, but I guess the situation should not be that much different from unit ball constraints.

Would your adaptive algorithms (FTSL, FTRL) generalize to the case of ellipsoid constraints (i.e. $w^\top Q w \le 1$)? If not, what breaks down?

Regarding "subgradient trick" discussed on page 3 (2nd paragraph of Sec. 2): I would be careful with applying your algorithm to general convex functions. This is because f_t's are then subgradients, which are produced by the algorithm itself, and it is hard to control the lower bound on L_n. For instance, if the optimal comparator (the minimizer of cumulative convex loss) is in the interior of $\mathcal{W}$ (which can easily happen for strictly convex losses), your algorithm only predicts with points on the boundary, and hence the regret cannot be sublinear. This means that FTL would produce subgradients which cancel each other so that Θ_t vanishes as $t \to \infty$.

Where do you use Proposition 3? While I found it to be a very elegant way of rewriting the regret, I do not see any application of this result in the paper. By the way, a very similar expression for the regret has been shown by Abernethy et al.: *Online Linear Optimization via Smoothing*, COLT 2014 (`https://arxiv.org/abs/1405.6076`), Lemma 2, in the context of Follow the Perturbed Leader (but with zero noise it actually boils down to Proposition 3).

Alternative proof of Eq. 4

The paper gives an equivalence between $\mathcal{W}$ being a λ-strongly convex set and $\mathcal{W}$ having curvature at least λ. Given that, and due to the fact that I have very little understanding of differential geometry, I was curious whether Theorem 4 could be proved without referring to the curvature at all (which I do not fully comprehend), and only using the strong convexity of $\mathcal{W}$ instead. Below is a simple proof of Equation (4) which does that, at the price of losing a constant factor of 4 in the bound. I hope this proof will be of interests to the authors (provided it is correct).

Define $\theta_\gamma = \gamma\theta_1 + (1-\gamma)\theta_2$ for some $\gamma \in (0,1)$. From convexity of Φ,

$$\Phi(\theta_\gamma) \;=\; \Phi(\gamma\theta_1 + (1-\gamma)\theta_2) \;\le\; \gamma\Phi(\theta_1) + (1-\gamma)\Phi(\theta_2) \;=\; \gamma\langle w^{(1)}, \theta_1\rangle + (1-\gamma)\langle w^{(2)}, \theta_2\rangle. \qquad (1)$$

From λ-strong convexity of $\mathcal{W}$:

$$w_\gamma \;\stackrel{\text{def}}{=}\; \gamma w^{(1)} + (1-\gamma)w^{(2)} + \gamma(1-\gamma)\frac{\lambda}{2}\|w^{(1)} - w^{(2)}\|^2 \frac{\theta_\gamma}{\|\theta_\gamma\|} \;\in\; \mathcal{W}.$$

From the definition of Φ and from $w_\gamma \in \mathcal{W}$:

$$\begin{aligned}\Phi(\theta_\gamma) \;\ge\; \langle w_\gamma, \theta_\gamma\rangle \;&=\; \gamma^2\langle w^{(1)}, \theta_1\rangle + \gamma(1-\gamma)\langle w^{(1)}, \theta_2\rangle + \gamma(1-\gamma)\langle w^{(2)}, \theta_1\rangle + (1-\gamma)^2\langle w^{(2)}, \theta_2\rangle \\ &\quad + \gamma(1-\gamma)\frac{\lambda}{2}\|w^{(1)} - w^{(2)}\|^2\|\theta_\gamma\|. \qquad (2)\end{aligned}$$

Combining (1) and (2), and dividing by $\gamma(1-\gamma)$ gives:

$$\frac{\lambda}{2}\|w^{(1)} - w^{(2)}\|^2\|\theta_\gamma\| \le \langle w^{(1)} - w^{(2)}, \theta_1 - \theta_2\rangle \le \|w^{(1)} - w^{(2)}\|\|\theta_1 - \theta_2\|,$$

where the last inequality is from Cauchy-Schwarz. Since this inequality holds for any $\gamma \in (0,1)$, and since $\theta_1, \theta_2 \neq 0$, it also holds for $\gamma = 0$ (i.e. $\theta_\gamma = \theta_2$), giving:

$$\|w^{(1)} - w^{(2)}\| \le \frac{2\|\theta_1 - \theta_2\|}{\lambda\|\theta_2\|}.$$

Now,

$$\langle w^{(1)} - w^{(2)}, \theta_1 \rangle \leq \langle w^{(1)} - w^{(2)}, \theta_1 \rangle + \underbrace{\langle w^{(2)} - w^{(1)}, \theta_2 \rangle}_{\geq 0} = \langle w^{(1)} - w^{(2)}, \theta_1 - \theta_2 \rangle$$
$$\leq \|w^{(1)} - w^{(2)}\| \|\theta_1 - \theta_2\| \leq \frac{2\|\theta_1 - \theta_2\|^2}{\lambda \|\theta_2\|}.$$

Some minor remarks

- Page 3, 1st paragraph of Sec. 2 "In round every round" $\rightarrow$ "In every round".
- Section 3.1: planar curve parametrization – there is a missing square root over $x'(s)^2 + y'(s)^2$.
- Section 3.1: you mix s and $\mathbf{s}$ as a parameter, also $\mathbf{0}$ should be replaced by 0. Moreover, you use norm $\|\cdot\|$ for some dot products (which are numbers). I would recommend a careful check of the paragraph starting from "Given a C^2...".
- Page 6, line 6: "interpreated" $\rightarrow$ "interpreted".
- Page 8, line 8: "By definition, $u_\gamma(0) = \tilde{\theta}_1$ and $u_\gamma(l) = \hat{\theta}_2$" – I think this requires a bit more explanation. I guess this is actually the place where you make use of the definition of $w^{(1)}$ and $w^{(2)}$.
- Page 8, below Eq. 7: "$\gamma'(s)$ is a unit vector parallel to P. Moreover $u'_\gamma(s)$ is parallel to $\gamma'(s)$ and $\lambda(s) = \|u'_\gamma(s)\|_2$" – are there any simple arguments you could provide to show why these statements hold? (maybe I am missing something trivial but I do not see it to be straightforward).
- The dependence on the dimension d in Corollary 8 could be improved by bounding the loss in the first τ trials by $MW\tau$, while for the remaining trials using $\sum_{t=\tau}^{n} e^{-\alpha t} \leq \frac{1}{\alpha} e^{-\alpha \tau}$ and finally tuning $\tau = \frac{1}{\alpha} \log(d/\alpha)$.
- Page 13, 1st paragraph of Sec. 4: change quotation marks in "Follow the Regularized Leader".
- The first three sentences of Theorem 10 should probably be placed before Theorem 10 (?).
- Page 16, below (18): "accumulate" $\rightarrow$ "accumulated".