

1 **Experiments and Baseline** Reviewer 1 is concerned that our paper is incomplete without experiments. We respect-
2 fully disagree, and would like to make several points:

3 **1.** Our paper asks and answers a basic theoretical question: is satisfying statistical fairness constraints at every round
4 compatible *even in principle* with obtaining the optimal \sqrt{T} rate-of-regret in one-sided feedback settings? Prior to our
5 work, this question had not been asked, and there was some reason for pessimism: Joseph et al. [20] had shown that
6 stronger individual notions of fairness were not necessarily compatible with optimal regret bounds.

7 **2.** One reason experiments are useful is to establish if a method outperforms previously studied algorithms for the same
8 problem. But as we are the first to study the problem of satisfying statistical fairness constraints in partial-information
9 online classification settings, there is no relevant prior work to compare to. We note that prior work on fair batch
10 classification does suggest one basic explore-then-exploit style baseline: during an “exploration round” classify all
11 examples as positive, and then run a fair batch classification algorithm to pick a hypothesis to play for all subsequent
12 rounds. Its not hard to see that this baseline, when optimized, obtains only a sub-optimal $T^{2/3}$ regret bound, as
13 compared to our optimal \sqrt{T} regret bound. We are happy to elaborate on this simple baseline in the paper.

14 **3.** Another reason for experimental results is to demonstrate that the proposed algorithm is implementable, and not
15 simply a theoretical “upper bound”. But as we note in the paper, our algorithm involves a careful combination of two
16 works: “A Reductions Approach to Fair Classification” from ICML 2018 and “Taming the Monster” from ICML 2014.
17 We need to make modifications to the details of these algorithms to make them work in our setting, but not in any way
18 that would affect their running time. Both algorithms are known to run efficiently in practice, with code available.

19 In summary, although we agree with the reviewer that experimental work could be interesting, we disagree that it would
20 be a major contribution of this work, or that the paper is incomplete without it. The main results of the paper cannot be
21 established experimentally, and there is no prior work to make empirical comparisons to.

22 **Clarity and Discussion** Thank you for your useful suggestions about clarity: we agree, and will try and follow your
23 advice. In particular, we will attempt to re-work Section 3 to focus less on technical details and more on giving intuition
24 for where and why the constraints of fairness and partial feedback require modification of existing techniques, deferring
25 more technical details to the supplement. Briefly, in addition to converting our partial feedback setting to the more
26 standard contextual bandits setting using the reduction from Section 2 of our paper, we face three high level obstacles:
27 1) We need to modify the “fair reductions” oracle to be able to handle fairness constraints arising from a *different data*
28 *set* than the error objective. 2) We need to modify the “mini-monster” contextual bandit algorithm to work with an
29 approximate oracle, rather than an exact one (because we cannot implement an exact oracle using step 1). 3) We need to
30 modify the analysis of “mini-monster” to be able to handle our infinite hypothesis class (because our fairness constraints
31 force our hypothesis class to consist of all convex combinations of base classifiers, even if our base class is finite). Each
32 of these steps requires some technical work, but we agree that focusing on the basic outline will be clearer.

33 We will also expand discussion at your suggestion about the distinction between three fairness constraints that one
34 could ask for, at increasing levels of strength: 1) Satisfying the equal false-positive-rate condition only asymptotically
35 (i.e. on average, in hindsight after the completion of the algorithm). 2) Satisfying the equal false-positive-rate condition
36 at every round (what we do in this paper). 3) Satisfying the equal false-positive-rate condition at an individual level (i.e.
37 over only the randomness of the classifier and not the population).

38 Prior work had only considered condition 3 (Joseph et al. [20]), and non-trivially satisfying this condition requires
39 realizability assumptions. Our paper asks whether we can achieve optimal regret bounds without any assumptions while
40 still satisfying 2. In particular, this means we are also the first paper to satisfy condition 1. Since we already obtain
41 the optimal \sqrt{T} regret bound satisfying the stronger condition 2, no asymptotic improvement would be possible if
42 we relaxed to condition 1. Condition 1 would also be meaningfully weaker, since it would offer no guarantees at any
43 particular time step: it would allow, for example, strong discrimination at every round, so long as the direction of that
44 discrimination varied with time.

45 Reviewer 2 asks if experiments would help establish if our method “ends up being fairer than some other standard
46 method.” We think the confusion here comes from the fact that we used asymptotic notation to state both the fairness
47 and regret guarantees for our algorithm. But our algorithm satisfies a hard fairness guarantee. We will clarify this, and
48 state our theorem using only asymptotic notation for the regret guarantee. (We also re-iterate that there are no existing
49 standard methods for the setting we consider).

50 Finally, sources of unfairness: This is a good discussion to add, thank you. As with all papers that take false positive
51 rates seriously, we implicitly assume that the labels we obtain are either correct, or have symmetric (“unbiased”) errors.
52 The sources of unfairness that we deal with relate to the differential ability of models in our class to fit both populations
53 (which we inherit from the batch setting), and the biased data collection inherent in online partial information settings.
54 We use EO only as a canonical example of a statistical fairness constraint, and don’t take the position that it is always
55 the right one. Our techniques also apply to other constraints like statistical parity.