Journal of Machine Learning Research 22 (2021) 1-35 Submitted 8/19; Revised 12/20; Published 3/21

Individual Fairness in Hindsight

Swati Gupta SWATIG@GATECH.EDU
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

Vijay Kamble KAMBLE@UIC.EDU
Department of Information and Decision Sciences
University of Illinois at Chicago
Chicago, IL 60607, USA

Editor: Moritz Hardt

Abstract

The pervasive prevalence of algorithmic decision-making in societal domains necessitates that these algorithms satisfy reasonable notions of fairness. One compelling notion is that of individual fairness (IF), which advocates that similar individuals should be treated similarly (Dwork et al. 2012). In this paper, we extend the notion of IF to online contextual decision-making in settings where there exists a common notion of conduciveness of decisions as perceived by the affected individuals. We introduce two definitions: (i) fairness-across-time (FT) and (ii) fairness-in-hindsight (FH). FT requires the treatment of individuals to be individually fair relative to the past as well as future, while FH only requires individual fairness of a decision *at the time of the decision*. We show that these two definitions can have drastically different implications when the principal needs to learn the utility model. Linear regret relative to optimal individually fair decisions is generally unavoidable under FT. On the other hand, we design a new algorithm: Cautious Fair Exploration (CaFE), which satisfies FH and achieves order-optimal sublinear regret guarantees for a broad range of settings.

Keywords: Individual Fairness, Online Learning, Contextual Bandits

1. Introduction

Algorithms facilitate decisions in increasingly critical aspects of modern life – ranging from search, social media, news, e-commerce, finance to determining credit-worthiness of consumers, estimating a felon's risk of reoffending, determining candidacy for clinical trials, etc. Their pervasive prevalence has motivated a large body of scientific literature in recent years that examines the effect of automated decisions on human well-being, and in particular, seeks to understand whether these effects are *fair* under various notions of fairness (Dwork et al. 2012, Sweeney 2013, Kleinberg et al. 2017, Angwin et al. 2016, Hardt et al. 2016, Chouldechova 2017, Chouldechova and G'Sell 2017, Corbett-Davies and Goel 2018).

In this context of automated decisions, fairness is often considered in a relative sense rather than an absolute sense. In his 1979 Tanner lectures, economist Amartya Sen eloquently argued that the heart of the issue rests on clarifying the "equality of what?" problem (Sen 2013). Equality can be desired with respect to opportunity (Hardt et al. 2016), outcomes (Phillips 2004), treatment (Dwork et al. 2012), or even mistreatment (Zafar et al. 2017). In this paper, we consider the notion of *individual fairness* (Dwork et al. 2012), that relies on the premise of equality of treatment, requiring that “similar” individuals must be treated “similarly”. This intuitively compelling notion of fairness was proposed in the influential work of Dwork et al. (2012) in the context of classification in supervised learning and has since been studied under several settings (see Yona and Rothblum 2018, Dwork and Ilvento 2018, Heidari and Krause 2018). The key idea is to introduce a Lipschitz condition on the decisions of a classifier, such that for any two individuals x, y that are at distance $d(x, y)$, the corresponding distributions over decisions $M(x)$ and $M(y)$ are also statistically close within a distance of some multiple of $d(x, y)$.

Individual fairness, as initially defined, is a static notion that pertains to offline or batch decisions. Because many algorithms for automated decision-making are sequential, in this paper, we propose an extension of individual fairness that explicitly accounts for the time at which decisions are made. We specifically focus on settings where there exists a common notion of conduciveness of decisions from the perspective of the individuals affected by these decisions; e.g., approval of a higher loan amount is more conducive to a loan applicant than the approval of a smaller amount, a shorter jail term is more conducive from the perspective of a convict, lower prices are more conducive for shoppers, etc.

As a motivating example, suppose two similar[1] persons A and B apply for a loan at a bank, and the bank approves a substantially higher loan amount to B than to A. This would be perceived as unfair (in the colloquial sense) by A, but maybe not by B. Under the classical definition of individual fairness, this distinction is irrelevant—implicitly, it is sufficient that either of the two similar individuals finds a drastically different treatment problematic, and hence the loan amounts approved for A and B must be similar.

The introduction of time allows for a richer treatment of the case above. For instance, if A applied for the loan earlier than B, then the treatment of A can still be defined to be fair as long as A got approved of a loan that is (approximately) at least as much as that approved for similar people *who applied before her*. In other words, A’s treatment by the bank can be deemed to be fair solely based on the history of decisions at the time when the loan was approved, despite the fact that this treatment turns out to be “individually unfair” in retrospect when B later gets approved for a substantially higher amount.

Armed with this basic intuition, we define *fairness-in-hindsight*: decisions are said to be fair in hindsight if the decisions for incoming individuals are individually fair relative to the past decisions for similar individuals, in the sense that they become more conducive over time by respecting a certain lower bound for rewards or an upper bound for penalties. To contrast this notion and to serve as a baseline, we also consider a more straightforward temporal extension of individual fairness, which we call fairness-across-time, where the treatments of individuals are required to be individually-fair relative to the past as well as the future. This means that similar individuals must always receive similar treatment irrespective of when they arrive: neither more conducive nor less conducive than what is justified by their degree of dissimilarity.

Our main technical contribution is to study the implications of these fairness constraints in situations where an algorithm operates under partial information and needs to learn good

[1] We discuss issues with defining such similarity in Section 2, but for now assume that this means two persons with exactly same observable attributes.

decisions over time. Formally, we consider a general online decision-making problem under uncertainty that falls into the class of stochastic contextual bandit problems that have been well-studied in literature (see, for example, Section 5 of Lattimore and Szepesvári (2020) for an overview). In this problem, individuals with observable contextual information (henceforth, we will refer to these as simply "contexts") arrive over time, and they have to be mapped to decisions by an algorithm. We assume that the contexts belong to a finite set. We further assume that the decisions are scalar and lie in a compact set identified with $[0, 1]$, e.g., a decision can represent the loan approved to a context or a price offered to a customer. The utility generated for each context and decision pair is random, whose distributional dependence on the pair is unknown to the algorithm; the specification of this dependence is henceforth referred to as the utility model. Our main assumption is that the uncertainty in the knowledge of the utility model is over a finite set of possible models. This restriction to a finite set allows a convenient abstraction to illustrate the key ideas in operationalizing fairness-in-hindsight. Continuous parameterizations of model uncertainty may have to be chosen dependent on the application in question and are interesting directions for future research; we discuss this in Section 6.

A standard performance metric in any such sequential decision-making settings featuring learning is the *regret* incurred by an *oblivious* algorithm that does not a priori know the utility model. This is defined as the difference between the optimal utility if the underlying utility model was known and the utility achieved under the algorithm. It is desirable that an algorithm ensures that the average asymptotic regret converges to 0, i.e., informally, it eventually learns the utility model and settles on optimal decisions.

The notion of treating similar individuals similarly, *always*, seems restrictive in such uncertain settings where it is a priori unclear which decisions are good. The large body of literature on sequential decision-making under uncertainty has shown that a certain degree of experimentation, at least in the early stages, could be fundamentally necessary to learn good decisions in the long run. But under the fairness-across-time constraint, bad decisions made in the early stages of experimentation may have to be repeated forever. This feature typically has dire consequences on regret. Under fairness-across-time, except for trivial settings, the expected regret against the benchmark of an optimal individually fair decision-rule grows linearly with the decision horizon. This is a fairly straightforward observation, illustrated in the example below, that is nevertheless worth formalizing.

Example 1 *Consider a bank making loan approval decisions in a new market that it has just entered. The decision space is $\mathcal{X} = [0, 1]$ representing the amount of loan sanctioned (normalized to 1). Consider a hypothetical setting where the only feature that the bank observes is the applicant's age, and it does not a priori know whether age is positively or negatively correlated with default probability. If the first applicant A is young, say 18 years old, and is given a loan of amount $\$M$, then any future applicant B aged 18 must be given $\$M$ to satisfy fairness-across-time. But this decision of $\$M$ loan is bound to be suboptimal and leads to a linear regret in either situation: (a) when M is small, and age is positively correlated with default probability, or (b) when M is large, and age is negatively correlated with default probability.*

In contrast, we demonstrate that the situation is not as pessimistic under fairness-in-hindsight: the possibility that decisions can become more conducive over time gives a powerful leeway that allows algorithms to learn *and* settle on good decisions over time. Formally, we design an algorithm that we call *Cautious Fair Exploration* (CaFE), which is individually fair in hindsight and attains sub-linear regret guarantees as compared to the optimal individually-fair benchmark in a wide range of settings.

CaFE operates in two phases, exploration and exploitation. In the exploration phase, the decisions are conservative, and the goal is to learn the utility model. Once the utility model is learned with the appropriate confidence, the algorithm enters the exploitation phase, when the decisions are then allowed to become more conducive for the appropriate individuals while ensuring individual fairness. The following example illustrates this point.

Example 2 *Consider Example 1. In this case, the bank can approve a small amount of loan, say ϵ, to each applicant in an initial exploration phase. Once the bank learns the correlation structure with appropriate confidence, it can start approving loans for larger amounts as appropriate in an individually fair manner while guaranteeing a loan of ϵ to everyone. This ensures fairness-in-hindsight.*

CaFE critically relies on the ability to learn with conservative decisions. But in many situations, learning is slow in the conservative regime. For instance, in the examples above, when the loan amount ϵ is small, the default probability is expected to be small irrespective of the individual's age. Hence, learning the correlation structure might take prohibitively long. On the other hand, if ϵ is large, then after having learned the model, the bank is forced to approve an amount of at least ϵ to individuals with a high default probability, leading to high regret. Thus, in these situations, there is a fundamental tradeoff between conservatism and the learning rate that is relevant to the overall regret incurred. Our technical results shed light on how decisions should be chosen in the exploration phase to balance this tradeoff to minimize regret. Our sublinear upper bounds on the regret of the resulting algorithm are accompanied by matching lower bounds that justify our design.

The paper is organized as follows. We build on the motivation for fairness-in-hindsight in the next section and survey relevant literature in Section 3. We then present the model and the fairness definitions in Section 4. We next present the learning problem in Section 5, where we present most of our main technical results. In Section 6, we discuss possible extensions and future directions. We conclude the paper in Section 7.

2. Motivation and Discussion

The notion of individual fairness over time is already ingrained in many of our societal systems, where it provides a dynamic, tangible frame of reference for our sense of social justice:

1. *Law:* In many legal systems, once a precedent is set by a ruling, decisions for similar cases observed in the future must follow these precedents. This is often referred to as *stare decisis*, i.e., to stand by things decided.[2] Further, rulings at any point in time affect future decisions by setting a precedent (prospective application of the law). In contrast, they seldom change the decisions of past rulings (retrospective application of the law)(Friedland

[2] See `https://www.law.cornell.edu/wex/stare_decisis`.

1974). Therefore, the effect of a ruling is not symmetric with respect to past and future decisions, thus bearing similarities to our unidirectional notion of fairness-in-hindsight.

2. *Whataboutism:* There has been a growing culture of *whataboutism* in society, which is an argumentative device used to prove mistreatment by pointing to another similar context that received a different, more conducive, treatment in the past. For example, politicians might defend their questionable actions or advocate no-penalty since a prior political scandal of an opposition party did not go through due process or face significant consequences.[3]

3. *Pricing:* Bolton et al. (2003) find that customers' impression of fairness of prices critically relies on past prices that act as reference points. In particular, they find that any price increases beyond that justified by an increase in perceived costs, e.g., due to inflationary effects, are perceived as price gouging and unfair by the customers, and moreover, to make matters worse for the firm, customers tend to underestimate these costs.

Given the prevalence of this notion in practice, it is natural to expect societal algorithms to uphold these same principles. An immediate question that arises then is: what are the implications of these constraints on the long-run performance of algorithms that typically operate under system uncertainty? As we show through CaFE, unlike the more stringent notion of fairness-across-time under which high regret is usually inevitable, fairness-in-hindsight does allow us to learn and settle on good decisions over time. The conservative exploration then exploitation structure of CaFE is intuitive and bears a resemblance to certain features seen in existing societal systems. For instance, it is typical for legal stances on new issues to be more conservative initially and then potentially become more liberal over time as the impact and nuances of these issues become clear, e.g., decriminalization laws remove penalties for actions perceived as crimes in the past. Another example is that of markdowns in retail, where the prices of new goods that did not see anticipated demand are decreased over time.

The issue of the metric. Dwork et al. (2012) admit that the existence and availability of a similarity metric between individuals for a particular decision-making problem is one of the most challenging aspects of the notion of individual fairness.

> *Our approach is centered around the notion of a task-specific similarity metric describing the extent to which pairs of individuals should be regarded as similar for the classification task at hand. The similarity metric expresses ground truth. When ground truth is unavailable, the metric may reflect the "best" available approximation as agreed upon by society. Following established tradition (Rawls 2001) the metric is assumed to be public and open to discussion and continual refinement. Indeed, we envision that, typically, the distance metric would be externally imposed, for example, by a regulatory body or externally proposed by a civil rights organization.* (Dwork et al. 2012)

[3] A common argument used by the defenders of the republican US President Richard Nixon's administration after the Watergate scandal in the United States was to point to an older unfortunate case of the Chappaquiddick accident that democrat Ted Kennedy was implicated in; see `https://bit.ly/2ORRj8L`. A similar instance is the "What about Hillary?" rhetoric used by supporters of US President Donald Trump; see `https://bit.ly/2GHRCC8`.

In the context of automated decisions, individuals typically appear as a vector of attributes to an algorithm. However, several challenges arise in defining a metric: it is a priori unclear which attributes of an individual should be considered to be relevant to a decision-making task from a fairness perspective, what is the relative importance of these attributes, which attributes must be ignored completely, etc. This choice is especially non-trivial because there could be seemingly non-controversial attributes, e.g., a preferred genre of music, education, zip code, etc., that are correlated with membership in protected population subgroups, which could be the basis for disparate treatment under an individually fair algorithm (Pedreshi et al. 2008).

We acknowledge that our proposal of fairness-in-hindsight inherits these concerns. However, recent works have offered compelling resolutions that attempt to achieve individual fairness in algorithmic decisions under weaker assumptions on the knowledge of the metric. For example, Jung et al. (2019) assumes access to a sample of pairs of individuals that must be treated approximately equally rather than having access to the similarity metric. Along similar lines, Gillen et al. (2018) and Bechavod et al. (2020) assume that only noisy feedback about the similarity metric is available to the algorithm via an auditor who detects fairness violations. We are optimistic that techniques in these works can be employed in conjunction with CaFE to achieve fairness-in-hindsight when the metric is not precisely known.

Time as a feature. Individual fairness advocates similar treatment of similar individuals. A natural question is whether the time at which a decision is taken can be incorporated as a feature, allowing increasingly different treatment of similar individuals as more time passes. This will make learning in the online framework easier since it will allow more flexibility of recourse in decisions. Our proposal of fairness-in-hindsight does not resort to such a technical workaround. It stems from the belief that *time is inherently not a valid basis for allowing discriminatory decisions* (in the sense of individual fairness). Such incorporation of time would allow, for example, giving conducive treatment to an individual from group A and then later giving an unfavorable treatment to a similar individual from group B, justified simply by the passage of time. Fairness-in-hindsight, on the other hand, guarantees that the treatment of a person is individually fair *at the time of the decision.*

3. Related Literature

Early research on fairness in machine learning focused on the offline setting of batch supervised learning from observational data (Pedreshi et al. 2008, Kamiran and Calders 2009, Calders and Verwer 2010, Kamishima et al. 2011, Dwork et al. 2012, Hardt et al. 2016, Kleinberg et al. 2017). Only recently has the literature started looking at the implications of fairness constraints in online learning settings. Our model and results complement a small but growing body of literature in this domain (Joseph et al. 2016b, Liu et al. 2017, Gillen et al. 2018, Heidari and Krause 2018, Celis et al. 2018, Joseph et al. 2016a, Elzayn et al. 2019). The two papers most related to our work are Joseph et al. (2016b) and Heidari and Krause (2018) that we discuss below.

Joseph et al. (2016b) pioneered the study of the impact of fairness constraints on learning in a contextual multi-armed bandit setting under a utility maximization objective. They propose a *meritocratic* notion of fairness so that with high probability over the entire decision horizon, the probability of picking an arm (i.e., a subgroup) at any time is monotonic in its underlying mean reward. Their notion of fairness is, however, limited to individuals that appear within each time period. In contrast, our notion of fairness is defined relative to all those who arrived in the past.

In a recent work of Heidari and Krause (2018), the authors extend individual fairness to account for the notion of time and study its impact on learning. They consider an online supervised learning problem from a class of model hypotheses under the probably-approximately-correct (PAC) learning framework and propose that decisions for individuals that arrive within a fixed number of time periods (say M) of each other must satisfy individual fairness. They design an algorithm that is asymptotically consistent, i.e., it learns and settles on the true hypothesis with high probability. In contrast, we propose fairness-in-hindsight by only requiring lower bounds on present decisions to satisfy individual fairness relative to *all* past decisions, motivated by the notion of conduciveness of decisions. We do not require an exogenous choice of the window of time periods (i.e., M) across which individual fairness must hold. Moreover, our focus is on learning for utility maximization as opposed to pure learning.

Several other works have recently appeared in this domain. Jabbari et al. (2017) extend the notion of fairness in Joseph et al. (2016b) to the setting of reinforcement learning. Liu et al. (2017) and Gillen et al. (2018) study the model of Joseph et al. (2016b) under different notions of individual fairness. Liu et al. (2017) require similar probabilities of picking two arms whose quality distribution is similar. They study calibration under this requirement since the definition is not restrictive enough, e.g., it does not require that these probabilities are monotonic in the average quality. Gillen et al. (2018) require that similar individuals must face a similar probability of being chosen by the algorithm, except that only noisy feedback about the distance metric between individuals is available. They study the problem of regret minimization compared to optimal individually fair policy relative to the true metric. Celis et al. (2018) consider a contextual bandit problem arising in personalization and address the problem of ensuring another notion of fairness called *group fairness*[4] across time. In all of these works, the settings, the models, and the fairness constraints are different from those we consider in the present work.

4. Static and Dynamic Models

We now define the models of contextual decision-making that we will focus on in the rest of the paper. The first model is static and does not feature time, while the second is a dynamic model explicitly incorporating time.

4.1 The static model

Consider a principal responsible for mapping contexts $c \in \mathcal{C}$ to scalar decisions $x \in \mathcal{X} = [0, 1]$, where $\mathcal{C}$ is a finite set with $|\mathcal{C}| = C$. We assume that the contexts are drawn from some distribution $\mathcal{D}$ over $\mathcal{C}$. For a context c and decision x, the principal observes a

[4] Group fairness tries to address the issue of *disparate impact* in automated decisions: which refers to practices that collectively allocate a more favorable outcome to one population subgroup compared to another. An algorithm satisfies group fairness, aka statistical parity, if its decisions are independent of membership in any subgroup.

random utility U drawn from some distribution $\mathcal{F}(x,c)$ over $\mathbb{R}$ and we will often work with the corresponding expected utility $\bar{u}(x,c) \triangleq \mathbb{E}_{\mathcal{F}(x,c)}(U)$. We assume that $\bar{u}(x,c)$ is continuous and concave in $x \in \mathcal{X}$ for each $c \in \mathcal{C}$. Continuity implies that that $\bar{u}(x,c)$ is uniformly bounded, i.e., $\max_{c\in\mathcal{C},x\in\mathcal{X}} |\bar{u}(x,c)| \leq B < \infty$. We first consider the case when the distribution $\mathcal{F}$ is known to the principal; i.e., no learning is required and we will later consider the case where this distribution needs to be learned in Section 5.

Example 3 *Suppose that the principal is a bank that is making loan approval decisions. The decision space is $\mathcal{X} = [0,1]$ representing the amount of loan sanctioned (normalized to 1). The probability of loan default depends on both the loan amount x and the type c of the applicant belonging to the finite set of types $\mathcal{C}$. Suppose that for a type c and a loan amount x, the probability of loan default is estimated to be $p(x,c)$. For a decision, x, the utility of the bank is $-x$ if there is a default and it is βx (the net present value of the interest) if there is no default, i.e.,*

$$U = \begin{cases} -x & w.p.\ p(x,c) \\ \beta x & w.p.\ 1-p(x,c). \end{cases}$$

Then, the expected utility is $\bar{u}(x,c) = -xp(x,c) + \beta x(1-p(x,c)) = x(\beta - p(x,c)(1+\beta))$.

Suppose that for any two contexts in $\mathcal{C}$, there exists a commonly agreed-upon distance between them as defined by a function $d_{\mathcal{C}} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}^+$. We assume that this function defines a metric on $\mathcal{C}$; in particular, it is non-negative, satisfies the triangle inequality, and the distance of a context to itself is zero. Consider the following definition of an *individually fair* decision-rule in the spirit of Dwork et al. (2012).

Definition 1 *(Dwork et al. 2012) A decision-rule ϕ is K-Lipschitz for $K \in [0,\infty)$ if*

$$|\phi(c) - \phi(c')| \leq K d_{\mathcal{C}}(c,c') \text{ for all } c,\, c' \in \mathcal{C}. \tag{1}$$

Let $\Phi_K : \mathcal{C} \to \mathcal{X}$ be the space of K-Lipschitz decision-rules that map contexts to decisions. The optimization problem of the principal is to choose a K-Lipschitz decision-rule that maximizes the expected utility. We define the maximum expected utility over K-Lipschitz decision-rules as:

$$U_K \triangleq \max_{\phi\in\Phi_K} \mathbb{E}_{\mathcal{D}}[\bar{u}(\phi(c),c)]. \tag{2}$$

Given the concavity of $\bar{u}(x,c)$ in $x \in \mathcal{X}$ for each $c \in \mathcal{C}$, this problem can be solved as a finite convex program, since $\mathcal{C}$ is assumed to be finite.

4.2 The dynamic model

Consider now a discrete time dynamic setting where time is denoted as $t = 1, \cdots, T$ and contexts $c_t \in \mathcal{C}$ are drawn i.i.d. from the distribution $\mathcal{D}$ over $\mathcal{C}$ at each time period. The principal makes a decision $x_t \in \mathcal{X} = [0,1]$, using a *policy* ψ that maps the sequence of contexts seen up to time t, the corresponding decisions up to time $t-1$, and the utility outcomes up to time $t-1$ to the decision x_t (for all $t \geq 1$). Note that a policy is distinct from a decision-rule: a decision-rule is a *static* object that maps every possible context to a decision, whereas a policy adaptively maps contexts to decisions as it encounters them, possibly mapping the same context to different decisions across time. As in the static case, for context c_t and decision x_t, the principal obtains a random utility U_t, drawn from the distribution $\mathcal{F}(x_t, c_t)$ independently of the past. Given the contexts observed and decisions taken, the expected utility of the principal until time T is given by $\sum_{t=1}^{T} \mathbb{E}[U_t] = \sum_{t=1}^{T} \bar{u}(x_t, c_t)$. We consider the following two definitions of fairness of policies.

Definition 2 (Fairness-across-time) *We say that a policy is fair-across-time (FT) with respect to the function $\mathcal{K}(s) : \mathbb{N} \to \mathbb{R}^+$ if the decisions it generates for any sequence of contexts satisfy,*

$$|x_t - x_{t'}| \le \mathcal{K}(|t' - t|) d_{\mathcal{C}}(c_t, c_{t'}) \text{ for all } t' \neq t. \tag{3}$$

When $\mathcal{K}(s) = K$ for some $K \in [0, \infty)$, we say that the policy is K-fair-across-time (K-FT).

Note that by setting $\mathcal{K}(\cdot)$ to be a monotone increasing function, one can model the scenario where the past decisions have a diminishing impact on the future decisions dependent on the amount of time passed. However, even if $\mathcal{K}$ is monotone increasing, fairness-across-time requires that any particular context must be mapped to the same decision irrespective of when it arrives in time (and this hinders learnability as we discuss further in Section 5).

Definition 3 (Fairness-in-hindsight) *We say that a policy is fair-in-hindsight (FH) with respect to the function $\mathcal{K}(s) : \mathbb{N} \to \mathbb{R}^+$ if the decisions it generates for any sequence of contexts satisfy,*

$$x_t \ge x_{t'} - \mathcal{K}(t - t') d_{\mathcal{C}}(c_t, c_{t'}) \text{ for all } t \ge t'. \tag{4}$$

When $\mathcal{K}(s) = K$ for some $K \in [0, \infty)$, we say that the policy is K-fair-in-hindsight (K-FH).

Note that fair-in-hindsight policies must make monotone (non-decreasing) decisions over time for any given context, assuming that higher decisions are more conducive[5]. In fact, setting $\mathcal{K}(s) = 0$ for all s, we recover policies that make monotone decisions over time irrespective of the context.

Let $\Psi^T_{K\text{-FT}}$ and $\Psi^T_{K\text{-FH}}$ be the space of T-horizon policies that are K-FT and K-FH respectively. Let the maximum attainable expected utility up to time T using K-FT and K-FH policies be $U^T_{K\text{-FT}}$ and $U^T_{K\text{-FH}}$:

$$U^T_{K\text{-FT}} := \max_{\psi \in \Psi^T_{K\text{-FT}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{D}}[\bar{u}(x_t, c_t)], \qquad U^T_{K\text{-FH}} := \max_{\psi \in \Psi^T_{K\text{-FH}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{D}}[\bar{u}(x_t, c_t)]. \tag{5}$$

It is clear that both $U^T_{K\text{-FT}} \ge TU_K$ and $U^T_{K\text{-FH}} \ge TU_K$, since one can simply use the optimal K-Lipschitz decision-rule at every stage. But for small horizons, one can potentially do better. Intuitively, this is because you may not expect to encounter all the contexts within a short horizon; hence the fairness constraints are expected to be less constraining, thus

[5] If lower decisions are more conducive from the perspective of an individual (e.g., x is the amount of penalty), then we can transform the decision space by mapping x to $1-x$ so that the higher transformed decisions are more conducive.

offering more flexibility in mapping contexts to decisions. For the interested reader, we present an example below to show that one can attain a higher utility under FH than under FT, and both these notions can lead to a higher utility that following the static optimal K-Lipschitz decision-rule when the horizon is small.

Example 4 *Let $T = 2$. Suppose there are two contexts: A, B where A is seen with probability 1/12 and B is seen with probability 11/12. Let the expected utility be $\bar{u}(A, x) = -x$ (defaulters) and $\bar{u}(B, x) = 1.2x$ (non-defaulters) for decisions $x \in [0, 1]$. Note that optimum unconstrained decision for A is 0 and optimum unconstrained decision for B is 1. Let $d(A, B) = 1$ and $K = 0.5$. In the optimal K-Lipschitz decision rule, the decision for A is 0.5 and the decision for B is 1.*

Now any K-FT policy ψ must give loans x_1 to A and x_2 to B such that $|x_1 - x_2| \leq 0.5$ irrespective of the time periods they arrive in. Any K-FH policy ψ' must ensure that if loan x_1 was given to context c_1 at $t = 1$, then at least the same amount of loan must be given to the same context, and at least $x_1 - 0.5$ must be given to the other context at $t = 2$, dependent on which context arrives. Suppose now that at $t = 1$, A arrives. Then, one can verify that the optimal K-FT policy must give a loan of 0.5 in anticipation of B in the next time step (so that a loan of 1 can be given to B), whereas the optimal K-FH policy can give a loan of 0 to A at $t = 1$. If A arrives at $t = 2$, then K-FH can still give a loan of 0; if B arrives, then it can give a loan of 1. Thus one can attain a higher utility under FH than under FT.

To see that both the notions FT and FH can lead to a higher utility than that under the static optimal K-Lipschitz decision-rule, observe that if there is only a single stage, i.e., $T = 1$, then the decision for A can be 0 under any FT or FH policy, whereas the optimal K-Lipschitz decision-rule must choose 0.5.

We can show, however, that when the horizon gets longer, one cannot do any better than achieving the static optimum average expected utility U_K (as defined in (8)).

Proposition 4 *For any $K \in [0, \infty)$, we have $U^T_{K\text{-FT}} \leq TU_K + 2BC$ and $U^T_{K\text{-FH}} \leq TU_K + 2BC$ where B is the upper bound on the possible expected utility, i.e., $\max_{c \in C,\, x \in X} |\bar{u}(x, c)| \leq B < \infty$, and $C = |\mathcal{C}|$. Hence,*

$$\lim_{T\to\infty} \frac{U^T_{K\text{-FT}}}{T} = \lim_{T\to\infty} \frac{U^T_{K\text{-FH}}}{T} = U_K.$$

Proof Note that $U^T_{K\text{-FT}} \leq U^T_{K\text{-FH}}$ since FT implies FH. Hence, we show the result only for $U^T_{K\text{-FH}}$. The corresponding result for $U^T_{K\text{-FT}}$ follows.

Fix an FH policy. At any given time t, let $\ell_t(c)$ be the tightest lower bound on the decision for c, for each $c \in \mathcal{C}$, based on decisions taken in the past. Note that if a decision x has been taken for a context c at any time before t, then $\ell_t(c) \geq x$ due to the FH constraint.

First, we show that ℓ_t specifies a K-Lipschitz decision-rule. To see this, consider two contexts c and c', and w.l.o.g., assume that $\ell_t(c) \geq \ell_t(c')$. If $\ell_t(c) = 0$, then clearly $\ell_t(c) = \ell_t(c') = 0$. Next, if for some time $t' < t$, the context c was mapped to decision $\ell_t(c)$, then from the FH constraint, it follows that $\ell_t(c') \geq \ell_t(c) - Kd_{\mathcal{C}}(c, c')$. Thus $|\ell_t(c) - \ell_t(c')| \leq Kd_{\mathcal{C}}(c, c')$. Finally, suppose that either the context c had never appeared before time t, or it had appeared and the highest decision taken for this context so far is some $x < \ell_t(c)$

(note again that the highest decision in the past for context c cannot be larger than $\ell_t(c)$). In this case, there is some other context c^* that was mapped to some decision x^* at some time in the past and $\ell_t(c) = x^* - Kd_{\mathcal{C}}(c^*, c)$ (since $\ell_t(c)$ is the tightest lower bound). But this also means that $\ell_t(c') \geq x^* - Kd_{\mathcal{C}}(c^*, c')$. Thus $\ell_t(c') - \ell_t(c) \geq K(d_{\mathcal{C}}(c^*, c) - d_{\mathcal{C}}(c^*, c'))$. But by the triangle inequality, we have $d_{\mathcal{C}}(c^*, c') \leq d_{\mathcal{C}}(c^*, c) + d_{\mathcal{C}}(c, c')$. Thus we have $\ell_t(c') \geq \ell_t(c) - Kd_{\mathcal{C}}(c, c')$. Thus again, $|\ell_t(c) - \ell_t(c')| \leq Kd_{\mathcal{C}}(c, c')$. This shows that ℓ_t is K-Lipschitz.

Now consider the decision rule ϕ_t chosen by the policy at time t (this depends on the history of actions and observations in the past). We will show that the total incremental utility obtained by the policy over time by replacing ℓ_t at each time t by ϕ_t is bounded by a constant independent of T.

To see this, let t_1^c, t_2^c, $\cdots$, $t_{F_c}^c$, be the (random) times until time T when the context that arrives is c, assuming that $F_c \geq 1$. Then the total additional utility obtained by the principal across all times by replacing ℓ_t at each time t by ϕ_t can be written as,

$$\sum_{t=1}^{T} \bar{u}(\phi_t(c_t), c_t) - \bar{u}(\ell_t(c_t), c_t) = \sum_{c \in \mathcal{C}} \mathbb{1}_{\{F_c \geq 1\}} \sum_{k=1}^{F_c} \bar{u}(\phi_{t_k^c}(c), c) - \bar{u}(\ell_{t_k^c}(c), c), \tag{6}$$

where c_t is the context that arrives at time t. However, due to the FH constraint, if $F_c \geq 1$, then $\phi_{t_k^c}(c) \geq \ell_{t_k^c}(c)$, and if $F_c \geq 2$ then $\ell_{t_k^c}(c) \geq \phi_{t_{k-1}^c}(c)$ for $F_c \geq k \geq 2$. Thus, the intervals $[\ell_{t_k^c}(c), \phi_{t_k^c}(c))$ for $1 \leq k \leq F_c$ are non-overlapping subsets of $[0, 1]$. Hence, we have

$$\sum_{k=1}^{F_c} \bar{u}(\phi_{t_k^c}(c), c) - \bar{u}(\ell_{t_k^c}(c), c) \leq 2B. \tag{7}$$

This inequality follows from concavity of $\bar{u}(x, c)$ in $x \in [0, 1]$ for each $c \in \mathcal{C}$ and the fact that $\max_{c \in C, x \in X} |\bar{u}(x, c)| \leq B$. Thus, to conclude, the total additional utility obtained by the principal by deviating from ℓ_t in each time period is at most $2B|\mathcal{C}|$. Since ℓ_t is a K-Lipschitz decision-rule, the expected utility under this decision-rule is at most U_K. Hence, we have an upper bound on the expected utility under any policy equal to $TU_K + 2BC$. ■

This, in particular, shows that relaxing the fairness-across-time constraint to only requiring fairness-in-hindsight does not lead to any long-run gains in objective. The policy of simply choosing the optimal static K-Lipschitz decision-rule at every stage is approximately optimal for a large horizon T. Next, we show that the situation is drastically different when there is learning involved.

5. Dynamic Model with Learning

Consider now a setting where the distribution of the utility given a context and a decision is unknown to the principal and must be learned. Formally, this distribution depends on an additional unknown parameter w, which we assume to belong to a finite set $\mathcal{W}$. With some abuse of notation, for each $w \in \mathcal{W}$, $c \in \mathcal{C}$ and $x \in \mathcal{X}$, the distribution of the utility of the principal is given by $\mathcal{F}(x, c, w)$. We assume that this distribution has a finite support $\mathcal{U}(x, c)$, i.e. $\max_{x,c} |\mathcal{U}(x, c)| < \infty$,[6] and for each $u \in \mathcal{U}(x, c)$, the probability of observing

[6] This is satisfied in Example 4.1 where there are only two possibilities: either the person defaults on the loan or doesn't.

u is given by $p(u \mid x, c, w)$. We assume that $p(u \mid x, c, w) > 0$, for each $u \in \mathcal{U}(x, c)$ for all $c \in \mathcal{C}$, $w \in \mathcal{W}$, and for all x in the interior of $\mathcal{X}$, i.e., $x \in (0, 1)$. Finally, we assume that the principal knows possible feasible parameters $\mathcal{W}$ but does not know the true parameter w, which must be learned by adaptively assigning decisions to contexts and observing the outcomes.

We now redefine some previously defined quantities (with some abuse of notation) to capture the dependence on the parameter w. First, we define $\bar{u}(x, c, w) \triangleq \mathbb{E}_{\mathcal{F}(x,c,w)}(U)$, i.e., the mean utility for a given $x \in \mathcal{X}$, $c \in \mathcal{C}$ and $w \in \mathcal{W}$. As before, we assume that $\bar{u}(x, c, w)$ is continuous and concave in $x \in \mathcal{X}$ for each $c \in \mathcal{C}$ and $w \in \mathcal{W}$. Continuity implies that $\bar{u}(x, c, w)$ is uniformly bounded, i.e., $\max_{c \in \mathcal{C},\, x \in \mathcal{X},\, w \in \mathcal{W}} |\bar{u}(x, c, w)| \le B < \infty$. Next, we define $U_K(w)$ to be the highest expected utility attainable under a K-Lipschitz decision rule, for a given parameter w, i.e.,

$$U_K(w) \triangleq \max_{\phi \in \Phi_K} \mathbb{E}_{\mathcal{D}}[\bar{u}(\phi(c), c, w)]. \tag{8}$$

Let ϕ^*_w denote the optimal K-Lipschitz decision rule that attains this maximum.

For a given horizon T, for any dynamic policy in $\Psi^T_{K\text{-FT}}$ or $\Psi^T_{K\text{-FH}}$ that is oblivious of w, we can define a notion of *regret* that compares its expected utility against the long-run optimal benchmark $TU_K(w)$. For any policy $\psi \in \Psi^T_{K\text{-FT}}$, for a fixed w, we denote its total utility at the end of the horizon as $U^T_{K\text{-FT}}(w, \psi)$ (similarly, $U^T_{K\text{-FH}}(w, \psi)$). Then, for any $\psi \in \Psi^T_{K\text{-FT}}$, let the regret be denoted as:

$$\text{Regret}^T_{K\text{-FT}}(w, \psi) \triangleq TU_K(w) - U^T_{K\text{-FT}}(w, \psi), \tag{9}$$

and similarly, for any $\psi \in \Psi^T_{K\text{-FH}}$, we define,

$$\text{Regret}^T_{K\text{-FH}}(w, \psi) \triangleq TU_K(w) - U^T_{K\text{-FH}}(w, \psi). \tag{10}$$

5.1 Learning under FT

We show that under the FT constraint, except for trivial settings, a regret that asymptotically grows linearly in T is unavoidable. The reason is that once a context is mapped to a decision, we are forced to map that context to the same decision forever under FT. We can show that with some positive probability, a bad decision in the first step for some w is inevitable, which then must be repeated forever, thus incurring linear regret. This intuition is illustrated in Example 1.1 in Section 1.

Proposition 5 *Suppose there is some pair $w', w'' \in \mathcal{W}$ such that a) $\phi^*_{w'}$ and $\phi^*_{w''}$ are the unique optimal (static) K-Lipschitz decision-rules for w' and w'' respectively,*[7] *and b) $\phi^*_{w'} \neq \phi^*_{w''}$. Then there exists an instance dependent constant $\kappa > 0$ and a time $T' \ge 1$ such that for any $T \ge T'$ and any K-*FT *policy $\psi \in \Psi^T_{K\text{-FT}}$,*

$$\max_{w \in \mathcal{W}} \text{Regret}^T_{K\text{-FT}}(w, \psi) \ge \kappa T.$$

[7] Note that the set of K-Lipschitz decision rules given a finite set of contexts is convex. By adding a small random perturbation to the expected utility function, we can ensure that the optimal decision rules are unique for each $w \in \mathcal{W}$.

Proof Fix some $T \geq 1$ and consider a K-FT policy $\psi \in \Psi^T_{K\text{-FT}}$. Since $\phi^*_{w'} \neq \phi^*_{w''}$, there exists some $c' \in \mathcal{C}$ such that $\phi^*_{w'}(c') \neq \phi^*_{w''}(c')$. Let ϕ^1 be the decision-rule followed by policy ψ at time 1, and let $\phi^1(c') = x'$. Consider a K-FT policy ψ^{o}, which knows w and maximizes the T-period utility for each w, except that it is constrained to use the decision rule ϕ^1 at time 1 irrespective of w, i.e., $x_1 = \phi^1(c_1)$ under ψ^{o}. It is then clear that for each $w \in \mathcal{W}$,

$$\text{Regret}^T_{K\text{-FT}}(w, \psi) \geq \text{Regret}^T_{K\text{-FT}}(w, \psi^{\text{o}}). \tag{11}$$

Consider now the event that the context seen at time 1 is c', i.e., event $\{c_1 = c'\}$ happens. On this event, we can show that the policy ψ^{o} can achieve at most $\delta_T = 2B(C+1)/T$ more than $U'(w, x')$ on an average, where

$$U'(w, x') = \max_{\phi \in \Phi_K;\, \phi(c') = x'} \mathbb{E}_{\mathcal{D}}[\bar{u}(\phi(c), c, w)]. \tag{12}$$

To see this, suppose that $\ell_t(c)$ is the lower bound on the decisions under the policy ψ^{o} at time $t \geq 2$ due to the FT constraint. We know that $\ell_t(c') = x'$. Also, by similar arguments as that in the proof of Proposition 4, ℓ_t can be argued to be a K-Lipschitz decision rule. Thus ℓ_t is feasible in problem (12) and hence the expected utility under this rule is at most $U'(w, x')$. Now consider the decision rule ϕ_t chosen by the policy ψ^{o} at time $t \geq 2$. We will show that the total incremental utility obtained by the policy over time by replacing ℓ_t at each time t by ϕ_t is bounded by a constant independent of T.

To see this, let t^c_1, t^c_2, $\cdots$, $t^c_{F_c}$, be the (random) times after time $t = 1$ and until time T, when the context that arrives is c, assuming that $F_c \geq 1$. Then the total additional utility obtained by the principal across all times by replacing ℓ_t at each time t by ϕ_t can be written as,

$$\sum_{t=2}^{T} \bar{u}(\phi_t(c_t), c_t, w) - \bar{u}(\ell_t(c_t), c_t, w) = \sum_{c \in \mathcal{C}} \mathbb{1}_{\{F_c \geq 1\}} \sum_{k=1}^{F_c} \bar{u}(\phi_{t^c_k}(c), c, w) - \bar{u}(\ell_{t^c_k}(c), c, w). \tag{13}$$

However, due to the FT constraint, if $F_c \geq 2$, then for any $F_c \geq k \geq 2$, we have that $\phi_{t^c_k}(c) = \ell_{t^c_k}(c)$ and $\ell_{t^c_k}(c) = \phi_{t^c_{k-1}}(c)$. Hence, if $F_c \geq 1$, then we have

$$\sum_{k=1}^{F_c} \bar{u}(\phi_{t^c_k}(c), c, w) - \bar{u}(\ell_{t^c_k}(c), c, w) = \bar{u}(\phi_{t^c_1}(c), c, w) - \bar{u}(\ell_{t^c_1}(c), c, w) \leq 2B. \tag{14}$$

This inequality follows from the fact that $\max_{c \in C,\, w \in \mathcal{W},\, x \in X} |\bar{u}(x, c, w)| \leq B$. Thus, to conclude, the total additional utility obtained by the principal by deviating from ℓ_t in each time period $t \geq 2$ is at most $2B|\mathcal{C}|$. Since ℓ_t is feasible in problem (12), the expected utility under this decision-rule is at most $U'(w, x')$. Moreover, the utility earned by the policy ψ^{o} at time $t = 1$ is at most B. Hence, we have an upper bound on the total expected utility of ψ^{o} on the event $\{c_1 = c'\}$, equal to $B + (T-1)U'(w, x') + 2BC$, which is at most $TU'(w, x') + 2B(C+1)$.

Next, we have,

$$\begin{aligned}
\max_{w \in \{w', w''\}} \text{Regret}^T_{K\text{-FT}}(w, \psi^{\text{o}}) &\geq \mathbb{P}(c_1 = c')\Big[\max_{w \in \{w', w''\}} T\Big(U_K(w) - U'(w, x') - \delta_T\Big)\Big] \\
&= \mathbb{P}(c_1 = c')\Big[\max_{w \in \{w', w''\}} T\Big(U_K(w) - U'(w, x')\Big) - 2B(C+1)\Big].
\end{aligned}$$

Clearly, $U_K(w) \geq U'(w, x')$ for any w, and hence $\max_{w \in \{w', w''\}}(U_K(w) - U'(w, x')) \geq 0$. If $\max_{w \in \{w', w''\}}(U_K(w) - U'(w, x')) = 0$, then this implies that $U_K(w) = U'(w, x')$ for both $w = w'$ and $w = w''$. But this implies the existence of K-Lipschitz decision-rules for w' and w'' that are optimal, and such that they both map c' to x'. This contradicts the fact that $\phi^*_{w'}$ and $\phi^*_{w''}$ are the unique optimal K-Lipschitz decision-rules, since they map c' to different decisions. Thus $\max_{w \in \{w', w''\}} U_K(w) - U'(w, x') > 0$ for each $x' \in \mathcal{X}$. This in turn implies that $\min_{x' \in \mathcal{X}} \max_{w \in \{w', w''\}} U_K(w) - U'(w, x') \triangleq \gamma' > 0$ (the minimum is well-defined since $U'(w, x')$ is a continuous function of x' for each w).

Hence, $\max_{w \in \{w', w''\}} \text{Regret}^T_{K\text{-FT}}(w, \psi^{\text{o}}) \geq \mathbb{P}(c_1 = c')[\gamma' T - 2B(C+1)]$. Using (11), we get the result. ■

This shows that fairness-across-time can result in significant losses in utility relative to the static optimum when learning is involved.

5.2 Learning under FH

The situation is not as bleak under the FH constraint as we now show. The problem with the FT constraint is that one is forced to repeat mistakes by mapping each context to the same decision that was made when the context was first encountered. The FH constraint allows for some flexibility: the decisions for every context can potentially *increase* over time. This presents the following possibility: one can try to *cautiously* learn w with small and equal decisions for all contexts, and then increase decisions for contexts as needed once w is learned with appropriate confidence. This was illustrated in Example 1.2 in Section 1. Formally, we propose an online learning algorithm that we call *Cautious Fair Exploration* or CaFE in Algorithm 1.

In order to describe our performance bound for CaFE, we first need to define a few quantities. For any w, $w' \in \mathcal{W}$, $x \in (0,1)$ and $c \in \mathcal{C}$, we define

$$\text{KL}(w, w'|x, c) \triangleq \sum_{u \in \mathcal{U}(x,c)} p(u \mid x, c, w) \log \frac{p(u \mid x, c, w)}{p(u \mid x, c, w')}. \tag{15}$$

This quantity is commonly known as the Kullback-Leibler (KL) divergence between the distributions $\mathcal{F}(x, c, w)$ and $\mathcal{F}(x, c, w')$.

Also, recall that $c \sim \mathcal{D}$ and for each $x \in (0,1)$, define:

$$L(x) \triangleq \min_{w, w'} \mathbb{E}_{\mathcal{D}}[\text{KL}(w, w'|x, c)]. \tag{16}$$

$L(x)$ captures how well we can distinguish between the different parameter values given a decision x: a low value of $L(x)$ implies that there is a pair $(w,\, w')$ that is difficult to distinguish at decision x. We also define for each $x \in \mathcal{X}$:

$$D(x) \triangleq \min_{c \in \mathcal{C},\, w \in \mathcal{W},\, u \in \mathcal{U}(x,c)} p(u \mid x, c, w). \tag{17}$$

$D(x)$ captures the smallest probability assigned to a utility value over all possibilities for w and c, as a function of the decision x. Finally, we make the following assumption.

Assumption 1 *$\bar{u}(x, c, w)$ is Lipschitz continuous on $\mathcal{X}$ for each $c \in \mathcal{C}$ and $w \in \mathcal{W}$, with a Lipschitz constant $R > 0$, i.e.,*

$$|\bar{u}(x, c, w) - \bar{u}(x', c, w)| \leq R|x - x'|,$$

for all x, $x' \in \mathcal{X}$, $c \in \mathcal{C}$, and $w \in \mathcal{W}$.

This assumption simply ensures that the loss in utility due to a small deviation from the optimal decision for a particular context isn't arbitrarily high. Without this requirement, one cannot hope to obtain non-trivial regret rates in general.[8]

Algorithm 1: Cautious Fair Exploration (CaFE)

Input: $T \in \mathbb{N}$, $(\mathcal{F}(x, c, w);\ c \in \mathcal{C},\ x \in \mathcal{X},\ w \in \mathcal{W})$, $\epsilon \in [0, 1]$, Lipschitz constant K.
Definitions: For $1 \leq t \leq T$, let $\lambda_t(w)$ be the likelihood of $w \in \mathcal{W}$ based on observations until time t, i.e.,

$$\lambda_t(w) = \prod_{i=1}^{t} p(U_i \mid x_i, c_i, w).$$

For each $w, w' \in \mathcal{W}$, define $\Lambda_0(w, w') = 0$ and $\Lambda_t(w, w') = \log \frac{\lambda_t(w)}{\lambda_t(w')}$ for $1 \leq t \leq T$.
Policy: While $1 \leq t \leq T$ do:

1. Explore**:** While there is no w such that for every $w' \neq w$, $\max\limits_{s<t} \Lambda_s(w, w') \geq \log T$, assign $x_t = \epsilon$. Note that there can only be one such w.

2. Exit from Explore**:**
   If there is a w such that for every $w' \neq w$, $\max\limits_{s<t} \Lambda_s(w, w') \geq \log T$, define $w^* \triangleq w$ and permanently enter the Exploit phase.

3. Exploit**:** Use the static optimal decision rule in $\mathcal{X}_\epsilon = [\epsilon, 1]$ assuming the model parameter is w^*, i.e., use the decision rule that solves:

$$\max_{\phi:\mathcal{C} \to \mathcal{X}_\epsilon} \mathbb{E}_{\mathcal{D}}[\bar{u}(\phi(c), c, w^*)] \tag{18}$$

$$\text{s.t. } |\phi(c) - \phi(c')| \leq K d_{\mathcal{C}}(c, c') \text{ for all } c,\ c' \in \mathcal{C}.$$

We present the following main result that characterizes the performance of CaFE.

[8] It turns out that this matters only around the $x = 0$ decision. One could alternatively assume a bounded derivative at 0 or a polynomial approximation for the utility function close to 0. Replacing Lipschitzness with these alternatives, however, does not change the nature of results obtained nor provide additional insights.

Theorem 6 *Fix a $K \in [0, \infty)$.*

1. CaFE *is a K-*FH *policy.*

2. *Suppose Assumption 1 holds. Also, suppose that $L(x) = \Omega(x^M)$ (defined in (16)) and $D(x) = \Omega(x^H)$ (defined in (17)) for some $M \geq 0$ and $H \geq 0$ as $x \to 0$, then,*

$$\text{Regret}^T_{K\text{-FH}}(w, \text{CaFE}_T) \leq \beta T^{\frac{M}{M+1}} \log T(1 + \text{o}(1)),$$

as $T \to \infty$, where CaFE_T is CaFE *initialized with $\epsilon = \epsilon_T = 1/T^{\frac{1}{M+1}}$, and $\beta > 0$ is a constant that depends on $L(.)$, $|\mathcal{W}|$, R, M, and H. In particular, if $M = 0$, then* CaFE *initialized with $\epsilon = \epsilon_T = 1/T$ attains a regret of $\text{O}(\log T)$.*

The intuition behind these upper bounds is as follows. In the Explore phase, the parameter w is learned with a probability of error $1/T$ by choosing $x_t = \epsilon_T$ irrespective of c_t.

From that point onwards, the policy enters the exploitation phase: it simply assumes the learned w^* to be the truth and chooses a K-Lipschitz decision rule defined on the decision space $\mathcal{X}_\epsilon = [\epsilon, 1]$. This defines a K-FH policy. Since the decisions are lower bounded by ϵ for every context during the exploitation phase, one incurs a loss of $\text{O}(\epsilon)$ per step. Thus we want to choose ϵ to be as small as possible; in fact, ideally, we would want to pick $\epsilon = 0$. But choosing a small ϵ may force us to learn w prohibitively slowly, thus increasing the length of the Explore phase and hence the overall regret. For example, consider again the setting where the bank is unsure whether age is positively or negatively correlated with loan default rates. It could be the case that overall default rates, irrespective of the context, are small when the loan amounts are small, making it difficult to learn the correlation structure with small loans.

If $L(x) = \Omega(1)$ (i.e., $M = 0$) as $x \to 0$ then we can indeed learn at an adequate rate by picking $\epsilon_T = 1/T$ (in fact, any smaller ϵ suffices), and in this case, we can achieve an overall regret of $\text{O}(\log T)$. On the other hand, if $L(x) = \Theta(x^M)$ for $M > 0$, then ϵ_T cannot be chosen to be too small so that the Explore phase is not prohibitively long; in this case, the choice of $\epsilon_T = 1/T^{1/(M+1)}$ optimizes the tradeoff between regret incurred during the Explore and Exploit phases, leading to an overall loss of $\tilde{\text{O}}(T^{\frac{M}{M+1}})$.

Proof (Theorem 6) We divide the proof into the following claims.

1. CaFE **is K fair-in-hindsight.** First, we show that $\underline{\text{CaFE is } K\text{-FH}}$. To see this, note that the policy is fixed irrespective of the context in the learning phase and hence FH. In the exploitation phase, it is FH with respect to any time in the exploitation phase since the exploitation phase uses a K-Lipschitz decision rule. Finally, it is also FH with respect to the Explore phase in the Exploit phase since decisions for each context only increase in going from Explore to Exploit.

2. **Bound on the expected time for exploration:** For a fixed T, define T' to be the minimum of T and the (random) time at which the Explore phase ends, i.e.,

$$T' = \max\{1 \leq t \leq T \mid \nexists\, w' \in \mathcal{W}, \max_{s<t} \Lambda_s(w', w'') \geq \log T \;\forall\; w'' \neq w'\}. \tag{19}$$

Note that if the Explore phase doesn't end before time T, then $T' = T$. We want to find an upper bound on the expected value of T'.

(a) **Bound on expected time to distinguish any fixed w from w':** We will first bound the expected time until any fixed parameter w gets distinguished from each $w' \neq w$. Consider a new coupled stochastic process $(\bar{\Lambda}_t(w, w'))_{t \in \mathbb{N}}$ that is identical to $(\Lambda_t(w, w'))_{t \leq T}$ up to time T' and then it continues as if the Explore phase did not end (i.e., the allocations are ϵ_T for all contexts forever). Define $T_w^{w'}$ to be the minimum of T and the random time at which w gets "distinguished" from w' in this new stochastic process, i.e., when the log-likelihood ratio of w relative to w' based on when the observations cross the threshold of $\log T$. Formally,

$$T_w^{w'} = \max\{1 \leq t \leq T \mid \max_{s<t} \bar{\Lambda}_s(w, w') < \log T\}. \tag{20}$$

Now, it is easy to show that the process

$$\left(\bar{\Lambda}_t(w, w') - t\mathbb{E}_{\mathcal{D}}(\mathrm{KL}(w, w'|\epsilon_T, c))\right)_{t \in \mathbb{N}}$$

is a martingale and $T_w^{w'}$ is a bounded stopping time (Ross 1996). Thus, by the optional stopping theorem,

$$\mathbb{E}(\bar{\Lambda}_{T_w^{w'}}(w, w') - T_w^{w'}\mathbb{E}_{\mathcal{D}}(\mathrm{KL}(w, w'|\epsilon_T, c)) = 0,$$

that is,

$$\begin{aligned}
\mathbb{E}(T_w^{w'}) = \frac{\mathbb{E}(\bar{\Lambda}_{T_w^{w'}}(w, w'))}{\mathbb{E}_{\mathcal{D}}(\mathrm{KL}(w, w'|\epsilon_T, c))} &\leq \frac{\log T + \mathbb{E}(\log \frac{p(U_{T_w^{w'}}|\epsilon_T, c_{T_w^{w'}}, w)}{p(U_{T_w^{w'}}|\epsilon_T, c_{T_w^{w'}}, w')})}{L(\epsilon_T)} \\
&\quad \text{(by the definition of } T_w^{w'}) \\
&\leq \frac{\log T + \mathbb{E}(\log \frac{1}{b\epsilon_T^H})}{L(\epsilon_T)} \\
&\quad \text{(for large enough } T \text{ for some } b > 0 \text{, since } D(x) = \Omega(x^H)) \\
&\leq \frac{\log T + b' - H \log \epsilon_T}{L(\epsilon_T)} \qquad \text{(for large enough } T\text{),}
\end{aligned}$$

where $b' = -\log b$.

(b) **Use $T_{w'}^{w}$ to bound T':** Define $T_w = \max_{w' \neq w} T_w^{w'}$, i.e., T_w is the minimum of T and the random time at which w gets "distinguished" from *all* $w' \neq w$. Then, the time it takes to differentiate the optimal parameter from all other parameters is in expectation $\mathbb{E}(T_w) = \mathbb{E}(\max_{w' \neq w} T_w^{w'}) \leq \sum_{w' \neq w} \mathbb{E}(T_w^{w'})$, which implies

$$\mathbb{E}(T') \leq (|\mathcal{W}| - 1|)\frac{\log T + b' - H \log \epsilon_T}{L(\epsilon_T)} \quad \text{for large enough } T. \tag{21}$$

3. **Bound on exploitation regret when parameter learnt is wrong:** Next, suppose that w^* is the parameter value that CaFE learns at the end of Explore and the true parameter is w. If we denote P_w^{err} to be the probability that $w^* \neq w$ when the true parameter is w, then we can show that $P_w^{err} \leq 1/T$. This follows from the fact that under the true w, the sequence of likelihood ratios $(\exp(-\Lambda_t(w, w')))_{0 \leq t \leq T}$ is a martingale and hence by Doob's martingale inequality (Ross 1996), $\mathbb{P}_w(\max_{t \leq T} \Lambda_t(w', w) > \log T) = \mathbb{P}_w(\max_{t \leq T} \exp(-\Lambda_t(w, w')) > T) \leq 1/T$ for any $w' \neq w$. This means that the expected regret during the exploitation stage in CaFE is bounded by

$$P_w(w^* \neq w)(T - \mathbb{E}[T' \mid w^* \neq w])R,$$

where R is the Lipschitz constant of the expected utility function.

4. **Bound on exploitation regret due to exploration at ϵ_T:** Finally, if we denote $U_K^{\epsilon_T}(w)$ to be the optimal value of the optimization problem (18) with exploration parameter ϵ_T when $w^* = w$, then we can show that $U_K^{\epsilon_T}(w) \geq U_K(w) - \epsilon_T R$. This is because we can take the optimal K-Lipschitz decision rule ϕ in $\mathcal{X} = [0, 1]$ that attains utility $U_K(w)$, and we can define a new decision rule ϕ' such that $\phi'(c) \triangleq \phi(c)$ if $\phi(c) \geq \epsilon_T$ and $\phi'(c) \triangleq \epsilon_T$ otherwise. It is easy to verify that this decision rule is K-Lipschitz and all the decisions are in $\mathcal{X}_{\epsilon_T}$; hence it is feasible in problem (18). Clearly, the expected utility of this decision rule is at least $U_K(w) - \epsilon_T R$ because of our assumption that $\bar{u}(x, c, w)$ is Lipschitz continuous in x for each c and w, with a Lipschitz constant R. This bounds the expected regret due to exploration at ϵ_T (in spite of $w = w^*$) to be:

$$P_w(w^* = w)(T - \mathbb{E}[T' \mid w^* = w])R\epsilon_T.$$

5. **Total expected regret:** Thus, we finally have that for a fixed model parameter w, the following upper bound holds for the total regret under CaFE:

$$\begin{aligned}
\text{Regret}^T_{K\text{-FT}}(w, \text{CaFE}) &\leq \underbrace{R\mathbb{E}(T')}_{\text{(a)}} + \underbrace{P_w(w^* \neq w)(T - \mathbb{E}[T' \mid w^* \neq w])R}_{\text{(b)}} \\
&\quad + \underbrace{P_w(w^* = w)(T - \mathbb{E}[T' \mid w^* = w])R\epsilon_T}_{\text{(c)}} \\
&\leq R\mathbb{E}(T') + \frac{1}{T}(TR) + TR\epsilon_T \\
&= R\mathbb{E}(T') + R + TR\epsilon_T \\
&\leq R(|\mathcal{W}| - 1|)\frac{\log T + b' - H \log \epsilon_T}{L(\epsilon_T)} + R + TR\epsilon_T,
\end{aligned}$$

for any T large enough, where (a) is the upper bound on the regret incurred during Explore, (b) is the upper bound on the regret incurred during Exploit on the event $\{w^* \neq w\}$, and (c) is the upper bound on the regret incurred during Exploit on the event $\{w^* = w\}$.

Now since $L(x) = \Omega(x^M)$ as $x \to 0$, there exists $a > 0$ such that $L(x) \geq ax^M$. Hence, we have

$$\text{Regret}^T_{K\text{-FT}}(w, \text{CaFE}) \leq R(|\mathcal{W}| - 1|)\frac{\log T + b' - H \log \epsilon_T}{a\epsilon_T^M} + R + TR\epsilon_T,$$

for any T large enough. Plugging in $\epsilon_T = 1/T^{\frac{1}{M+1}}$, we obtain that for a large enough T,

$$\mathrm{Regret}^T_{K\text{-FT}}(w, \textsc{CaFE}_T) \leq R(|\mathcal{W}| - 1|)T^{\frac{M}{M+1}} \frac{\log T + b' + \frac{H}{M+1}\log T}{a} + R + T^{\frac{M}{M+1}} R.$$

Thus,

$$\limsup_{T\to\infty} \frac{\mathrm{Regret}^T_{K\text{-FT}}(w, \textsc{CaFE}_T)}{T^{\frac{M}{M+1}} \log T} \leq \frac{R(|\mathcal{W}| - 1|)(1 + \frac{H}{M+1})}{a}.$$

This proves the statement of the theorem.

■

Remark 1. If there exists a set of optimal K-Lipschitz decision-rules $(\phi^*_w)_{w\in\mathcal{W}}$ such that

$$\min_{c\in\mathcal{C}, w\in\mathcal{W}} \phi^*_w = x^* \in (0, 1),$$

then, although the performance in Theorem 6 still holds in this case, one can obtain a better performance by transforming the decision space to identify x^* with 0. Formally, one can transform the decision space $[x^*, 1]$ by mapping any decision x in this set to $(x-x^*)/(1-x^*)$, thus ensuring that the transformed decisions lie in $[0, 1]$, and the smallest optimal decision across all parameters and contexts is 0.

Remark 2. We believe that in most practical situations, for any closed set $\mathcal{S}$ in the relative interior of $[0, 1]$, i.e., $\mathcal{S} \subseteq [0, 1] \setminus \{0, 1\}$, we will have $\min_{x\in\mathcal{S}} L(x) > 0$, i.e., any decision in the set $\mathcal{S}$ will be able to distinguish between any pair w, w'. Thus, informally, only the extreme decisions {0,1} are problematic and hence the distinguishability between w, w' as one approaches either 0 or 1 impacts performance. An important class of settings, where this assumption on $L(x)$ is trivially true is, when there are two actions A and B, and a decision $x \in [0, 1]$ represents the probability with which action B is chosen (say, B is more conducive than A). For example, the decision x could simply be the probability with which an applicant gets approved for a pre-determined amount of loan (A = denial, B = approval).

In these cases, one can, without loss of generality, assume that for every possible pair w, w', either action A or B can distinguish this pair under some context (although, any single action may not allow distinguishability of all pairs). This is because if there is a pair of parameters w, w' that cannot be distinguished by either of the actions under any context, then one can simply aggregate this pair into a single parameter value. Hence, any decision x that chooses both the actions with a positive probability, i.e., x is in the relative interior of $[0, 1]$, can distinguish between all pairs of parameter values. For example, denial of the loan application (action A) may not provide any information about the utility model. However, approval (action B) would provide relevant information. In such settings, if $\min_{c\in\mathcal{C}, w\in\mathcal{W}} \phi^*_w = x^* > 0$, then $\mathrm{O}(\log T)$ regret is achievable using CaFE by the transformation of decision space mentioned in Remark 1.

Remark 3. It is easy to show that if it is necessary to distinguish some pair w, w' because the corresponding optimal individually fair decision rules are different, and low decisions do not allow such a distinction, then linear regret may be inevitable if high decisions are sub-optimal for either w or w'. For instance, suppose there are only two possible parameter values w and w', and there is a single context. The optimal decision under w for this context is 0 and optimal decision under w' is 1. Suppose that there is a $\delta > 0$ such such that $\mathrm{KL}(w, w'|x) = 0$ for any $x \in [0, \delta]$. Then in order to get $\mathrm{o}(T)$ regret for w, one must take $\Omega(T)$ decisions in $[0, \delta]$. But since these decisions offer no distinction between w and w', $\Omega(T)$ regret is inevitable under w'.

Remark 4. The linear dependence of the regret on $|\mathcal{W}|$ can be eliminated for the case where $L(x) = \Omega(1)$ as $x \to 0$. The source of this dependence is the union bound that was utilized in bounding $\mathbb{E}(T_w)$ in the proof. Recall that T_w is the minimum of T and the time taken until the random walks of the log-likelihood ratios of observations under w and w', for the different values of w', all cross the threshold of $\log T$. Our proof bounds this quantity by the sum of the times taken by each of the random walks to cross this threshold, leading to the $|\mathcal{W}|$ dependence. In the case where $L(x) = \Omega(1)$, we can directly bound $\mathbb{E}(T_w)$ by arguing that the time taken by all the random walks to cross this threshold is essentially governed by the slowest random walk, i.e., the one corresponding to the parameter value w' for which the drift $\mathbb{E}_{\mathcal{D}}(\mathrm{KL}(w, w'|\epsilon, c))$ is the smallest. This is expressed in the following result, which follows from Lemma 4.3 in Agrawal et al. (1989). It crucially leverages the fact that as $T \to \infty$, the drift $\mathbb{E}_{\mathcal{D}}(\mathrm{KL}(w, w'|\epsilon_T, c))$ is bounded below by a positive constant for each $w' \neq w$.

Proposition 7 *(Agrawal et al. 1989) For each $w' \in \mathcal{W}\backslash\{w\}$, consider the stochastic process $(\bar{\Lambda}_t(w, w'))_{t\in\mathbb{N}}$ of log-likelihood ratios of observations under w and under w' generated by choosing the decision $\epsilon_T = 1/T$ at each time. Let $T_w^{w'}$ be as defined in* (20) *and define $T_w = \max_{w'\neq w} T_w^{w'}$. Suppose that $L(x) \geq a$ for any small enough $x > 0$. Then,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}(T_w)}{\log T} \leq \frac{1}{a}. \tag{22}$$

In the general case where $L(x) = \Omega(x^M)$ for $M > 0$ as $x \to \infty$, the underlying argument doesn't naturally extend, because, informally, the drifts of these random walks shrink to 0 as time $T \to \infty$, and thus the growing variation in the time taken by each random walk to cross the threshold of $\log T$ starts to become the dominating factor in the determination of the expected time taken for all walks to cross the threshold. In this case, it is unclear if the dependence of the regret on $|\mathcal{W}|$ can be improved, and we leave this as an open question.

5.3 Lower bounds on regret under FH

The upper bound on the performance of CaFE in Theorem 6 depends on the exponent (i.e., M) of some polynomial lower bound that is known for $L(x)$ for $x \to 0$. This exponent determines the order of T in the bound. We now construct examples and corresponding lower bounds on regret, which demonstrate that this dependence of CaFE's regret on T is order-optimal.

The proofs of these bounds rely on the following high-level argument. Suppose there are two hypotheses about the state of the world: one in which high decisions are optimal and the other in which low decisions are optimal. If a policy raises decisions too quickly (and importantly, irrevocably, because of the FH constraint) without obtaining sufficiently strong empirical evidence for the hypothesis that it is indeed in a situation where high decisions are optimal, then (a "change of measure" argument shows that) there is a good chance that the hypothesis is incorrect, and because of the FH constraint, the policy will incur a high regret. The fact that a policy is *good*, i.e., it does not incur high regret irrespective of the true hypothesis, implies an upper bound on the probability of this policy raising decisions too quickly without obtaining sufficient empirical evidence to justify it. This means that any good policy must obtain sufficient evidence before raising decisions. But *it takes time* to obtain this evidence, which means that it is *inevitable* that the policy will incur some regret in the case that high decisions are optimal. In the case where $\lim_{x\to 0} L(x) = 0$ (i.e., $M > 0$), the situation is worse: while the decisions are low, it takes *even longer* to obtain sufficient empirical evidence to justify higher decisions, thus increasing the amount of inevitable regret.

5.3.1 Case: $L(x) = \Theta(x^M)$.

First, we construct an instance with $L(x) = \Theta(x^M)$ as $x \to 0$ where a regret of $\omega(T^{\frac{M}{M+1}-\beta})$ is inevitable for any $\beta > 0$. Suppose that $\mathcal{C} = \{0\}$, i.e., there is only one context. Hence the choice of the Lipschitz constant K is immaterial, and the FH constraint simply means that the decisions must be non-decreasing over time. Let $\mathcal{W} = \{A, B\}$ and let the distribution of utility given the decision x and parameter w be defined as:

$$U \overset{(w=A)}{=} \begin{cases} x^{1-M/2} & \text{w.p. } 0.5(1+x^{M/2}), \\ -x^{1-M/2} & \text{w.p. } 0.5(1-x^{M/2}), \end{cases}$$

and

$$U \overset{(w=B)}{=} \begin{cases} x^{1-M/2} & \text{w.p. } 0.5(1-x^{M/2}), \\ -x^{1-M/2} & \text{w.p. } 0.5(1+x^{M/2}). \end{cases} \tag{23}$$

It is convenient to define $p(u \mid x, w)$ as the probability of observing $u \in \{x^{1-M/2}, -x^{1-M/2}\}$ for a fixed $w \in \mathcal{W}$, and $x \in \mathcal{X}$.[9] It is easy to see that the mean utilities are: $\bar{u}(x, A) = x$ and $\bar{u}(x, B) = -x$ (where we have suppressed the dependence on the context). Clearly, the optimal decision is $x = 1$ if $w = A$ and $x = 0$ if $w = B$. For any $p, q \in (0, 1)$, let $D_{\mathrm{KL}}(p\|q)$ denote the the K-L divergence of a Bernoulli(p) distribution relative to a Bernoulli(q) distribution, i.e., $D_{\mathrm{KL}}(p\|q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$. Recall that $L(x)$ was defined to be the minimum expected KL-divergence over any two parameters w, w' (see (16)). For this instance, note that $D_{\mathrm{KL}}(0.5(1+x^{M/2})\|0.5(1-x^{M/2}) = D_{\mathrm{KL}}(0.5(1-x^{M/2})\|0.5(1+x^{M/2})) = x^{M/2}\log\frac{1+x^{M/2}}{1-x^{M/2}}$, and hence

$$\begin{aligned} L(x) &= \min(D_{\mathrm{KL}}(0.5(1+x^{M/2})\|0.5(1-x^{M/2})), D_{\mathrm{KL}}(0.5(1-x^{M/2})\|0.5(1+x^{M/2}))) \\ &= x^{M/2}\log\frac{1+x^{M/2}}{1-x^{M/2}}. \end{aligned} \tag{24}$$

[9] Here $p(x^{1-M/2} \mid x, A) = p(-x^{1-M/2} \mid x, B) = 0.5(1 + x^{\frac{M}{2}})$, $p(-x^{1-M/2} \mid x, A) = p(x^{1-M/2} \mid x, B) = 0.5(1 - x^{\frac{M}{2}})$ and 0 otherwise.

Note that $L(x)$ is increasing in x and one can show that $L(x) = \Theta(x^M)$ as $x \to 0$. This means that for x small enough and for some $0 < v < V$, we have

$$vx^M \leq L(x) \leq Vx^M. \tag{25}$$

Proposition 8 *Consider the instance defined above. Suppose there is a sequence of K-FH policies $(\psi_T)_{T\in\mathbb{N}}$ that satisfy,*

$$\text{Regret}^T_{K\text{-FH}}(w, \psi_T) = \mathrm{o}(T^\alpha),$$

as $T \to \infty$ for each $\alpha > \frac{M}{M+1}$ and each $w \in \{A, B\}$. Then

$$\text{Regret}^T_{K\text{-FH}}(A, \psi_T) = \omega(T^\beta),$$

as $T \to \infty$ for each $\beta < \frac{M}{M+1}$.

Proof For any underlying w, the policy ψ_T along with the distributions of utilities conditioned on decisions given in (23), induce a probability distribution on the sequence of decision and utility pairs $(X_1, U_1, X_2, U_2, \cdots, X_T, U_T)$. Let $\mathbb{P}_w$ and $\mathbb{E}_w$ denote the probabilities of events and expectations, respectively, under $w = A$ and $w = B$. Here X_t is measurable with respect to the σ-algebra generated by $(X_1, U_1, X_2, U_2, \cdots, X_{t-1}, U_{t-1})$, and U_t is conditionally independent of the past given X_t. This, in particular, means that for any function $f : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, and each $w \in \{A, B\}$,

$$\mathbb{E}_w[f(U_t, X_t) \mid X_1, U_1, X_2, U_2, \cdots, X_t] = \mathbb{E}_w[f(U_t, X_t) \mid X_t]. \tag{26}$$

Define the (random) sequence of empirical log-likelihood ratios of $w = A$ relative to $w = B$, $(\Lambda_t)_{t\leq T}$, where

$$\Lambda_t = \sum_{s=1}^{t} \log \frac{p(U_s \mid X_s, A)}{p(U_s \mid X_s, B)}. \tag{27}$$

Also, we will be using a "centered" sequence $(\overline{\Lambda}_t)_{0\leq t\leq T}$, where $\overline{\Lambda}_t = \Lambda_t - \mu_t$ and $(\mu_t)_{0\leq t\leq T}$ is the mean process defined as $\mu_0 = 0$ and for any $1 \leq t \leq T$,

$$\mu_t = \sum_{s=1}^{t} \mathbb{E}_A \left[\log \frac{p(U_s \mid X_s, 1)}{p(U_s \mid X_s, 0)} \,\middle|\, X_s \right], \tag{28}$$

Using (26), it is easy to see that $(\overline{\Lambda}_t)_{0\leq t\leq T}$ is a martingale.

The proof is now divided into two parts.

Part A. **One cannot raise decisions too quickly without obtaining sufficient empirical evidence to justify it.**

Let us call a policy ψ_T *good* if $\text{Regret}^T_{K\text{-FH}}(B, \psi_T) = \mathrm{o}(T^\alpha)$ as $T \to \infty$ for each $\alpha > \frac{M}{M+1}$. In this part, we will define a threshold $\tau(T) \in \mathcal{X} = [0, 1]$. When $w = A$, we will show that under any good policy ψ_T, the probability of raising the decision beyond

$\tau(T)$ despite the fact that there is insufficient evidence for $w = A$ (i.e., log-likelihood ratio of A relative to B is low) is $\mathrm{o}(1)$.

Let the threshold be $\tau(T) = T^{\frac{M}{M+1}+\gamma-1}$ for any $\gamma \in (0, 1/(M+1))$. Note that $\tau(T) = \mathrm{o}(1)$ as $T \to \infty$. Define k_T to be the first time that the policy raises decisions higher than $\tau(T)$ ($k_T = T$ if it doesn't), i.e., $k_T = \max\{1 \leq k \leq T \text{ s.t. } X_k < \tau(T)\}$. Define the event:

$$C_T \triangleq \left\{ k_T \leq \frac{\gamma \log T}{2L(\tau(T))} \text{ and } \Lambda_{k_T} \leq \frac{3\gamma}{4} \log T \right\}.$$

Informally, this event says that the decision crosses $\tau(T)$ before time $\frac{\gamma \log T}{2L(\tau(T))}$ and at the time of crossing, the empirical log-likelihood ratio of A relative to B is low, i.e., it is below $\frac{3\gamma}{4} \log T$. Then we have

$$\mathbb{P}_B(C_T) = \mathbb{E}_B(\mathbb{1}_{C_T}) \overset{(a)}{=} \mathbb{E}_A(\mathbb{1}_{C_T} \exp(-\Lambda_{k_T})) \geq \mathbb{P}_A(C_T) T^{-3\gamma/4}. \tag{29}$$

Here, (a) is the standard change of measure identity. Hence,

$$\begin{aligned}
\mathrm{Regret}^T_{K\text{-FH}}(B, \psi_T) &\overset{(a)}{\geq} \mathbb{P}_B(C_T)\tau(T)\,(T - k_T) \\
&\overset{(b)}{\geq} \mathbb{P}_B(C_T)\tau(T)\left(T - \frac{\gamma \log T}{2L(\tau(T))}\right) \\
&\overset{(c)}{\geq} \mathbb{P}_A(C_T)T^{-3\gamma/4}\tau(T)\left(T - \frac{\gamma \log T}{2L(\tau(T))}\right) \\
&\overset{(d)}{\geq} \mathbb{P}_A(C_T)T^{-3\gamma/4}\tau(T)\left(T - \frac{\gamma \log T}{2v\ \tau(T)^M}\right),
\end{aligned}$$

for a large enough T. Here (a) follows from the fact that when $w = B$, then on the event C_T, the policy will incur a regret of at least $\tau(T)$ in each time period after time k_T. (b) is simply using the definition of C_T, (c) follows from (29), and (d) follows from (25). Now, since,

$$\frac{1}{\tau(T)^M} = T^{(-M^2/M+1)-M\gamma+M} \leq T^{(-M^2/M+1)+M} = T^{\frac{M}{M+1}},$$

we have that $\frac{\gamma \log T}{2v\tau(T)^M} = \mathrm{o}(T)$. Thus, we finally have,

$$\begin{aligned}
\mathrm{Regret}^T_{K\text{-FH}}(B, \psi_T) &\geq \mathbb{P}_A(C_T)T^{-3\gamma/4+(\frac{M}{M+1})+\gamma-1}(T - \mathrm{o}(T)) \\
&= \mathbb{P}_A(C_T)T^{\gamma/4+(\frac{M}{M+1})}(1 - \mathrm{o}(1)).
\end{aligned}$$

Since $\mathrm{Regret}^T_{K\text{-FH}}(B, \psi_T) = \mathrm{o}(T^\alpha)$ for each $\alpha > \frac{M}{M+1}$, we have $\mathbb{P}_A(C_T) = \mathrm{o}(1)$. Thus, for $\overline{C}_T$, the complement of the event C_T, we get:

$$\mathbb{P}_A(\overline{C}_T) = \mathbb{P}_A\left(\underbrace{k_T > \frac{\gamma \log T}{2L(\tau(T))}}_{(\star)} \text{ or } \underbrace{\Lambda_{k_T} > \frac{3\gamma}{4} \log T}_{(\dagger)} \right) = 1 - \mathrm{o}(1). \tag{30}$$

This shows that when $w = A$, any good policy must wait for a sufficiently long time before raising the decisions beyond $\tau(T)$ ($\star$), *or* there must be sufficient empirical evidence for $w = A$ at the time when the decision is raised beyond $\tau(T)$ (†). We next show that it is inevitable that one has to wait sufficiently long before raising the decision beyond $\tau(T)$. This will follow from the fact that it takes time for the log-likelihood ratio of $w = A$ relative to $w = B$ to grow sufficiently.

Part B. **It takes time (and hence regret) to gather sufficient empirical evidence.**

We now show that $\mathbb{P}_A\left(k_T \leq \dfrac{\gamma \log T}{2L(\tau(T))} \text{ and } \Lambda_{k_T} > \dfrac{3\gamma}{4}\log T\right) = \mathrm{o}(1)$. Denote $z(T) = \lfloor \frac{\gamma \log T}{2L(\tau(T))} \rfloor$. Then we have,

$$\begin{aligned}
&\mathbb{P}_A\left(k_T \leq \frac{\gamma \log T}{2L(\tau(T))} \text{ and } \Lambda_{k_T} > \frac{3\gamma}{4}\log T\right)\\
&\quad = \mathbb{P}_A\left(k_T \leq z(T) \text{ and } \Lambda_{\min(k_T, z(T))} > \frac{3\gamma}{4}\log T\right)\\
&\quad \leq \mathbb{P}_A\left(\Lambda_{\min(k_T, z(T))} > \frac{3\gamma}{4}\log T\right)\\
&\quad = \mathbb{P}_A\left(\overline{\Lambda}_{\min(k_T, z(T))} > \frac{3\gamma}{4}\log T - \mu_{\min(k_T, z(T))}\right)\\
&\quad \overset{(a)}{\leq} \mathbb{P}_A\left(\overline{\Lambda}_{\min(k_T, z(T))} > \frac{3\gamma}{4}\log T - z(T)L(\tau(T))\right)\\
&\quad \leq \mathbb{P}_A\left(\overline{\Lambda}_{\min(k_T, z(T))} > \frac{\gamma}{4}\log T\right).
\end{aligned}$$

Here, (a) follows from the definition of μ_t in equation (28) and from the fact that

$$\mathbb{E}_A\left[\log \frac{p(U_s \mid X_s, 1)}{p(U_s \mid X_s, 0)} \mid X_s\right] = L(X_s) \leq L(\tau(T)),$$

almost surely for all $s \leq k_T$, since $X_s < \tau(T)$ for all $s \leq k_T$, and $L(\cdot)$ is increasing.

We now define a new policy ψ'_T with associated random variables that will be differentiated from the corresponding random variables under ψ_T by adding a "prime" superscript, e.g., $X \to X'$. This new policy follows the prescriptions of ψ_T until one of the two events happen:

(a) $\overline{\Lambda}'_t > \frac{\gamma}{4}\log T$, in which case it increases decision to $\tau(T)$ and chooses $\tau(T)$ until the end of the horizon.

(b) ψ_T prescribes raising the decision from some $x' < \tau(T)$ to some $x'' \geq \tau(T)$, in which case it continues to play x' until either condition (1) is satisfied, or until the end of the horizon.

Effectively, this policy raises decision to $\tau(T)$ exactly when $\Lambda'_t > \frac{\gamma}{4}\log T$ and then fixes the decisions at $\tau(T)$. Define the random variable

$$k'_T \triangleq \max\left\{k \le T \text{ s.t. } X'_k < \tau(T)\right\} = \max\left\{k \le T \text{ s.t. } \sup_{s<k} \overline{\Lambda}'_s \le \frac{\gamma}{4}\log T\right\}. \tag{31}$$

Suppose that the sample paths under the two policies are coupled until the point that these two policies have identical prescriptions. Then, by the construction of ψ'_T, it is clear that on each sample path that $\overline{\Lambda}_{\min(k_T,z(T))} > \frac{\gamma}{4}\log T$ occurs, $\overline{\Lambda}'_{\min(k'_T,z(T))} > \frac{\gamma}{4}\log T$ occurs as well. Thus,

$$\begin{aligned}
\mathbb{P}_A\left(\overline{\Lambda}_{\min(k_T,z(T))} > \frac{\gamma}{4}\log T\right)) &\le \mathbb{P}_A\left(\overline{\Lambda}'_{\min(k'_T,z(T))} > \frac{\gamma}{4}\log T\right) \\
&= \mathbb{P}_A\left(k'_T \le z(T)\right) \\
&\le \mathbb{P}_A\left(\sup_{s\le z(T)} \overline{\Lambda}'_s > \frac{\gamma}{4}\log T\right) \\
&\le \frac{\operatorname{var}(\overline{\Lambda}'_{z(T)})}{(\log T)^2} \\
&\quad \text{(by Kolmogorov's maximal inequality (Ross 1996))} \\
&= \frac{\sum_{t=1}^{z(T)} \operatorname{var}\big(\log \frac{p(U'_t|X'_t,w)}{p(U'_t|X'_t,w')} - \mathbb{E}_A\left[\log \frac{p(U'_t|X'_t,w)}{p(U'_t|X'_t,w')} \mid X'_t\right]\big)}{(\log T)^2} \\
&= \frac{\sum_{t=1}^{z(T)} \mathbb{E}_A\left(\operatorname{var}\left(\log \frac{p(U'_t|X'_t,w)}{p(U'_t|X'_t,w')} \mid X'_t\right)\right)}{(\log T)^2}.
\end{aligned}$$

The random variable $\log \frac{p(U'_t|X'_t,w)}{p(U'_t|X'_t,w')}$ lies in $[-\log \frac{1+(X'_t)^{M/2}}{1-(X'_t)^{M/2}}, \log \frac{1+(X'_t)^{M/2}}{1-(X'_t)^{M/2}}]$. Under policy ψ'_T, $X'_t \le \tau(T)$ almost surely for all $t \le T$. Thus the range of $\log \frac{p(U'_t|X'_t,w)}{p(U'_t|X'_t,w')}$ is at most $\left[-\log \frac{1+\tau(T)^{M/2}}{1-\tau(T)^{M/2}}, \log \frac{1+\tau(T)^{M/2}}{1-\tau(T)^{M/2}}\right]$. Hence, by Popoviciou's inequality for the variances,[10]

$$\operatorname{var}\left(\log \frac{p(U'_t \mid X'_t, w)}{p(U'_t \mid X'_t, w')} \,\middle|\, X'_t\right) \le \left(\log \frac{1+\tau(T)^{M/2}}{1-\tau(T)^{M/2}}\right)^2 = \mathrm{O}(\tau(T)^M),$$

as $\tau(T) \to 0$. Hence, we finally have

$$\mathbb{P}_A\left(\overline{\Lambda}'_{\min(k_T,z(T))} > \frac{\gamma}{4}\log T\right) \le \frac{z(T)\mathrm{O}(\tau(T)^M)}{(\log T)^2} \le \frac{\gamma \mathrm{O}(\tau(T)^M)}{L(\tau(T))\log T} \overset{(a)}{\le} \frac{\gamma \mathrm{O}(\tau(T)^M)}{v\tau(T)^M \log T} = \mathrm{o}(1),$$

where (a) follows from (25). Thus, to reiterate, we have shown that

$$\mathbb{P}_A\left(k_T \le \frac{\gamma \log T}{2L(\tau(T))} \text{ and } \Lambda_{k_T} > \frac{3\gamma}{4}\log T\right) = \mathrm{o}(1),$$

thus showing part B.

[10] If a random variable takes values in $[a, b]$, then its variance is at most $(b-a)^2/4$.

Coupled with (30), this implies that $\mathbb{P}_A\left(k_T > \frac{\gamma \log T}{2L(\tau(T))}\right) = 1 - \mathrm{o}(1)$. Hence,

$$\begin{aligned}
\mathrm{Regret}^T_{K\text{-FH}}(A, \psi_T) &\geq (1 - \mathrm{o}(1))(1 - \tau(T))\frac{\gamma \log T}{2L(\tau(T))} \\
&\geq (1 - \mathrm{o}(1))(1 - \mathrm{o}(1))\frac{\gamma \log T}{2V\tau(T)^M} \text{ (for a large enough } T\text{)} \\
&\geq (1 - \mathrm{o}(1))\frac{\gamma}{2V}T^{\frac{M}{M+1} - M\gamma} \log T \text{ (for a large enough } T\text{)}.
\end{aligned}$$

Thus $\mathrm{Regret}^T_{K\text{-FH}}(A, \psi_T) = \omega(T^{\frac{M}{M+1} - M\gamma})$ for every $\gamma > 0$, hence proving the claim.

■

Remark 5. If there was no FH constraint, a cumulative regret of at most 1 can be achieved in this instance: one can simply choose $x_1 = 1$, which immediately reveals whether $w = 1$ (if $U_1 = 1$) or $w = 0$ if ($U_1 = -1$). This demonstrates the stark impact of the FH constraint on regret.

This lower bound, however, does not resolve whether in the setting where $L(x) = \Theta(1)$, a regret of $\mathrm{O}(\log T)$ is necessary, as our upper bound suggests. We show next that this is indeed the case.

5.3.2 Case: $L(x) = \Theta(1)$.

We now construct an instance with $L(x) = \Theta(1)$ as $x \to 0$ where an expected regret of $\Omega(\log T)$ is inevitable. Suppose that $\mathcal{C} = \{0\}$ and let $\mathcal{W} = \{A, B\}$. Again, since there is only one context, the choice of the Lipschitz constant K is immaterial, and the FH constraint simply means that the decisions must be non-decreasing over time. The utility at time t given the decision x_t and parameter w is given by $U_t = x_t F_t$ where F_t is i.i.d. across time, distributed as:

$$F_t \overset{(w=A)}{=} \begin{cases} 1 & \text{w.p. } 0.75, \\ -1 & \text{w.p. } 0.25 \end{cases} \quad \text{and } F_t \overset{(w=B)}{=} \begin{cases} 1 & \text{w.p. } 0.25, \\ -1 & \text{w.p. } 0.75 \end{cases} .$$

Clearly, the optimal decision is $x = 1$ if $w = A$ and $x = 0$ if $w = B$.

Proposition 9 *Consider the instance defined above. Suppose there is a sequence of K-FH policies $(\psi_T)_{T\in\mathbb{N}}$ that satisfy,*

$$\mathrm{Regret}^T_{K\text{-FH}}(w, \psi_T) = \mathrm{o}(T^\alpha)$$

as $T \to \infty$ for each $\alpha > 0$ and each $w \in \{A, B\}$. Then as $T \to \infty$,

$$\mathrm{Regret}^T_{K\text{-FH}}(A, \psi_T) \geq \Omega(\log T).$$

The structure of the proof for this case is similar to that of Proposition 3; in fact, the arguments are simpler since the KL-divergence of $w = A$ relative to $w = B$ (or vice-versa) is independent of the decisions taken by the policy for non-zero decisions. We detail the proof below for completeness.

Proof The policy ψ_T along with the distribution of F_t induces a probability distribution on the sequence of decision and utility pairs $(X_1, U_1, X_2, U_2, \cdots, X_T, U_T)$. Here X_t is measurable with respect to the σ-algebra generated by $(X_1, U_1, X_2, U_2, \cdots, X_{t-1}, U_{t-1})$, and U_t is conditionally independent of the past given X_t. Let $\mathbb{P}_w$ and $\mathbb{E}_w$ denote the probabilities of events and expectations, respectively, under $w = A$ and $w = B$.

Define the (random) sequence of empirical log-likelihood ratios of $w = A$ relative to $w = B$, $(\Lambda_t)_{0 \le t \le T}$, where $\Lambda_0 = 0$ and for $1 \le t \le T$,

$$\Lambda_t = \sum_{s=1}^{t} \mathbb{1}_{\{X_t > 0\}} \log \frac{0.75\mathbb{1}_{\{F_t=1\}} + 0.25\mathbb{1}_{\{F_t=-1\}}}{0.75\mathbb{1}_{\{F_t=-1\}} + 0.25\mathbb{1}_{\{F_t=1\}}}. \tag{32}$$

This is the sequence seen by the policy whenever a non-zero decision is taken (otherwise F_t is not observed). We denote the *complete* sequence of empirical log-likelihood ratios by $(\Lambda_t^c)_{0 \le t \le T}$, where $\Lambda_0^c = 0$ and for $1 \le t \le T$,

$$\Lambda_t^c = \sum_{s=1}^{t} \log \frac{0.75\mathbb{1}_{\{F_t=1\}} + 0.25\mathbb{1}_{\{F_t=-1\}}}{0.75\mathbb{1}_{\{F_t=-1\}} + 0.25\mathbb{1}_{\{F_t=1\}}}. \tag{33}$$

Note that this sequence only depends on F_t and is independent of the decisions X_t taken by the policy. Similar to Proposition 8, the proof is now divided into two parts.

Part A. **One cannot raise decisions too quickly without obtaining sufficient empirical evidence to justify it.**

Define a threshold $\tau(T) = 1/T^{\gamma/4}$ for $\gamma > 0$ and let $k_T = \max\{1 \le k \le T \text{ s.t. } x_k < 1/T^{\gamma/4}\}$. Consider the following event:

$$C_T \triangleq \left\{ k_T \le \frac{(1-\gamma)\log T}{D_{\text{KL}}(0.75\|0.25)} \text{ and } \Lambda_{k_T} \le (1 - \frac{\gamma}{2}) \log T \right\}. \tag{34}$$

By the change of measure identity, we have that

$$\mathbb{P}_B(C_T) = \mathbb{E}_B(\mathbb{1}_{C_T}) = \mathbb{E}_A(\mathbb{1}_{C_T} \exp(-\Lambda_{k_T})) \ge \mathbb{P}_A(C_T) T^{-1+\gamma/2}. \tag{35}$$

Hence,

$$\begin{aligned}
\text{Regret}_{K\text{-FH}}^T(B, \psi_T) &\overset{(a)}{\ge} \mathbb{P}_B(C_T) \frac{1}{T^{\gamma/4}} (T - k_T) \\
&\overset{(b)}{\ge} \mathbb{P}_A(C_T) \frac{T^{-1+\gamma/2}}{T^{\gamma/4}} (T - k_T) \\
&\overset{(c)}{\ge} \mathbb{P}_A(C_T) \frac{T^{-1+\gamma/2}}{T^{\gamma/4}} \left(T - \frac{(1-\gamma)\log T}{D_{\text{KL}}(0.75\|0.25)} \right).
\end{aligned}$$

Here, (a) follows from the fact that for $w = B$, on event C_T, one incurs a regret of $T^{-\gamma/4}$ per time step after k_T. (b) follows from (35), and (c) follows from the fact that on event C_T, $k_T \leq \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}$. Since $\mathrm{Regret}^T_{K\text{-FH}}(w, \psi_T) = \mathrm{o}(T^\alpha)$ for each $\alpha > 0$ and for each w, we have $\mathbb{P}_A(C_T) = \mathrm{o}(1)$. Thus, if we denote $\overline{C}_T$ to be the complement of the event C_T, then we have,

$$\mathbb{P}_A(\overline{C}_T) = \mathbb{P}_A\left(k_T > \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)} \text{ or } \Lambda_{k_T} > (1-\frac{\gamma}{2})\log T\right) = 1 - o(1). \tag{36}$$

Part B. **It takes time (and hence regret) to gather sufficient empirical evidence.**

Next, denote $b(T) = \lfloor \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)} \rfloor$. Then

$$\begin{aligned}
&\mathbb{P}_A\left(k_T \leq \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)} \text{ and } \Lambda_{k_T} > (1-\frac{\gamma}{2})\log T\right) \\
&= \mathbb{P}_A\left(k_T \leq b(T) \text{ and } \Lambda_{k_T} > (1-\frac{\gamma}{2})\log T\right) \\
&\leq \mathbb{P}_A\left(k_T \leq b(T) \text{ and } \sup_{t \leq b(T)} \Lambda_t > (1-\frac{\gamma}{2})\log T\right) \\
&\leq \mathbb{P}_A\left(\sup_{t \leq b(T)} \Lambda_t > (1-\frac{\gamma}{2})\log T\right) \\
&\leq \mathbb{P}_A\left(\sup_{t \leq b(T)} \Lambda^c_t > (1-\frac{\gamma}{2})\log T\right) \\
&\leq \mathbb{P}_A\left(\frac{1}{b(T)} \sup_{t \leq b(T)} \Lambda^c_t > \frac{(1-\frac{\gamma}{2})}{1-\gamma} D_{\mathrm{KL}}(0.75\|0.25)\right)) \\
&= \mathrm{o}(1).
\end{aligned} \tag{37}$$

The last inequality results from the fact that, by the maximal version of the strong law of large numbers (see Theorem 2.2 of Bubeck et al. (2012)),

$$\frac{1}{b(T)} \sup_{t \leq b(T)} \Lambda^c_t \xrightarrow{a.s.} D_{\mathrm{KL}}(0.75\|0.25) \text{ as } T \to \infty.$$

Combining (36) and (37), we finally have,

$$\mathbb{P}_A\left(k_T \geq \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}\right) = 1 - \mathrm{o}(1).$$

Hence,

$$\mathrm{Regret}^T_{K\text{-FH}}(A, \psi_T) \geq (1 - o(1))\left(1 - \frac{1}{T^{\gamma/4}}\right)\frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}.$$

This implies the result. ■

Remark 6. In the absence of FH constraint, we can show that a $\mathrm{O}(1)$ cumulative regret can be guaranteed in the instance above. This can, for instance, be achieved by the following policy that always chooses $x_t > 0$. Let $\bar{F}_t = (1/t)\sum_{s=1}^{t} U_t/x_t = (1/t)\sum_{s=1}^{t} F_t$. Then choose $x_1 = 1$ and for $t \geq 2$, choose $x_t = 1$ if $\bar{F}_{t-1} \geq 0$ and $x_t = e^{-t}$ if $\bar{F}_{t-1} < 0$. Thus the total expected regret on the event $w = 1$ is upper bounded by $\sum_{t=1}^{T} \mathbb{P}_A(\bar{F}_{t-1} < 0)$, and the total expected regret on the event $w = 1$ is upper bounded by

$$\sum_{t=1}^{T} e^{-t}\mathbb{P}_B(\bar{F}_{t-1} < 0) + \mathbb{P}_B(\bar{F}_{t-1} \geq 0) \leq \sum_{t=1}^{T} e^{-t} + \mathbb{P}_B(\bar{F}_{t-1} \geq 0).$$

Both these quantities are $\mathrm{O}(1)$ as $T \to \infty$ since by the Hoeffding bound, $\mathbb{P}_B(\bar{F}_t \geq 0) \leq e^{-\nu_1 t}$ and $\mathbb{P}_A(\bar{F}_t < 0) \leq e^{-\nu_2 t}$ for some instance dependent constants ν_1, $\nu_2 > 0$.

5.4 Knowledge of T is necessary to obtain a good FH algorithm

When the time horizon T is unknown, a technique known as the "doubling trick" often allows one to achieve the same order of regret as that what is possible with the knowledge of T (for example, see Besson and Kaufmann (2018)). However, since the constraints for individual fairness in hindsight do not allow changes to a less conducive decision over time, the doubling trick does not work in our setting. In fact, it is simply not possible to obtain an algorithm with good guarantees without knowing the horizon T. We show this in the context of the instance defined in Section 5.3.2. In particular, we show that for this instance, there is no sequence of T horizon policies that can simultaneously achieve a $\mathrm{o}(\log T)$ regret on the horizon of duration $\Theta(\log T)$ and a $\mathrm{O}(T^{\alpha})$ regret for the horizon of duration T for any $\alpha < 1$. The informal argument is as follows. When the parameter is $w = A$ (the decision 1 is optimal in this case), in order to guarantee a regret of at most $\mathrm{o}(\log T)$ for the $\Theta(\log T)$ horizon, one can spend at most $\mathrm{o}(\log T)$ time choosing low decisions, e.g., decisions less than 1/2. However, raising the decisions that quickly does not allow sufficient distinction from the case where $w = B$, when the decision 0 is optimal; in particular, it doesn't allow a distinction that is sufficient enough to guarantee $\mathrm{O}(T^{\alpha})$ regret for the longer time horizon T for any $\alpha < 1$.

Proposition 10 *Consider the instance defined in Section 5.3.2. Define $D = 1/D_{\mathrm{KL}}(0.75\|0.25)$. Suppose that there is a sequence of K-FH policies $(\psi_T)_{T\in\mathbb{N}}$ that satisfy,*

$$\mathrm{Regret}_{K\text{-FH}}^{D\log T}(A, \psi_T) \leq \mathrm{o}(\log T),$$

as $T \to \infty$. Then

$$\mathrm{Regret}_{K\text{-FH}}^{T}(B, \psi_T) \geq \Omega(T^{\alpha}),$$

for every $\alpha < 1$ as $T \to \infty$.

Proof Let κ_T be the (possibly random) time until which the policy ψ_T chooses decisions below 0.5. Then, since the policy incurs a regret of at least 0.25 for each time step in which the decision is less than 0.5 and the parameter is $w = A$, we have that $\mathrm{Regret}_{K\text{-FH}}^{D\log T}(A, \psi_T) \geq \mathbb{E}_A(\min(\kappa_T, D\log T)) \times 0.25$. Since $\mathrm{Regret}_{K\text{-FH}}^{D\log T}(A, \psi_T) \leq \mathrm{o}(\log T)$, we have that $\mathbb{E}_A(\min(\kappa_T, D\log T)) \leq \mathrm{o}(\log T)$. Thus, for a fixed $\gamma \in (0,1)$, by Markov's inequality, we have,

$$\mathbb{P}_A\left(\min(\kappa_T, D\log T) \geq \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}\right) \leq \frac{D_{\mathrm{KL}}(0.75\|0.25) \times \mathrm{o}(\log T)}{(1-\gamma)\log T} = \mathrm{o}(1). \tag{38}$$

We thus have that,

$$\mathbb{P}_A\left(\kappa_T < \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}\right) = \mathbb{P}_A\left(\min(\kappa_T, D\log T) < \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}\right) \geq 1 - \mathrm{o}(1). \tag{39}$$

The equality follows by the definition of D and because $D\log T > D(1-\gamma)\log T$.

Next, fix a $\beta > 0$. Then by an identical argument to that leading to Equation 37 in the proof of Proposition 9, we have that

$$\mathbb{P}_A\left(\kappa_T < \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)} \text{ and } \Lambda_{\kappa_T} \geq (1 - \frac{\gamma}{1+\beta})\log T\right) \leq \mathrm{o}(1). \tag{40}$$

Thus (39) and (40) together imply that

$$\mathbb{P}_A\left(\kappa_T < \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)} \text{ and } \Lambda_{\kappa_T} < (1 - \frac{\gamma}{1+\beta})\log T\right) \geq 1 - \mathrm{o}(1). \tag{41}$$

Let C denote the event $\left\{\kappa_T < \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)} \text{ and } \Lambda_{\kappa_T} < (1 - \frac{\gamma}{1+\beta})\log T\right\}$. We then have the following inequality by the change of measure identity.

$$\mathbb{P}_A(C) = \mathbb{E}_B(\mathbf{1}_C \exp(\Lambda_{\kappa_T})) \geq 1 - \mathrm{o}(1). \tag{42}$$

Since $\Lambda_{\kappa_T} < (1 - \frac{\gamma}{1+\beta})\log T$ on the event C, we thus have that,

$$\mathbb{E}_B\left(\mathbf{1}_C \exp((1 - \frac{\gamma}{1+\beta})\log T\right) \geq \mathbb{E}_B(\mathbf{1}_C \exp(\Lambda_{\kappa_T})) \geq 1 - \mathrm{o}(1). \tag{43}$$

Thus we have,

$$\mathbb{P}_B(C) \geq \frac{1 - \mathrm{o}(1)}{\exp((1 - \frac{\gamma}{1+\beta})\log T)} = T^{-1+\frac{\gamma}{1+\beta}}(1 - \mathrm{o}(1)). \tag{44}$$

However, when $w = B$, the optimal decision is 0, and hence $\kappa_T < \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}$ implies that the per time-step regret is at least 0.25 after $\frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}$ time periods due to the FH constraint that doesn't allow the decisions to be lower than 0.5 after κ_T. Hence, we have that

$$\mathrm{Regret}^T_{K\text{-FH}}(B, \psi_T) \geq 0.25 \times \left(T - \frac{(1-\gamma)\log T}{D_{\mathrm{KL}}(0.75\|0.25)}\right)(1 - \mathrm{o}(1))\, T^{-1+\frac{\gamma}{1+\beta}} = \Omega(T^{\frac{\gamma}{1+\beta}}). \tag{45}$$

Since $\gamma \in (0,1)$ and $\beta > 0$ are arbitrary, the result follows. ■

6. Extensions and future directions

We now discuss some extensions of our proposal and comment on interesting directions for future work.

1. *Continuous parameter settings.* As we mentioned in the introduction, the assumption of a finite set of possible utility models, allows for a convenient abstraction to demonstrate the operational properties of individual fairness in hindsight in an otherwise fairly general setting. However, extensions to continuous parameter settings may be a practical necessity to operationalize this notion in specific applications. Such extensions are interesting directions for future work. We expect that the broad approach of conservative exploration then exploitation would lead to sublinear regret guarantees in these settings. However, achieving optimal regret rates may necessitate algorithmic innovation. The parametric model would typically depend on the application in question. For instance, a relevant application of notions based on individual fairness is to online personalized pricing. A recent model introduced in Ban and Keskin (2018) is as follows. Individuals with contextual information or *features* arrive over time. These contexts are represented by a sequence of feature vectors (X_t), where $X_t \in \mathbb{R}^d$. They are presented with a sequence of prices (p_t) by the seller (who is the principal in this setting) in an online fashion. The probability of a context X_t accepting a price p_t is given by

$$f(X_t'\alpha - (X_t'\beta)p_t),$$

where f is a known function (e.g., the logistic function) and $\alpha, \beta \in \mathbb{R}^d$ are unknown parameters lying in some known compact set. If a context accepts the price p_t, then the seller obtains a revenue of p_t. Traditionally (e.g., in Ban and Keskin (2018)), the goal of the seller is to maximize the expected revenue over a time horizon. In our setting, we will have the same goal while ensuring that the prices additionally satisfy fairness-in-hindsight, i.e., e.g., for some $K > 0$, and some $r > 0$,

$$p_t \leq p_{t'} + K\|X_t - X_{t'}\|_r \text{ for all } t \geq t'. \tag{46}$$

It would be interesting to characterize “good” fair-in-hindsight pricing policies that can maximize revenue.

2. *Strategic concerns.* A possible criticism of CaFE (and more generally the notion of fairness-in-hindsight) is that, since early decisions are conservative, individuals may prefer to arrive later when the algorithm has entered the “Exploit” phase. We argue that in many practical settings, individuals can only afford a limited delay from the time they require a decision to the time they approach the principal for a decision. For instance, when an individual needs a loan, it is for some time-sensitive project which cannot be delayed arbitrarily. In such situations, strategic arrival within some limited window of requiring a decision will not impact the performance of CaFE. On the other hand, there are situations where individuals may have heterogeneous costs for the delay, which may correlate with their contextual information and uncertain model parameters. In these situations, understanding the tradeoffs between individual fairness and robustness to strategic behavior are interesting directions for future work.

On a similar note, there could be heterogeneity across population groups with regard to their arrival times. For instance, when a new loan program is introduced by a bank, it could be the case that early applicants are the ones with dire financial need, and such need could potentially be correlated with membership in certain groups. In such cases, CaFE could be perceived to be unfair due to its disparate impact across groups since groups arriving earlier will receive more conservative decisions compared to those that arrive later. In the presence of such heterogeneity in arrivals across time, it seems unlikely that it is possible to learn good decisions while ensuring group fairness as well as individual fairness in hindsight, even in cases where the principal's utilities do not depend on group membership.[11] We leave such investigations of the precise tradeoffs between utility maximization and notions of group and individual fairness in dynamic contextual decision-making as an important open direction for research.

3. *Vector-valued decisions.* Extension to vector-valued decisions, i.e., where decisions lie in the set $\mathcal{X} = [0,1]^d$ for some $d > 1$, is relatively straightforward assuming that the distance between decisions is measured under the L_∞ norm. In this case, a K-Lipschitz decision rule is the one that satisfies:

$$\|\phi(c) - \phi(c')\|_\infty \leq K d_{\mathcal{C}}(c, c') \text{ for all } c,\, c' \in \mathcal{C}. \tag{47}$$

A policy is K-fair-in-hindsight for some $K > 0$ if the decisions $(\mathbf{x}_t)_{t=1,\cdots,T}$ it generates for any sequence of contexts, where $\mathbf{x}_t = (x_t^{(i)})_{i=1,\cdots,d}$, satisfy,

$$x_t^{(i)} \geq x_{t'}^{(i)} - \mathcal{K}(t-t') d_{\mathcal{C}}(c_t, c_{t'}) \text{ for all } t \geq t' \text{ and for all } i = 1, \cdots, d. \tag{48}$$

CaFE can be easily adapted to this setting by defining the decision to be ϵ on all dimensions during the Explore phase, which in turn leads to a lower bound of ϵ on decisions on each dimension during the Exploit phase. A sublinear regret bound similar to Theorem 6 can be proved for CaFE in this setting under similar assumptions.

7. Conclusion

In this paper, we proposed a new notion of fairness, fairness-in-hindsight, that extends the concept of individual fairness to account for temporal considerations. Our proposal is simple, intuitive, and importantly, we show that it assimilates well with sequential decision-making settings that involve learning, unlike the more straightforward notion of fairness-across-time. This latter aspect inspires optimism since it suggests that similar temporal fairness notions that are already embedded in our critical societal systems like law need not necessarily hinder learning of good policies over time; as we pointed out earlier, conservative exploration then exploitation structure of our fair-in-hindsight learning algorithm CaFE is already observed in these contexts. Thus, to summarize, fairness-in-hindsight can be a practical, first-order safeguard against claims of discrimination in modern algorithmic deployments.

In many settings, however, the utility model of the principal itself changes over time (i.e., what were perceived to be good policies are not good anymore), or the distance metric changes over time (i.e., contexts that seemed different are actually closer to each other). Our

[11] Relevant paper in this regard is Blum et al. (2018), which shows the incompatibility of ensuring certain notions of group fairness and achieving good classifications in the online learning problem of combining fair expert classifiers when the arrivals are adversarial.

model and results are inapplicable in such settings since we assume that the utility model and the distance metric remain fixed throughout the time horizon. Defining appropriate notions of fairness in such settings and optimizing the related utility-fairness tradeoffs is an important direction for future research in this area.

Acknowledgments

A two page extended abstract of this paper appeared in the 20th ACM Conference on Economics and Computation (EC), 2019. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing during the Real-Time Decision-Making Program (Spring 2018). The authors would like to thank the participants of the Fairness Cluster (Summer 2019) at the Simons Institute, the anonymous referees, and the area editor for thoughtful comments and valuable feedback.

References

R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. IEEE Transactions on Automatic Control, 34(3), 1989.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 2016.

G.-Y. Ban and N. B. Keskin. Personalized dynamic pricing with machine learning. Available at SSRN 2972985, 2018.

Y. Bechavod, C. Jung, and Z. S. Wu. Metric-free individual fairness in online learning. arXiv preprint arXiv:2002.05474, 2020.

L. Besson and E. Kaufmann. What doubling tricks can and can't do for multi-armed bandits. arXiv preprint arXiv:1803.06971, 2018.

A. Blum, S. Gunasekar, T. Lykouris, and N. Srebro. On preserving non-discrimination when combining expert advice. In Advances in Neural Information Processing Systems, pages 8376–8387, 2018.

L. E. Bolton, L. Warlop, and J. W. Alba. Consumer perceptions of price (un)fairness. Journal of consumer research, 29(4):474–491, 2003.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning, 5(1):1–122, 2012.

T. Calders and S. Verwer. Three naíve Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2):277–292, 2010.

L. E. Celis, S. Kapoor, F. Salehi, and N. K. Vishnoi. An algorithmic framework to control bias in bandit-based personalization. arXiv preprint arXiv:1802.08674, 2018.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2):153–163, 2017.

A. Chouldechova and M. G'Sell. Fairer and more accurate, but for whom? arXiv preprint arXiv:1707.00046, 2017.

S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023, 2018.

C. Dwork and C. Ilvento. Individual fairness under composition. Proceedings of Fairness, Accountability, Transparency in Machine Learning (FATML), 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science (ITCS), pages 214–226. ACM, 2012.

H. Elzayn, S. Jabbari, C. Jung, M. Kearns, S. Neel, A. Roth, and Z. Schutzman. Fair algorithms for learning in allocation problems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 170–179, 2019.

M. L. Friedland. Prospective and retrospective judicial lawmaking. The University of Toronto Law Journal, 24(2):170–190, 1974.

S. Gillen, C. Jung, M. Kearns, and A. Roth. Online learning with an unknown fairness metric. In Advances in Neural Information Processing Systems, pages 2600–2609, 2018.

M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, pages 3315–3323, 2016.

H. Heidari and A. Krause. Preventing disparate treatment in sequential decision making. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 2248–2254. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1617–1626, 2017.

M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. Fair algorithms for infinite and contextual bandits. arXiv preprint arXiv:1610.09559, 2016a.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In Advances in Neural Information Processing Systems, pages 325–333, 2016b.

C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu. Eliciting and enforcing subjective individual fairness. arXiv preprint arXiv:1905.10660, 2019.

F. Kamiran and T. Calders. Classifying without discriminating. In Computer, Control and Communication, pages 1–6. IEEE, 2009.

T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pages 643–650, 2011.

J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In 8th Innovations in Theoretical Computer Science Conference, ITCS, pages 43:1–43:23, 2017.

T. Lattimore and C. Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes. Calibrated fairness in bandits. arXiv preprint arXiv:1707.01875, 2017.

D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 560–568. ACM, 2008.

A. Phillips. Defending equality of outcome. Journal of political philosophy, 12(1):1–19, 2004.

J. Rawls. Justice as fairness: A restatement. Harvard University Press, 2001.

S. Ross. Stochastic processes. Wiley series in probability and statistics: Probability and statistics. Wiley, 1996. ISBN 9780471120629. URL https://books.google.com/books?id=ImUPAQAAMAAJ.

A. Sen. Equality of what? Globalization and International Development: The Ethical Issues, page 61, 2013.

L. Sweeney. Discrimination in online ad delivery. Queue, 11(3):10, 2013.

G. Yona and G. Rothblum. Probably approximately metric-fair learning. In International Conference on Machine Learning (ICML), pages 5666–5674, 2018.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, pages 1171–1180, 2017.