Temporal Aspects of Individual Fairness

Swati Gupta
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332
swatig@gatech.edu

Vijay Kamble
Dept. of Information and Decision Sciences
University of Illinois at Chicago
Chicago, IL 60607
kamble@uic.edu

Abstract

The concept of individual fairness advocates similar treatment of similar individuals to ensure equality in treatment [DHP+12]. In this paper, we extend this notion to account for the *time* at which a decision is made, in settings where there exists a notion of "conduciveness" of decisions as perceived by individuals. We introduce two definitions: (i) fairness-across-time and (ii) fairness-in-hindsight. In the former, treatments of individuals are required to be individually fair relative to the past as well as future, while in the latter we only require individual fairness relative to the past. We show that these two definitions can have drastically different implications in the setting where the principal needs to learn the utility model: one can achieve a vanishing asymptotic loss in long-run average utility relative to the full-information optimum under the fairness-in-hindsight constraint, whereas this asymptotic loss can be bounded away from zero under the fairness-across-time constraint.

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 the 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 [DHP+12, Swe13, KMR16, ALMK16, HPS+16, Cho17, CG17, CDG18] .

In this context of automated decisions, fairness is often considered in a relative sense rather than an absolute sense. In his 1979 Tanner Lectures, Amartya Sen noted that since nearly all theories of fairness are founded on an equality of some sort, the heart of the issue rests on clarifying the "equality of what?" problem [HC18, and references therein]. Equality can be desired with respect to outcomes [HPS+16], treatment [DHP+12], or even mistreatment [ZVR+17]. In this paper, we consider the equality of treatment and take the contextual (or individual) view of fairness where "similar" individuals are treated "similarly". This notion of fairness was proposed in the influential work of Dwork et al. [DHP+12] and has since been studied under several settings, e.g., see [YR18, DI18b, DI18a]. The key idea described in this work is to introduce a "Lipschitz" condition on the decisions of a classifier, such that for any two individuals x, y that are at distance $d(x, y) \in [0, 1]$, the corresponding distributions over decisions $M(x)$ and $M(y)$ are also statistically close within a distance of some multiple of $d(x, y)$.

In this work, we extend this notion of individual fairness to account for the time at which decisions are made, in settings where there exists a commonly agreed upon notion of conduciveness of decisions from the perspective of an individual; 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 to a

Preprint. Work in progress.

convict than a longer term, etc. As a motivating example, consider two similar persons A and B who have both applied for a loan at a bank. Suppose that the bank approves a substantially higher loan amount to A than to B. This would be perceived as unfair (in the colloquial sense) by B, but maybe not by A. Under the classical definition of individual fairness, this distinction is irrelevant – implicitly, it is sufficient that either of the two similar individuals find 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, for a particular individual who has arrived at some point in time, it may be reasonable to require that the treatment of this individual is perceived to be fair (again, in the colloquial sense) relative to the treatment of past individuals, but not relative to treatment of individuals that will arrive in the future. In our example above, if B applied for the loan earlier than A, then the treatment of B can still be defined to be fair as long as B 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, B'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 A later gets approved for a substantially higher amount.

Armed with this basic intuition, we introduce two definitions that extend individual fairness to incorporate the notion of time: (i) *fairness-across-time* and (ii) *fairness-in-hindsight*. In the former, treatments of individuals are required to be individually fair relative to the past as well as future, while in the latter we only require individual fairness relative to the past. In particular, fairness-in-hindsight allows decisions to become more conducive over time by providing a lower bound for rewards or an upper bound for penalties. In both these definitions, we also incorporate the possibility of having Lipschitz constants that depend on differences in time.

In our technical results, we find that these two definitions can have drastically different implications in the setting where the utility models are needed to be learned. In particular, we show that one can achieve a vanishing asymptotic loss in long-run average utility relative to the full-information optimum under the fairness-in-hindsight constraint, whereas under the more stringent constraint of fairness-across-time, the asymptotic loss can be bounded away from zero. These results complement a small but growing body of literature on learning in multi-armed bandit problems under various fairness constraints [JKMR16, LRD$^+$17, GJKR18].

2 Model

A static model. Consider a principal responsible for mapping contexts to decisions. Contexts c lie in the finite set $\mathcal{C} \subseteq \mathbb{R}^m$ and are drawn from some distribution $\mathcal{D}$ over $\mathcal{C}$. Decisions x are scalar lie in the set $\mathcal{X} = [0, 1]$. For a context c and decision x, the principal observes a random utility $U = xF$, where F is a random variable drawn from some distribution $\mathcal{F}_c$ defined on a finite set $\mathcal{S} \subseteq \mathbb{R}$. For each $c \in \mathcal{C}$, define $f(c) \triangleq \mathbb{E}_{\mathcal{F}_c}(F)$. We assume that the distribution $\mathcal{F}_c$, for each c, is known to the principal. A *decision-rule* is a function $\phi : \mathcal{C} \to \mathcal{X}$ that maps each context $c \in \mathcal{C}$ to a decision in $\mathcal{X}$.

Example 1. *Suppose the principal is a bank who is making loan approval decisions. The probability of loan default depends on the type c of the applicant belonging to the finite set of types $\mathcal{C}$. Suppose that for a type c, the probability of loan default is estimated to be $p(c)$. The decision space is $x \in [0, 1]$ representing the amount of loan sanctioned (normalized to 1). 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., $U = xF$, where,*

$$F = \begin{cases} -1 & \text{w.p. } p(c) \\ \beta & \text{w.p. } 1 - p(c) \end{cases}$$

Thus the expected utility is $\mathbb{E}(U) = -xp(c) + \beta x(1 - p(c)) = x(\beta - p(c)(1 + \beta)) = xf(c)$, where $f(c) = \beta - p(c)(1 + \beta)$.

Suppose that for any two contexts in $\mathcal{C}$, we can talk about a distance between them defined by a commonly agreed-upon distance function $d_{\mathcal{C}} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}^+$. We assume that this distance function defines a metric on $\mathcal{C}$; in particular, it satisfies the triangle inequality. Consider the following definition of an *individually fair* decision-rule in the spirit of [DHP$^+$12].

Definition 1. *[DHP$^+$12] 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 Φ_K be the space of K-Lipschitz decision-rules that map $\mathcal{C}$ to $\mathcal{X}$. The optimization problem of the principal is to choose a K-Lipschitz decision-rule that maximizes the expected utility. Define,

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

Since $\mathcal{C}$ is finite, it is easy to see that this problem can be solved as a finite linear program.

A 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 $\mathcal{C}$. The decisions of the principal, x_t at any time t, lie in the set $\mathcal{X} = [0, 1]$. For a context c_t and corresponding decision x_t, the principal obtains a random utility $U_t = x_t F_t$, where F_t is drawn from the distribution $\mathcal{F}_{c_t}$ independently of the past. The average expected utility of the principal until time T is given by $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[U_t] = \sum_{t=1}^{T} x_t f(c_t)$. A *policy* for the principal maps the sequence of contexts seen upto time $t-1$, the corresponding decisions, and the utility outcomes, to a decision $x_t \in \mathcal{X}$ for all t. 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. 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,*

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

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

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,*

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

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

Let Ψ^T be the space of all policies for a fixed horizon T, and 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. Consider the following two optimization problems for the principal.

$$P^T_{K\text{-FT}}: \max_{\psi \in \Psi^T_{K\text{-FT}}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[x_t f(c_t)] \quad \text{and} \quad P^T_{K\text{-FH}}: \max_{\psi \in \Psi^T_{K\text{-FH}}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[x_t f(c_t)]. \tag{5}$$

The expectations are with respect to the randomness in the sequence $(c_t)_{t \geq 1}$. Define the optimal values of these problems as $U^T_{K\text{-FT}}$ and $U^T_{K\text{-FH}}$, respectively. It is clear that both $U^T_{K\text{-FT}} \geq U_K$ and $U^T_{K\text{-FH}} \geq U_K$, since one can simply use the optimal K-Lipschitz decision-rule at every stage. But for small horizons, potentially, one can 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 offering more flexibility in mapping contexts to decisions. But we can show that when the horizon gets longer, one can't do any better than achieving U_K on average as defined in (2).

Proposition 2.1. *For any $K \in [0, \infty)$, $\lim_{T\to\infty} U^T_{K\text{-FT}} = \lim_{T\to\infty} U^T_{K\text{-FH}} = U_K$.*

This, in particular, shows that relaxing the fairness-across-time constraint to only requiring fairness-in-hindsight doesn't 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. In the next section, we show that the situation is drastically different when there is learning involved.

A dynamic model with learning. Now consider a situation where the distribution of the utility, given a context and a decision, is unknown to the principal and must be learned. Formally, suppose that the distribution $\mathcal{F}_c$ equals $\mathcal{G}_{c,w}$ for each $c \in \mathcal{C}$, where $w \in \mathcal{W}$ for some finite set $\mathcal{W}$. We assume that for any w and w' in $\mathcal{W}$, there is a $c \in \mathcal{C}$ such that $\mathcal{G}_{c,w} \neq \mathcal{G}_{c,w'}$. Suppose that the set $\mathcal{W}$ is known, but w is unknown to the principal and must be learned by adaptively assigning decisions to contexts and observing the outcomes. Define $g(c, w)$ to be the mean of $\mathcal{G}_{c,w}$. With some abuse of

notation, we define $U_K(w)$ to be the maximum value of the expected utility under the optimal K-Lipschitz decision rule, given w. For a large enough horizon T, for any dynamic policy in $\Psi^T_{K\text{-FT}}$ or $\Psi^T_{K\text{-FH}}$ that doesn't assume the knowledge of w, we can compare its average expected utility against the long-run optimal benchmark $U_K(w)$. Again, with some abuse of notation, for any w-agnostic policy $\psi \in \Psi^T_{K\text{-FT}}$, we denote its average utility for a fixed w as $U^T_{K\text{-FT}}(w, \psi)$. We similarly, define $U^T_{K\text{-FH}}(w, \psi)$ to be the performance of a w-agnostic policy $\psi \in \Psi^T_{K\text{-FH}}$.

In this case, it is easy to construct an example where for any $\psi \in \Psi^T_{K\text{-FT}}$, $\max_{w\in\mathcal{W}} U_K(w) - U^T_{K\text{-FT}}(w, \psi)$ is bounded away from 0 for any T large enough. A formal example is given in the appendix. Here we provide an intuition.

Example 2. *Suppose the context for each loan applicant simply denotes whether they are aged below 45 or aged above 45, and the bank does not know whether age is positively or negatively correlated with default probability. If the first applicant is aged above 45 and is given a loan of amount \$M, then any future applicant aged above 45 must be given \$M to satisfy fairness-across-time. But this decision of \$M loan is bound to be suboptimal when \$M is small but age is negatively correlated with default probability or when \$M is large and age is positive correlated with default probability.*

This shows that the FT constraint can result in significant losses in utility relative to the static optimum when learning is involved. But the situation is not as bleak under the FH constraint as we show in following main result.

Theorem 1. *Fix a $K \in [0, \infty)$. Then for every $\epsilon \in (0, 1]$, there exists a sequence of K-*FH *policies $(\psi^T_\epsilon)_{T\in\mathbb{N}}$ such that*

$$\max_{w\in\mathcal{W}} \left(U_K(w) - \liminf_{T\to\infty} U^T_{K\text{-FT}}(w, \psi^T_\epsilon) \right) \leq \epsilon|\mathcal{C}| \max_{c\in\mathcal{C};w'\in\mathcal{W}} |g(c, w')|. \tag{6}$$

The idea behind this result is to choose the following K-FH policy ψ^T_ϵ. Until a random time T' (which can be thought of as a "learning" phase) defined such that at T', the parameter w is learned with a probability of error of $1/T$, the principal chooses $x_t = \epsilon$ irrespective of c_t. We can show that $E(T') = \mathrm{O}(\log T)$ and hence the expected loss in utility relative to $U_K(w)$ in this learning phase is $\mathrm{o}(1)$. From that point on, the policy simply assumes the learned w^* to be the truth and chooses a K-Lipschitz decision rule defined on the decision space $\mathcal{X}^\epsilon = [\epsilon, 1]$, assuming $f(c) = g(c, w^*)$. The per-period expected loss in utility relative to $U_K(w^*)$ from that point on is at most $\epsilon|\mathcal{C}| \max_{c\in\mathcal{C};w'\in\mathcal{W}} |g(c, w')|$. These facts together imply the result.

Example 3. *Consider Example 2, where the bank is initially unaware whether age is positively or negatively correlated with default probability. In this case, the bank can approve a small amount of loan, say ϵ to each applicant in an initial "learning" phase. Once the bank learns the correlation structure with appropriate confidence, it can start approving loans for larger amounts as appropriate, while guaranteeing a loan of ϵ to everyone. This ensures that fairness holds in the sense of FH and there is a vanishing loss in long run average utility relative to the case where the correlation structure is known.*

3 Conclusion

Temporal aspects of fairness are highly nuanced and prominent in many areas. For example, drastically different legal decisions can be taken for similar individuals if the cases are tried at different times (think centuries). Decriminalization laws remove penalties for actions perceived as crimes in the past and such amendments are perceived as only being fair and more conducive, in line with our notion of fairness-in-hindsight.[1] Directionality of constraints on decisions in time has also been acknowledged by the law wherein laws can be applied prospectively (affect decisions of future cases) or retrospectively (affect decisions of pending or past cases) [Fri74]. In contrast, once a precedent is set by a ruling, decisions for similar contexts observed in the future must follow these precedents; whereas such rulings seldom effect past rulings [FJ08, Las13]. In the field of revenue management and pricing, price experimentation for consumers is often perceived as unfair as similar consumers can be charged very different prices in the process of exploration to learn demand (also observed by

[1] `https://www.nytimes.com/2018/09/06/world/asia/india-gay-sex-377.html`

[BBC+16]). However, if prices are slowly increased over time (for e.g. in rent control), or slowly decreased over time (for e.g. markdown sales of fashion items), this is largely deemed fair. Our work is a first step towards modeling such temporal aspects of fairness that are applicable in many such settings.

References

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There?s software used across the country to predict future criminals. *And it?s biased against blacks. ProPublica*, 2016.

[BBC+16] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. 2016.

[CDG18] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[CG17] Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*, 2017.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science conference*, pages 214–226. ACM, 2012.

[DI18a] Cynthia Dwork and Christina Ilvento. Group fairness under composition, 2018.

[DI18b] Cynthia Dwork and Christina Ilvento. Individual fairness under composition, 2018.

[FJ08] James H Fowler and Sangick Jeon. The authority of supreme court precedent. *Social networks*, 30(1):16–30, 2008.

[Fri74] Martin L Friedland. Prospective and retrospective judicial lawmaking. *The University of Toronto Law Journal*, 24(2):170–190, 1974.

[GJKR18] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *arXiv preprint arXiv:1802.06936*, 2018.

[HC18] Lily Hu and Yiling Chen. Welfare and distributional impacts of fair classification. *arXiv preprint arXiv:1807.01134*, 2018.

[HPS+16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3315–3323, 2016.

[JKMR16] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[Las13] Kurt T Lash. The cost of judicial error: Stare decisis and the role of normative theory. *Notre Dame L. Rev.*, 89:2189, 2013.

[LRD+17] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

[Ros96] S.M. Ross. *Stochastic processes*. Wiley series in probability and statistics: Probability and statistics. Wiley, 1996.

[RTA89] A Rajeev, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(3), 1989.

[Swe13] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[YR18] Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning (ICML)*, pages 5666–5674, 2018.

[ZVR+17] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

A Proofs

Proof of Proposition 2.1. 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 $u_t(c)$ 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 for any decision x taken for a context c in the past, $u_t(c) \geq x$.

First, we show that $u_t(\cdot)$ specifies a K-Lipschitz decision-rule. To see this, consider two contexts c and c' and w.l.o.g., assume that $u_t(c) \geq u_t(c')$. First, if $u_t(c) = 0$, then clearly $u_t(c) = u_t(c') = 0$. Next, if for some time $t' < t$, the context c was mapped to decision $u_t(c)$, then from the FH constraint, it follows that $u_t(c') \geq u_t(c) - Kd_{\mathcal{C}}(c, c')$. Thus $|u_t(c) - u_t(c')| \leq Kd_{\mathcal{C}}(c, c')$. 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 $x < u_t(c)$ (note again that the highest decision in the past for context c cannot be larger than $u_t(c)$). In this case, there is some other context c^* that was mapped to some decision x^* at some time in the past and $u_t(c) = x^* - Kd_{\mathcal{C}}(c^*, c)$ (since $u_t(c)$ is the tightest lower bound). But this also means that $u_t(c') \geq x^* - Kd_{\mathcal{C}}(c^*, c')$. Thus $u_t(c') - u_t(c) \geq K(d_{\mathcal{C}}(c^*, c) - d_{\mathcal{C}}(c^*, c'))$. 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 $u_t(c') \geq u_t(c) - Kd_{\mathcal{C}}(c, c')$. Thus again, $|u_t(c) - u_t(c')| \leq Kd_{\mathcal{C}}(c, c')$. This shows that $u_t(\cdot)$ is K-Lipschitz.

Now consider the decision rule ϕ_t chosen by the policy at time t. Our overall proof strategy is as follows. We will bound from above the expected utility under ϕ_t at time t by the expected utility of the decision rule $u_t(.)$ plus a side-payment. Since $u_t(.)$ is a K-Lipschitz decision-rule, the expected utility under this decision-rule is at most U_K. Additionally, we will show that over time, the total side-payments are bounded by a constant independent of T.

To see this, suppose that ϕ_t is replaced by $u_t(.)$. Now any loss in expected utility due to this switch can be compensated by a side-payment of $(\phi_t(c) - u_t(c))|f(c)|$ to the principal in the event that c arrives at time t, for each $c \in \mathcal{C}$. Moreover, if c arrives at time t, then at time $t+1$, $u_{t+1}(c) = \phi_t(c)$. Thus, the total side-payment over all arrivals of context c, irrespective of T, is at most $|f(c)|$. Thus we have an upper bound on the expected utility under any policy, equal to $TU_K + \sum_{c\in\mathcal{C}} |f(c)|$. This implies the result.

□

Proof of Theorem 1. Let $l_t(w)$ be the likelihood of $w \in \mathcal{W}$ based on observations until time t. Define $\Lambda_t(w, w') = \log(l_t(w)/l_t(w'))$. Let $w^*_t = \arg\max_{w\in\mathcal{W}} l_t(w)$ be the maximum likelihood estimate of the model parameter at time t. Fix $\epsilon \in (0, 1]$. The K-FH policy ψ^T_ϵ is defined as follows.

1. **Learning phase:** While $\min_{w\in\mathcal{W};w\neq w^*_t} \Lambda_t(w^*_t, w) \leq T$, assign $x_t = \epsilon$.
2. **Exit from learning phase:** If $\min_{w\in\mathcal{W};w\neq w^*_t} \Lambda_t(w^*_t, w) > T$ define $w^* = w^*_t$ and permanently enter the exploitation phase.
3. **Exploitation phase:** 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:

$$\max_{\phi:\mathcal{C}\to\mathcal{X}^\epsilon} \mathbb{E}[\phi(c)g(c, w^*)] \tag{7}$$

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

First, observe that this policy is 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 learning phase in the exploitation phase since decisions for each context only increase in going from learning to exploitation.

Next, for a fixed T, if we define $T' \leq T$ be the random time at which learning phase ends, then we can show that $\mathbb{E}(T') = \mathrm{O}(\log T)$ (this follows from Lemma 4.3 in [RTA89]). Moreover, if we denote $P_w(w^* \neq w)$ to be the probability of learning an incorrect model parameter w^* when the true parameter is w, then we can show that $P_w(w^* \neq w) \leq 1/T$. This follows from the fact that under the true w, the sequence of likelihood ratios $\Lambda_t(w, w')$ is a martingale and hence by Doob's martingale inequality [Ros96], $P(\max_{t \leq T} \Lambda_t(w, w') > T) \leq 1/T$ for any $w' \neq w$.

Finally, if we denote $U_K^\epsilon(w)$ to be the optimal value of the optimization problem (7) when $w^* = w$, then we can show that $U_K^\epsilon(w) \geq U_K(w) - \epsilon|\mathcal{C}| \max_{c \in \mathcal{C}; w' \in \mathcal{W}} |g(c, w')|$. 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 $\phi(c) \geq \epsilon$ and $\phi'(c) \triangleq \epsilon$ otherwise. It is easy to verify that this decision rule is K-Lipschitz and all the decisions are in $\mathcal{X}^\epsilon$; hence it is feasible in problem (7). Clearly, the expected utility of this decision rule is at least $U_K(w) - \epsilon|\mathcal{C}| \max_{c \in \mathcal{C}; w' \in \mathcal{W}} |g(c, w')|$. This implies the claim.

Thus, we finally have that for a fixed model parameter w, the total expected utility $U^T_{Q\text{-FT}}(w, \psi^T_\epsilon)$ under the Q-FH policy is at least

$$\left(1 - \frac{1}{T}\right)\left(U_K(w) - \epsilon|\mathcal{C}| \max_{c \in \mathcal{C}; w' \in \mathcal{W}} |g(c, w')|\right)(T - E(T')). \tag{8}$$

Dividing by T and taking the limit as $T \to \infty$ implies the result.

□

B Formal version of Example 2

Suppose that $\mathcal{C} = \mathcal{W} = \{0, 1\}$. $g(c, w)$ is expressed in the matrix below, where the first row (column) corresponds to $c = 0$ ($w = 0$).

$$[g(c, w)]_{\mathcal{C} \times \mathcal{W}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{9}$$

Suppose that $c_t = 0$ with probability 0.9 and $c = 1$ with probability 0.1 independently for all t. Suppose that $K = 1/2$. In this case, it is easy to see that the K-Lipschitz constraint is superfluous if w is known: if $w = 0$, then the optimal K-Lipschitz decision-rule chooses $x_t = 1$ when $c_t = 0$ and $x_t = 0$ when $c_t = 1$; if $w = 1$, then it chooses $x_t = 0$ when $c_t = 0$ and $x_t = 1$ when $c_t = 1$. Now consider a w-agnostic policy $\psi \in \Psi^T_{\text{K-TF}}$ for some large T. The average utility under this policy cannot be worse than in the setting where w is revealed to the policy by an oracle in time period 2. In this setting, suppose that if $c_1 = 0$ then the policy chooses some $x_1 \in [0, 1]$. We can show that whatever x_1 may be, it forces a long-run loss that is bounded away from 0 relative to $U_K(w)$ for at least one of the two possible w, even in the case where w is revealed by an oracle in time period 2.

To see this, note that for a fixed x_1, if w is revealed to be 0 at time 2, a near-optimal policy (for a large T) for any $t \geq 2$ under the constraint that $x_t = x_1$ if $c_t = 0$ is to map $c_t = 1$ to $x_t = 0$. If instead w is revealed to be 1 at time 2, a near-optimal policy (for a large T) under the constraint that $x_t = x_1$ if $c_t = 0$ is to map $c_t = 1$ to $x_t = 1$. The per period loss relative to $U^*_K(0)$ from time 2 onwards in the first case is $0.9 \times 1 - x_1$ and the per period loss relative to $U^*_K(1)$ from time 2 onwards in the second case is $0.9 \times x_1$. Thus in expectation over the randomness in c_1, $\max_{w \in \mathcal{W}} U_K(w) - U^T_{\text{K-TF}}(w, \psi) \gtrapprox (T-1)/T \times 0.9 \times 0.9 \times \max(x_1, 1 - x_1) = 0.9^2 \times 0.5 \times (1 - 1/T)$.