Energy Shields for Fairness

FILIP CANO, Institute of Science and Technology Austria, Austria
THOMAS A. HENZINGER, Institute of Science and Technology Austria, Austria
KONSTANTIN KUEFFNER, Institute of Science and Technology Austria, Austria

Runtime fairness is not a one-time constraint but a dynamic property evaluated over a sequence of decisions. To ensure fairness at runtime, it is necessary to account for past decisions, information neglected by conventional, static classifiers. Traditional fairness shields enforce runtime fairness abruptly, by intervening *deterministically* whenever a sequence of decisions violates the target for a running fairness measure. This motivates our *main conceptual contribution:* ***energy shields****.* An energy shield is a novel, lightweight, adaptive controller that monitors a sequence of decisions and intervenes *probabilistically* to ensure runtime fairness smoothly, by utilizing physics-inspired energy functions to nudge the sequence toward fairness: the more unfair the decisions, the stronger the nudging force becomes. This makes energy shields the ***first*** fairness shields to provide both *short-term safety and long-term liveness guarantees.* Safety ensures that the running fairness measure stays within a running target interval with high probability, and liveness ensures that the limit of the fairness measure lies within the limit target interval. Intuitively, the short-term specifies the tolerated fairness values and the long-term specifies the desired fairness values. We also provide a synthesis procedure for constructing the least intrusive energy shield for a given target specification, and demonstrate its efficiency experimentally. We evaluate our energy shields against existing fairness shields through the lens of short- and long-term fairness.

ACM Reference Format:
Filip Cano, Thomas A. Henzinger, and Konstantin Kueffner. 2026. Energy Shields for Fairness. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 33 pages. https://doi.org/10.1145/3805689.3806807

1 Introduction

Algorithmic decision-making is ubiquitous in modern life, from hiring and lending to online advertising. In these settings, binary decisions, e.g., approving or denying a loan or displaying one ad over another, are often made sequentially [19, 37]. Much research has focused on designing *statically fair* algorithms, which ensure fairness in expectation over a fixed distribution [16]. This guarantees that in the long term the decision sequence will be fair (if the distribution remains static). However, this perspective fails to address unfair behavior that arises in the short term from the cumulative history of decisions, which can be arbitrarily biased even for a fair classifier [13]. We start to illustrate our work using a simplified binary online advertisement setting, which we generalize later.

Example 1.1. Consider a static online advertising setting where two companies bid for advertising space on an online platform. The company A places a bid of value c_A and the company B places a bid of value c_B. At each point in time $t \in \mathbb{N}$, a user accesses the platform and will be shown an ad. The choice of ad X_t is made probabilistically $X_t \sim \text{Bernoulli}(p)$ by a decision maker, if the realization x_t evaluates to 1, the ad from company A will be selected. Intuitively, the decision maker is fair if p matches the bid ratio, i.e., $p \in c_A/(c_A + c_B) \pm \varepsilon$ for $\varepsilon > 0$.

Authors' Contact Information: Filip Cano, Institute of Science and Technology Austria, Klosterneuburg, Austria, filip.cano@ist.ac.at; Thomas A. Henzinger, Institute of Science and Technology Austria, Klosterneuburg, Austria, tah@ist.ac.at; Konstantin Kueffner, Institute of Science and Technology Austria, Klosterneuburg, Austria, konstantin.kueffner@ist.ac.at.

This work is licensed under a Creative Commons Attribution 4.0 International License.
FAccT '26, Montreal, QC, Canada
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3806807

In a sequential setting, a sequence $x_1, \ldots, x_t$ of ad placements is fair at time t, if the empirical running average decision $\mu(x_1, \ldots, x_t) = \frac{1}{t}\sum_{i=1}^{t} x_i$ lies within a target interval centered around the bid ratio $c_A/(c_A + c_B) \pm \varepsilon$ for $\varepsilon > 0$. This translates into a short-term objective where the average is required to remain in the interval most of the time with high-probability, and a long-term objective where the limit average is required to converge to a point in the target interval.

Running vs. limit fairness: safety and liveness. Ex. 1.1 illustrates two *fairness* properties defined w.r.t. a running and a limit target interval. The short-term property requires the running target interval to be met by the running average at a finite time, and the long-term property requires the limit target to be met by the limit average of the decision sequence. Intuitively, the running target expresses what fairness values we tolerate in the short-term, and the limit target expresses what fairness values we desire in the long-term. This distinction matches safety-liveness classification of properties studied in formal verification [7]. The violation of a safety property can be determined at a finite time, thus matching our short-term requirement for fairness, called *running fairness*. The satisfaction of a liveness property can be determined in the limit only, thus matching our long-term requirement for fairness, called *limit fairness*. Our objective is to ensure that a decision sequence meets given target intervals for both running and limit fairness.

Traditional vs. energy shielding. To enforce dynamic fairness requirements, *runtime shielding* has emerged as a promising approach [14]. Originating in formal verification, a shield is a lightweight controller that monitors a decision sequence and intervenes minimally to correct decisions in order to enforce desired fairness requirements [5]. Unfortunately, existing fairness shields act deterministically, alternating abruptly between full intervention and no intervention, to enforce highly restrictive short-term fairness requirements, and they fail to provide long-term fairness guarantees altogether [14]. We introduce probabilistic shielding mechanisms that have short- and long-term guarantees.

Example 1.2 (Ex. 1.1 cont.). Suppose both companies bid equally, resulting in a target ratio of 0.5 and we allow a tolerance of 0.1. A shield will observe the sequence of the decision maker one decision at a time. After each new observation, the shield must decide, based on the history, whether to keep or flip the original decision. For example, if the average decision risks exceeding 0.6 and the current decision is 1. A naive classical shield would deterministically flip the decision to remain within the target interval, thus switching from being completely idle to being maximally invasive. An energy shield would randomly flip the decision based on a chosen enforcement probability, thus it can increase its intervention strength with the unfairness of the past decision average.

Group fairness. Group fairness metrics, such as demographic parity, equal opportunity, equalized odds, and others, quantify the fairness between demographic groups in binary decision-making by comparing the group conditional acceptance probabilities [39]. For example, demographic parity ensures that the ratio of positive decisions is independent of the group. The motivational Example 1.1 considers a setting where decisions are mutually exclusive, i.e., accepting company A implies rejecting company B. In this setting, the empirical demographic parity degenerates to the arithmetic mean. In Section 7 we demonstrate that our energy shields generalize to the general setting, as motivated below.

Example 1.3 (Ex. 1.2 cont.). As in Example 1.1 the decision maker chooses between company A or company B. Afterwards, a second decision maker places the chosen ad in a regular ($S = R$) or a premium ($S = P$) ad space. This setting admits a second fairness measure between companies based on the premium ad placement rate, the empirical equivalent of demographic parity, i.e., $\mathbb{P}(S = P \mid \text{Company} = A) - \mathbb{P}(S = P \mid \text{Company} = B)$.

Generalizations. The motivational Example 1.1 considers a static setting where the companies bid a constant amount at every point in time. The setting of this example is ideal for explaining the main concepts in energy shielding, and will be the main focus in Section 3 to 6. In Section 9 we study two generalizations that may be

useful for realistic settings: unknown and dynamic settings. In the *unknown* setting, the decision probability is fixed, but not communicated to the shield. The shield can estimate it from previous decisions and adapt the enforcement effect to the current parameter estimation. In the *dynamic* setting, the decision probability may be subject to change over time (for example, when the decision maker is retrained). Here, energy shields remain useful, although with weaker formal guarantees.

Contributions. Our main *conceptual contribution* is the introduction of ***energy shields***, a probabilistic shielding framework inspired by physics-based energy functions to ensure both running and limit fairness requirements. An energy shield smoothly nudges a sequence of decisions toward fairness by assigning higher "energy" to unfair sequences and intervening, proportional to the energy, to enter a lower-energy state. As a consequence, our *energy shields are the **first** fairness shields capable of providing short- and long-term fairness guarantees* To demonstrate that energy shields provide **short-term fairness** guarantees, we provide exponentially decaying tail bounds on the probability and expected value of running fairness violations with respect to a given target interval. To demonstrate that energy shields provide **long-term fairness** guarantees, we characterize the limit behavior of the shielded process, deriving conditions on the energy function to ensure convergence of the fairness measure to a given target. Additionally, we quantify the long-run expected cost of intervention and prove that steeper energy functions yield fewer violations, an important monotonicity property. We exploit the monotonicity property of energy functions to propose a **synthesis procedure** that combines binary search and dynamic programming with tail bounds to find the least intrusive energy shield satisfying the desired running and limit fairness properties. We validate the effectiveness and efficiency of the synthesis procedure experimentally. Moreover, we benchmark our energy shields against the seminal fairness shields by [14], empirically supporting the claim that energy shields are the first shields to provide both short- and long-term guarantees. As the bulk of our contribution is theoretical, we focus on presenting the results and leave the proofs in the appendix.

2 Related Work

Most algorithmic fairness research focuses on fairness in a static setting. This includes: measures for the fairness of a decision maker at the level of groups [23, 27] and individuals [21]; pre-, in-, and post-processing techniques to synthesize fair decision maker [1, 25, 27, 45, 47]; verification techniques to check whether static decision makers are fair [3, 9, 24, 36, 40, 44]. Among the existing techniques for static fairness, our shields could be classified as a post-processing technique. The key difference is that those methods modify decision-makers *once* before deployment, whereas our work addresses fairness *during* deployment via runtime intervention.

We are not the first to be concerned with algorithmic fairness over a sequence of decisions [2, 13]. A large body of work focuses on detecting unfair behavior at runtime, both for individual [26] and for group fairness [4, 10, 28]. Beyond detection, [14] is the only work enforcing fairness at runtime. Their shields adopt the sequential fairness definition from Parand et al. [2], ensuring that a sequence of decisions will be fair with probability 1 at predefined periodic intervals. Our shields soften this condition, providing high-probability short-term guarantees and are the first to provide limit guarantees. Although the monitoring and enforcement of fairness has only recently emerged as a topic of interest, classical runtime monitoring and enforcement have long been studied in the runtime verification community, with monitors [6, 8, 20, 22, 38, 43] and shields [5, 15, 18] developed for Linear Temporal Logic specifications. Shielding has also been explored in probabilistic settings [31, 46], using probabilistic model checking techniques [33]. We build on stochastic approximation results [12, 32] to design probabilistic shields with almost sure convergence guarantees.

3 Setting

In this section, we formally introduce the setting, the fairness properties, and the notion of a shield.

Decision process. We model the setting in Ex. 1.1 using a single coin with decision probability $p \in [0,1]$. At each point in time $t \in \mathbb{N}$ the coin is tossed, resulting in a decision $x_t \in \mathbb{B}$, where $\mathbb{B} = \{0,1\}$, which is the realization of the random variable $X_t \sim \text{Bernoulli}(p)$. Combined they generate the decision process $X = (X_t)_{t\in\mathbb{N}}$ of i.i.d. Bernoulli(p) random variables. A realization $x = (x_t)_{t\in\mathbb{N}} \in \{0,1\}^\omega$ of X is an infinite sequence of binary values.

Fairness. We are interested in measuring the fairness of the process defined above. The measure we use is called average outcome fairness [13]. Formally, given an infinite sequence of binary decisions $x \in \{0,1\}^\omega$, e.g., a realization of the decision process, we measure the fairness of a finite prefix $x_{1:t} = (x_1, \dots, x_t)$ as its average decision, denoted by $\mu(x_{1:t})$, or μ_t if clear from the context. If it exists, $\mu(x)$ denotes the fairness measure of the realization in the limit:

$$\mu(x_{1:t}) = \frac{1}{t}\sum_{i=1}^{t} x_i, \qquad \text{and} \qquad \mu(x) = \lim_{t\to\infty} \mu(x_{1:t}). \tag{1}$$

A *fairness target* is a tuple $\varphi = (\tau, \mathcal{S}, \mathcal{L})$ consisting of a burn-in time $\tau \in \mathbb{N}$, a running target $\mathcal{S} \subseteq [0,1]$, and a limit target $\mathcal{L} \subseteq \mathcal{S} \subseteq [0,1]$. The fairness target specifies the acceptable fairness measure values at every finite time greater than the burn-in τ and at the limit. For the target to be satisfiable we require that the burn-in time τ is sufficiently large to account for initial high variance (see Section 11 for further discussion on how to choose the burn-in time parameter). Given a fairness target, an infinite sequence $x \in \mathbb{B}^\omega$ satisfies:

- *point fairness* at time $t \in \mathbb{N}$, if the fairness measure is in the running target at t, i.e., $\mu(x_{1:t}) \in \mathcal{S}$;
- *running fairness*, if point fairness is always satisfied after the burn-in τ, i.e., $\forall t \geq \tau\colon \mu(x_{1:t}) \in \mathcal{S}$;
- *limit fairness*, if the fairness measure is in the limit target in the limit, i.e., $\lim_{t\to\infty} \mu(x_{1:t}) \in \mathcal{L}$;
- *fairness*, if both running fairness and limit fairness are satisfied.

Intuitively, the running target expresses what fairness values are tolerated in the short-term, and the limit target expresses what fairness values are desired in the long-term.

Example 3.1 (Ex. 1.1 cont.). Assume the bid ratio is 0.5. In the long-term we require a fairness measure of 0.5, and in the short-term we accept a tolerance of 0.1 after some burn-in τ. The corresponding fairness target is $\varphi = (\tau, [0.4, 0.6], \{0.5\})$. Note, the decision X_t follows a Bernoulli distribution, thus the fairness measure $\mu(X_{1:t})$ follows a binomial distribution scaled by $1/t$. Since $(1/t)Bin(t,p) \xrightarrow{t\to\infty} p$, limit fairness requires $p = 0.5$. The probability of satisfying point fairness at t is $\mathbb{P}[\mu(X_{1:t}) \in \mathcal{S}] = \mathbb{P}[Bin(t, 0.5) \in t(0.5 \pm 0.1)] \approx \sum_{i=\lfloor 0.4t \rfloor}^{\lceil 0.6t \rceil} \binom{t}{i} p^i (1-p)^{t-i}$.

Shielding. As illustrated in Ex. 3.1, without control of p, the only possible intervention on the process is to overwrite individual decisions. This is called shielding. A deterministic shield for a decision process is a program with the power to flip the decisions made at runtime. Formally, a deterministic shield $\mathfrak{S}\colon (\mathbb{B}\times\mathbb{B})^* \times \mathbb{B} \to \mathbb{B}$ uses the history of decisions $x_1, \dots, x_t \in \mathbb{B}^t$ at time $t \in \mathbb{N}$ and the history of intervention $y_1, \dots, y_{t-1} \in \mathbb{B}^{t-1}$ to compute the next intervention y_t. The intervention indicates whether the decision x_t is flipped, i.e., if $y_t = 1$ then the shielded decision is $z_t = 1 - x_t$, otherwise the shielded decision is $z_t = x_t$. The shield should aid the satisfaction of the fairness target, evaluated over the sequence shielded decisions $z = (z_t)_{t\in\mathbb{N}}$ with as little interference as possible. This is quantified by the average interference cost $v_t \coloneqq \sum_{i=1}^{t} y_i/t$ over a intervention sequence $y_1, \dots, y_t$.

Example 3.2 (Ex. 3.1 cont.). A trivial shield guaranteeing running and limit fairness, ensures that the company ads alternate, i.e., the shielded decision sequence is $(01)^\omega$, thus $\mu(z_{1:2t}) = 0.5$ for all t. A less restrictive shield that satisfies running fairness, but not the limit fairness, is one that only interferes if the point fairness is about to be violated. We call this shield the *naive shield*. For example, assume at $t = 100$ we have $\mu(z_{1:t}) = 40/100$, if $x_{t+1} = 0$, then the fairness measure would be $40/101 < 0.4$, so the shield enforces, i.e., $z_{t+1} = 1$.

Probabilistic shields. Deterministic shields act aggressively and abruptly once the fairness measure is at risk of leaving the target and remain idle most of the time. This leads to two regimes, one where the shield has full control over the decision and one where it has none. We introduce probabilistic shields, i.e., shields in which interventions are executed probabilistically, allowing for a much more gentle approach to shielding. Formally, a probabilistic shield is a function $\mathfrak{S}\colon (\mathbb{B}\times\mathbb{B})^*\times\mathbb{B}\to[0,1]$ mapping a history of decisions $x_1,\dots,x_t\in\mathbb{B}^t$ at time $t\in\mathbb{N}$ and the history of interventions $(y_1,\dots,y_{t-1})$ into an nudging probability $q_t=\mathfrak{S}(x_1,y_1,\dots,y_{t-1},x_t)$, defining the distribution from which the next intervention is sampled, i.e., $Y_t\sim\text{Bernoulli}(q_t)$.

At each time step t, we distinguish three binary random variables: X_t, the original decision produced by the environment; Y_t, the intervention indicator chosen by the shield; and Z_t, the released shielded decision. $Y_t=0$ means that the shield keeps the original decision, while $Y_t=1$ means that it flips it. Therefore, it holds that

$$Z_t=(1-X_t)\cdot Y_t+X_t\cdot(1-Y_t).$$

Problem definition. We consider four processes: the decision process $X=(X_t)_{t\in\mathbb{N}}$, the intervention process $Y=(Y_t)_{t\in\mathbb{N}}$, the shielded decision process $Z=(Z_t)_{t\in\mathbb{N}}$, and the shielded fairness process $M=(M_t)_{t\in\mathbb{N}}$. The processes are defined at every time $t\in\mathbb{N}$ as follows:

$$X_t\sim\text{Bernoulli}(p),\quad Y_t\sim\text{Bernoulli}(\mathfrak{S}(X_1,Y_1,\dots,Y_{t-1},X_t)),$$
$$Z_t=(1-X_t)\cdot Y_t+X_t\cdot(1-Y_t),\quad\text{and}\quad M_t=\mu(Z_1,\dots,Z_t).$$

Intuitively, the shielded fairness process, i.e., the fairness measure evaluated over the shielded decision process, should satisfy a given fairness target $\varphi=(\tau,\mathcal{S},\mathcal{L})$ with an error probability less than $\delta\in(0,1)$, i.e.,

$$\mathbb{P}(\forall t\geq\tau\colon M_t\in\mathcal{S})\geq 1-\delta,\quad\text{and}\quad\mathbb{P}\left(\lim_{t\to\infty}M_t\in\mathcal{L}\right)\geq 1-\delta\quad\text{for }\delta\in(0,1).$$

Because both the environment and the shield are stochastic, the short-term specification can only be guaranteed with high probability. We use $\delta\in(0,1)$ to denote the tolerated failure probability, equivalently a confidence level of $1-\delta$. Smaller δ yields stronger guarantees, but typically leads to more conservative shields. Like the burn-in time τ, the failure tolerance δ is an application-dependent design parameter.

4 Energy-based Shields

We introduce energy shields, a family of physics inspired probabilistic shields. The shields acts by computing an energy state for the process and nudging the system toward lower-energy configurations. The energy state is given by an energy function and the current fairness measure.

Energy function. An energy function is bowl-shaped with minimum at its pivot point.

Definition 4.1. An *energy function* with *pivoting point* $\kappa\in[0,1]$ is a function $\zeta\colon[0,1]\to[0,1]$ satisfying: (i) ζ is continuously differentiable, i.e., ζ is differentiable, and ζ' is continuous; (ii) $\zeta'(c)\leq 0$ for $c<\kappa$, and $\zeta'(c)\geq 0$ for $c>\kappa$; (iii) $\zeta(\kappa)=\zeta'(\kappa)=0$, and $\zeta(c)>0$ for $c\in\{0,1\}$.

Example 4.2. Two families of energy functions are even-polynomials and exponential energy functions (see Fig. 1) defined, as $\zeta^{\text{Pol}}_{\kappa,\alpha,\beta}(x)=\alpha|x-\kappa|^\beta$ and $\zeta^{\text{Exp}}_{\kappa,\rho,\sigma}(x)=\rho(1-e^{-\sigma(x-\kappa)^2})$, where $\kappa\in(0,1)$, $\beta\in(1,\infty)$, $\alpha\in(0,1/\max(\kappa,1-\kappa)^\beta$, $\sigma\in(0,\infty)$, and $\rho\in(0,1/(1-e^{-\sigma(\min\{\kappa,1-\kappa\})^2}))$. The parameter ranges ensure that the energy function does not exceed 1 without clipping.

Energy shield. An energy shield is defined w.r.t. an energy function $\zeta\colon[0,1]\to[0,1]$ with pivoting point $\kappa\in[0,1]$. The pivot point determines the favored decision, while the energy function determines the nudging probability. Formally, assume we have observed the decisions $x_1,\dots,x_t$ and accumulated the interventions $y_1,\dots,y_{t-1}$, which determined the shielded decisions $z_1,\dots,z_{t-1}$ at time $t\in\mathbb{N}$. Then if $\mu_{t-1}\leq\kappa$, the shield

accepts $x_t = 1$ and flips $x_t = 0$ with probability $\zeta(\mu_{t-1})$, and if $\mu_{t-1} > \kappa$, the shield accepts $x_t = 0$, and flips $x_t = 1$ with probability $\zeta(\mu_{t-1})$. This determines the distribution of the next shielded decision Z_t and fairness value M_t.

Claim 1 (Shielded decision process). *A decision process Z shielded by an energy shield forms a sequence of Bernoulli random variables with evolving bias, i.e, $Z_t \sim \text{Bernoulli}(p_t)$. The biases are defined recursively as $p_1 = p$ and $p_{t+1} = f(\mu_t)$ for a given history $z_1, \dots, z_t$, where*

$$f(\mu) = \begin{cases} p + (1-p)\zeta(\mu) & \text{if } \mu \le \kappa, \\ p \cdot (1 - \zeta(\mu)) & \text{if } \mu > \kappa \end{cases}. \tag{2}$$

Moreover, the resulting shielded fairness process update can be written as

$$M_t = \mu_{t-1} + \frac{1}{t}(Z_t - \mu_{t-1}) \quad \text{(with } \mu_0 = 0\text{)}. \tag{3}$$

Intuitively, Eq. 2 shows that the decision maker pulls the fairness value μ toward p, while the shield exerts an opposing pull toward κ. Stronger energy values amplify this effect, shifting the process further to κ. Eq. 3 makes this dynamic explicit: the sequence of μ_t's evolves as a stochastic approximation process drifting in the direction of $\mathbb{E}[Z_{t+1}|\mu_t] - \mu_t = f(\mu_t) - \mu_t$. At a convergence point, this drift should vanish, which occurs at a fixpoint of f, i.e., a value of μ satisfying $f(\mu) = \mu$.

5 Short-term: Safety Guarantees

For the short-term running fairness property we require the shielded fairness process to stay within the running target $\mathcal{S} \subseteq [0, 1]$ at all finite times after the burn-in τ (see Ex. 3.1). To prove that our shield satisfies running fairness with high-probability we develop upper bounds on the probability and the expected number of point fairness violations over an interval. For a target $\mathcal{S} = [L, U]$, we set the burn-in to $\tau_{\mathcal{S}} = 4/\min(|L - \mu^*|, |U - \mu^*|)$.

Definition 5.1. Let M be the shielded fairness process generated by the decision probability p and an energy shield with energy function ζ. For a running target interval $\mathcal{S} \subseteq [0, 1]$, we define two violation measures, the probability of violating point fairness $\mathcal{P}_{\mathcal{S}}$ and the expected number of point fairness violations $\mathcal{E}_{\mathcal{S}}$ over a time interval $[T, T'] \subseteq \mathbb{N} \cup \{\infty\}$ as

$$\mathcal{P}_{\mathcal{S}}(M_{T:T'}) = \mathbb{P}\left[\exists t \in [T, T'] \colon M_t \notin \mathcal{S}\right] \quad \text{and} \quad \mathcal{E}_{\mathcal{S}}(M_{T:T'}) = \sum_{t=T}^{T'} \mathbb{E}\left[\mathbf{1}[M_t \notin \mathcal{S}]\right]. \tag{4}$$

5.1 Upper-bounds

Our main result is a bound for the probability of the shielded fairness process violating point fairness at time T. This result follows from constructing a martingale that upper-bounds the distance of the process to its converging point μ^*, and using Azuma-Hoeffding's inequality on said martingale.

Theorem 5.2. *Let $\mathcal{S} = [L, U]$ such that $\kappa, p \in \mathcal{S}$, and $K = 1/32$. Then for every $t \ge \tau_{\mathcal{S}}$*

$$\mathbb{P}[M_t \notin \mathcal{S}] \le \exp\left(-Kt|L - \mu^*|^2\right) + \exp\left(-Kt|U - \mu^*|^2\right). \tag{5}$$

Property bounds. Utilizing Theorem 5.2, the expected number of violations and the probability of a single violation in an interval, follows from Boole's inequality.

Corollary 5.3. *Let $r_L = e^{-K|L-\mu^*|^2}$, $r_U = e^{-K|U-\mu^*|^2}$. For every $T, T' \in \mathbb{N}$ s.t. $\tau_{\mathcal{S}} \le T < T'$, then*

$$\mathcal{E}_{\mathcal{S}}(M_{T:T'}) \le \sum_{t=T}^{T'} (r_L^t + r_U^t), \qquad \text{and} \qquad \mathcal{E}_{\mathcal{S}}(M_{T:\infty}) \le \frac{r_L^T}{1 - r_L} + \frac{r_U^T}{1 - r_U}.$$

Hence, it follows that $\mathcal{P}_{\mathcal{S}}(M_{T:T'}) \le \mathcal{E}_{\mathcal{S}}(M_{T:T'})$ for $T' \in \mathbb{N} \cup \{\infty\}$ s.t. $T' \ge T$.

5.2 Monotonicity with Respect to the Energy Function

In this section, we define a partial order among energy functions given by their steepness, and show how our violation properties are monotone with respect to steepness.

Definition 5.4. Let ζ_1, ζ_2 be two energy functions with a common pivot point κ. We say that ζ_1 is *steeper* than ζ_2, and denote it by $\zeta_1 \succeq \zeta_2$ if $\zeta_1(y) \geq \zeta_2(y)$ for all $y \in [0,1]$.

Intuitively, a steeper energy function (see Fig. 1) constitutes a more "aggressive" shield, leading to fewer safety violations, both in probability and in expectation.

Theorem 5.5. *Let $\zeta_1 \succeq \zeta_2$ be two energy functions, and let M^{ζ_1} and M^{ζ_2} be the shielded fairness process generated by enforcing the decision process with ζ_1 and ζ_2, respectively. For all $T, T' \in \mathbb{N}$ such that $\tau_{\mathcal{S}} \leq T < T'$ we have*

$$\mathcal{E}_{\mathcal{S}}\left(M^{\zeta_1}_{T:T'}\right) \leq \mathcal{E}_{\mathcal{S}}\left(M^{\zeta_2}_{T:T'}\right), \qquad \text{and} \qquad \mathcal{P}_{\mathcal{S}}\left(M^{\zeta_1}_{T:T'}\right) \leq \mathcal{P}_{\mathcal{S}}\left(M^{\zeta_2}_{T:T'}\right).$$

6 Long-term: Liveness Guarantees

For the long-term limit fairness property, we require the shielded fairness process to converge to a point in the limit target $\mathcal{L} \subseteq [0,1]$. To prove that our shield satisfies limit fairness, we identify the conditions on the energy function to ensure that the shielded fairness process converges to a target value μ^* under the decision probability p with an expected average interference cost of $|p - \mu^*|$.

Main result. It is remarkable that both the convergence of M_t and the expected cost depend *only* on p and the value of ζ at the target point μ^* and not on the pivot point κ.

Theorem 6.1. *Let $\mu^* \in [0,1]$. Given the shielded decision process from Eq. 3, with bias p and energy function ζ, the shielded fairness process $(M_t)_{t\in\mathbb{N}}$ converges almost surely (a.s.) to μ^* if and only if*

$$\zeta(\mu^*) = \begin{cases} (\mu^* - p)/(1-p) & \text{if } p < \mu^*, \\ (p - \mu^*)/p & \text{otherwise} \end{cases}. \tag{6}$$

Furthermore, the expected interference cost $(\mathbb{E}[v_t])_{t\in\mathbb{N}}$ converges almost surely to $|\mu^* - p|$.

Proof intuition. We establish that the shielded fairness process $M = (M_t)_{t\in\mathbb{N}}$ converges at the fixpoint of f. First, we show f has indeed a unique fixpoint located between p and the pivot κ.

Lemma 6.2. *The function $f\colon [0,1] \to [0,1]$ defined as in Eq. 2 is continuously differentiable, and has a unique point $\mu^* \in [0,1]$ such that $f(\mu^*) = \mu^*$. Furthermore, μ^* sits between p and κ.*

Once we know f has a unique fixpoint, we use stochastic approximation theory to prove that the shielded fairness process M converges almost surely to the fixpoint.

Lemma 6.3. *The fairness process, defined in Eq. 3, converges a.s. to the unique fixpoint μ^* of f, as defined in Eq. 2. The error $(M_t - \mu^*)^2$ converges a.s. at the rate of $o(1/t^\lambda)$ for all $\lambda \in (0,1)$.*

Proof sketch. Let $g(x) = f(x) - x$, then we rewrite the update rule in Eq. 3 as

$$M_{t+1} = M_t + \gamma_t\big(g(M_t) + \xi_{t+1}\big), \tag{7}$$

where $\xi_t = Z_t - f(M_{t-1})$, and $\gamma_t = 1/t$. This is a classical Robbins-Monro form for stochastic approximation [12]. From stochastic approximation theory we know that, under certain regularity conditions, M_t from Eq. 7 approximates the zero value of g, which is fixpoint of f. We use [32] to bound the convergence rate. □

Equation 6 in Theorem 6.1 follows from Lemma 6.3 by noticing that the fixpoint μ^* has to satisfy Eq. 8. Since the fixpoint lies between p and κ, the branch in Eq. 8 is chosen based on $\mu^* \leq p$.

$$\mu^* = \begin{cases} p + (1-p) \cdot \zeta(\mu^*) & \text{if } p \leq \kappa, \\ p \cdot (1 - \zeta(\mu^*)) & \text{if } p > \kappa \end{cases}. \tag{8}$$

The expected cost of intervention at step $t+1$ is the probability of not seeing the favorable decision (1 if the current average is below κ and 0 otherwise) and having an energy high enough to intervene. The expected intervention cost $h(\mu_t)$ converges $h(\mu_t) \to h(\mu^*)$, because the fairness value converges $\mu_t \to \mu^*$, where h is defined as

$$h(\mu) = \begin{cases} (1-p) \cdot \zeta(\mu) & \text{if } \mu \leq \kappa \\ p \cdot \zeta(\mu) & \text{otherwise} \end{cases}. \tag{9}$$

Lemma 6.4. *For the process described in Eq. 3 , the corresponding sequence of average interference $(v_t)_{t \in \mathbb{N}}$ converges to $h(\mu^*)$, where μ^* is the fixpoint of f (Eq. 2) and h is as defined in Eq. 9.*

Finally, Eq. 6 and Lemma 6.4 imply that the expected intervention cost converges to $|p - \mu^*|$.

7 Energy Shields for Group Fairness

We apply our fairness shields to enforce the difference between two decision probabilities, which is necessary to enforce common group fairness metrics. Group fairness metrics quantify the fairness between demographic groups in binary decision-making by comparing the group conditional acceptance probabilities [39]. For example, demographic parity ensures that the ratio of positive decisions is independent of the group.

$$\mathbb{P}(X = 1 \mid \text{Group} = A) - \mathbb{P}(X = 1 \mid \text{Group} = B) \in [-\varepsilon, +\varepsilon] \quad \text{with tolerance of } \varepsilon > 0. \tag{10}$$

The energy shields introduced so far focus on a setting where the decision maker must accept exactly one of the groups at each point in time (see Sec. 3). We generalize this in the following.

Two-group setting. We assume that at each point in time $t \in \mathbb{N}^+$ only one group $g_t \in \mathbb{G} = \{A, B\}$ is presented to the decision maker with probability $r_{g_t} \in (0,1)$, where $r_A + r_B = 1$, which subsequently makes the decision $x_t \in \{0,1\}$ with group conditional decision probability $p_{g_t} \in (0,1)$. We use $d = p_A - p_B$ to denote the parameter that will have an analogous role as p in the single-group setting. The objective is to control the empirical group conditional acceptance rate, which is defined for every sequence of group-decision pairs $w_{1:t} = (g_1, x_1), \ldots, (g_t, x_t)$ of length $t \in \mathbb{N}^+$ as

$$\mu_A(w_{1:t}) - \mu_B(w_{1:t}) \quad \text{where} \quad \mu_g(w_{1:t}) \coloneqq \frac{\sum_{i=1}^{t} x_i \cdot \mathbf{1}[g_i = g]}{\sum_{i=1}^{t} \mathbf{1}[g_i = g]} \quad \text{for } g \in \mathbb{G}.$$

In this setting a shield is a function $\mathfrak{S} \colon (\mathbb{G} \times \mathbb{B} \times \mathbb{B})^* \times (\mathbb{G} \times \mathbb{B}) \to [0,1]$ is a mapping from a group-decision-intervention history $(g_1, x_1, y_1), \ldots, (g_{t-1}, x_{t-1}, y_{t-1})$ and an input $(g_t, x_t) \in \mathbb{G} \times \mathbb{B}$ to a nudging probability q_t, defining the distribution from which the next intervention y_t is sampled. Formally, we consider the group process $G = (G_t)_{t \in \mathbb{N}}$, the decision process $X = (X_t)_{t \in \mathbb{N}}$ the intervention process $Y = (Y_t)_{t \in \mathbb{N}}$, the shielded input process $Z = (Z_t)_{t \in \mathbb{N}}$, and the shielded fairness process $M = (M_t)_{t \in \mathbb{N}}$ defined as

$$G_t \sim \text{Bernoulli}(r_A), \quad X_t \sim \text{Bernoulli}(p_{G_t}), Y_t = \text{Bernoulli}(\mathfrak{S}(G_1, X_1, Y_1, \ldots, Y_{t-1}, G_t, X_t))$$
$$Z_t = (1 - X_t) \cdot Y_t + X_t \cdot (1 - Y_t), \quad M_t = \mu_A(G_1, Z_1, \ldots, G_t, Z_t) - \mu_B(G_1, Z_1, \ldots, G_t, Z_t)$$

Given a probability $\delta \in [0,1]$ and a fairness target $\varphi = (\tau, \mathcal{S}, \mathcal{L})$, with $\tau \in \mathbb{N}^+$ and $\mathcal{L} \subseteq \mathcal{S} \subseteq [-1, +1]$, the objective is to find a shield that guarantees: $\mathbb{P}(\forall t \geq \tau \colon M_t \in \mathcal{S}) \geq 1 - \delta$, and $\mathbb{P}(\lim_{t \to \infty} M_t \in \mathcal{L}) \geq 1 - \delta$.

Implementation of the energy shield. We extend the domain of the energy function in Def. 4.1 from $[0,1]$ to $[-1,+1]$, i.e., $\zeta\colon [-1,+1] \to [0,1]$. The shield monitors the evolution of the shielded fairness process M. At each time t, the current fairness value is M_t. As before, the shield favors the decision that reduces the distance between M_t and κ. This is always possible because when the group is A accepting increases and rejecting decreases the fairness measure. This is inverted for the group B. Moreover, the shield flips the original decision with probability $\zeta(M_t)$, if this decision increases the distance to κ. This results in similar long-term guarantees as in Thm. 6.1.

Theorem 7.1. *Let $\mu^* \in [-1,+1]$. The two-group shielded fairness process $(M_t)_{t\in\mathbb{N}}$ converges a.s. to μ^** iff

$$\zeta(\mu^*) = \begin{cases} (\mu^* - (p_A - p_B))/(1 - (p_A - p_B)) & \text{if } p_A - p_B < \mu^*, \\ (p_A - p_B - \mu^*)/(1 + p_A - p_B) & \text{otherwise.} \end{cases} \tag{11}$$

The expected interference cost $(\mathbb{E}[v_t])_{t\in\mathbb{N}}$ converges to a value smaller than $|\mu^* - (p_A - p_B)|$.

For the short-term guarantees, the results depend on a burn-in condition for both groups. Using $N_{g,t} := \sum_{i=1}^{t} \mathbf{1}[G_i = g]$, we express the condition with the following event:

$$E_T := \left\{ \forall t \geq T :\ N_{A,t} \geq \tfrac{r_A}{2} t \ \text{ and } \ N_{B,t} \geq \tfrac{r_B}{2} t \right\}.$$

Lemma 7.2. *For every $\eta \in (0,1)$, let $\tau_G(\eta) := \frac{8}{r_{\min}} \log \frac{4}{\eta}$. Then for all $T \geq \tau_G(\eta)$, it holds that $\mathbb{P}(E_T) \geq 1 - \eta$.*

Theorem 7.3 (2-group analogue of Thm. 5.2). *Fix an interval $I = [L,U] \subseteq [-1,1]$ such that $\kappa, d \in I$ and let $\eta \in (0,1)$. Let $K = \min\{r_A, r_B\}/32$, then for all $t \geq \tau_G(\eta)$*

$$\mathbb{P}(M_t \notin I) \ \leq \ \eta \ + \ \exp\!\left(-K\, t\, |L - \mu^*|^2\right) \ + \ \exp\!\left(-K\, t\, |U - \mu^*|^2\right). \tag{12}$$

Theorem 7.4 (2-group analogue of Thm. 5.5). *Let $\zeta_1 \succeq \zeta_2$ be two energy functions on $[-1,1]$ with common minimum at κ. Fix $I = [L,U] \subseteq [-1,1]$ and let M^{ζ_1}, M^{ζ_2} be the corresponding 2-group shielded fairness processes. Then for every $\eta \in (0,1)$, for all $T < T'$ with $T \geq \tau_G(\eta)$,*

$$\mathcal{E}_I\left(M^{\zeta_1}_{T:T'} \mid E_T\right) \ \leq \ \mathcal{E}_I\left(M^{\zeta_2}_{T:T'} \mid E_T\right), \quad \text{and} \quad \mathcal{P}_I\left(M^{\zeta_1}_{T:T'} \mid E_T\right) \ \leq \ \mathcal{P}_I\left(M^{\zeta_2}_{T:T'} \mid E_T\right).$$

8 Shield Synthesis

In this section, we state the energy shield synthesis problem and propose Alg. 1 as a solution. The synthesis procedure is virtually identical for the simple and general group fairness setting introduced in Section 3 and Section 7 respectively.

Problem instance. A problem instance $(\mathcal{V}, \Xi, \varphi, \delta, \varepsilon, \boldsymbol{p})$ consists of: a violation measure $\mathcal{V} \in \{\mathcal{P}, \mathcal{E}\}$, a totally ordered set of energy functions $\Xi = \{\zeta_k\}_{k\in R}$ indexed by some interval $R \subset \mathbb{R}$ where for all $i \leq j \in R$ we have $\zeta_i \preceq \zeta_j$, a fairness target $\varphi = (\tau, \mathcal{S}, \mathcal{L})$, a violation threshold $\delta > 0$, and an approximation tolerance $\varepsilon > 0$; a tuple $\boldsymbol{p}$ of setting parameters. The simple setting in Section 3 is described by decision probability $\boldsymbol{p} = (p)$ and requires energy functions with domain $[0,1]$. The general setting in Section 7 is described by the group appearance probability and the group specific decision probabilities $\boldsymbol{p} = (r_A, p_A, p_B)$ and requires energy function with domain $[-1,1]$.

Problem 1. *Given $(\mathcal{V}, \Xi, \varphi, \delta, \varepsilon, \boldsymbol{p})$ find the least invasive energy function $\zeta \in \Xi$ satisfying the fairness constraints:*
(i) the shielded fairness process M^ζ with parameter $\boldsymbol{p}$ satisfies $\mathcal{V}_\mathcal{S}(M^\zeta_{\tau:\infty}) \leq \delta$ and $\lim_{t\to\infty} M^\zeta_t \in \mathcal{L}$ a.s., and
(ii) ζ is ϵ-minimal, i.e., if $\zeta^* \in \Xi$ is the smallest valid element, then $|\mathcal{V}_\mathcal{S}(M^{\zeta^*}_{\tau:\infty}) - \mathcal{V}_\mathcal{S}(M^\zeta_{\tau:\infty})| \leq \epsilon$.

We remark that if the violation measure is $\mathcal{P}$, then running fairness should be satisfied with probability greater than $1-\delta$, and if it is $\mathcal{E}$, then the total number of point fairness violations after τ should be bounded by δ. In Alg. 1

we show a synthesis procedure based on having a family of energy functions $\Xi = (\zeta_r)_{r\in R}$ that is indexed by some interval $R \subset \mathbb{R}$ and is monotonic with respect to the index. In App. B we describe such a family $(\zeta^{\text{Mon}}_{r;\boldsymbol{p},\mathcal{S},\mathcal{L}})_{r\in(0,1)}$, which is defined w.r.t. $\boldsymbol{p}$ and a specification φ, as a piecewise exponential and polynomial function.

Algorithm 1 Shield synthesis

Require: problem instance $(\mathcal{V}, \Xi, p, \varphi, \delta, \varepsilon)$.

```
l ← inf R; u ← sup R                          ▷ Set lower and upper bound for energy function w.r.t. ≺.
T_DP ← min{t ∈ ℕ | Bound(t) ≤ ε}              ▷ smallest t s.t. prob. bound (Thrm. 5.2&7.3) satisfies tolerance
if Condition(ζ_u, T_DP) > δ then return Fail   ▷ The most strict energy function is not enough
while l ≠ u do
    m ← (l + u)/2 ;   d ← Condition(ζ_m, T_DP)
    if |d − δ| < ϵ then return ζ_m
    if d ≤ δ then l ← m  else  u ← m
```

Algorithm. The algorithm exploits monotonicity of the energy-function family to identify the least steep function that satisfies the violation condition, yielding the least invasive shield. Inside the binary search, it approximates the violation measure $\mathcal{V}$ via Condition, which combines dynamic programming with the tail bounds. Concretely, we split $[\tau, \infty)$ into a prefix $[\tau, T_{\text{DP}}]$ and a tail (T_{DP}, ∞). The threshold is chosen using probability bounds from Thm. 5.2 or 7.3, for the simple and general setting, respectively. For the simple setting, Cor. 5.3 gives directly $\text{Bound}(t) \coloneqq \frac{r_-^t}{1-r_-} + \frac{r_+^t}{1-r_+}$. For the general setting, in line with Thm. 7.3, we need to add a term η to the bound, indicating that $\mathbb{P}(E_T)$ is large enough. A standard choice would be to set $\eta = \varepsilon/2$, and the rest of the bound to be smaller than $\varepsilon/2$. For the prefix, we compute the exact violation measure with standard dynamic programming techniques. For the tail term, we use again setting specific tail bounds from Thrm. 5.2 or 7.3. Since the bounds for a violation in the tail can be made arbitrarily small for large enough T_{DP}, we use them to approximate the exact violation value with as much precision as required—a time-precision trade-off.

9 Generalizations

The assumption of a known and stationary decision probability p made in Section 3, might not hold in reality. We generalize to setting with unknown and/or changing decision probability. Specifically, we consider a setting where (i) "p is fixed and unknown" and (ii) "$(p_t)_{t\in\mathbb{N}}$ changes and is unknown". The setting "$(p_t)_{t\in\mathbb{N}}$ changes and is known" is less interesting as it reduces to the shield setting the decision bias to a point within a target.

Setting (i): p is fixed and unknown. This setting requires the estimation of the unknown bias p using a consistent estimate $\hat{P}_t = \mu_t(X_1, \ldots, X_t)$ at each point in time $t \in \mathbb{N}^+$, which is used to update the energy function instead of p. Because $\hat{P}_t$ converges almost surely to p, the energy function converges and enforces the long-term target.

Theorem 9.1. *Let $\mu^* \in [0,1]$ be a fairness target. Let $(\zeta_q)_{q\in[0,1]}$ be a family of energy functions s.t.*

$$\zeta_q(\mu^*) = \begin{cases} (\mu^* - q)/(1-q) & \text{if } q < \mu^*, \\ (q - \mu^*)/q & \text{otherwise.} \end{cases} \tag{13}$$

Then the shielded process that takes at each step the energy function $\zeta_{\hat{p}_t}$ converges a.s. to μ^*.

Setting (ii): $(p_t)_{t\in\mathbb{N}}$ changes and is unknown. This setting makes it impossible for a non-trivial shield to enforce the convergence to a target point. However, an energy shield with sufficiently steep energy function ensures that the shielded process remains within a target interval almost surely.

Theorem 9.2. *Let $\zeta\colon [0,1] \to [0,1]$ be an energy function, and L, R be points satisfying: $L < \kappa$, $\zeta(L) = L$, $R > \kappa$, and $\zeta(R) = 1-R$. Then, for the shielded process $(M_t)_{t\in\mathbb{N}}$ it holds a.s. that: $\liminf_{t\to\infty} M_t \geq L$ and $\limsup_{t\to\infty} M_t \leq R$.*

10 Experiments

We evaluate the impact of energy shields on the fairness across all settings, explore the time-precision trade-off in Alg. 1, and compare against the existing fairness shields from Cano et al. [14] using MacBook with the M2 chip.

10.1 Impact of Energy Functions

Here we investigate how different energy functions affect running and limit fairness, and what happens if we go beyond the assumption of a known and stationary decision probability p (see Section 9).

Energy functions. In Fig. 1 we consider a fair decision maker with $p = 0.5$ in R1, and a biased decision maker $p = 0.65$ in R2. We set our fairness target $\varphi = (0, [0.4, 0.6], \{0.5\})$, as indicated by the dotted lines in C3. We deploy the shield with a polynomial ζ^{Pol} and exponential ζ^{Exp} energy functions (see Ex. 4.2), as well as an energy function that induces an idle shield ζ^{Idle}, which is always 0, and an energy function that deterministically intervenes ζ^{Naive}, which is 0 in the interior of the running target and 1 elsewhere. The energy function parametrization depends on the decision maker and is depicted in C1. In C2 we can observe the effect of the energy function on the decision probability, where the fixpoint of the functions is indicated by the 45° line. In C3 we depict a lower and upper bound containing 95% of the simulated runs, and in C4 we depict the accumulated average violations

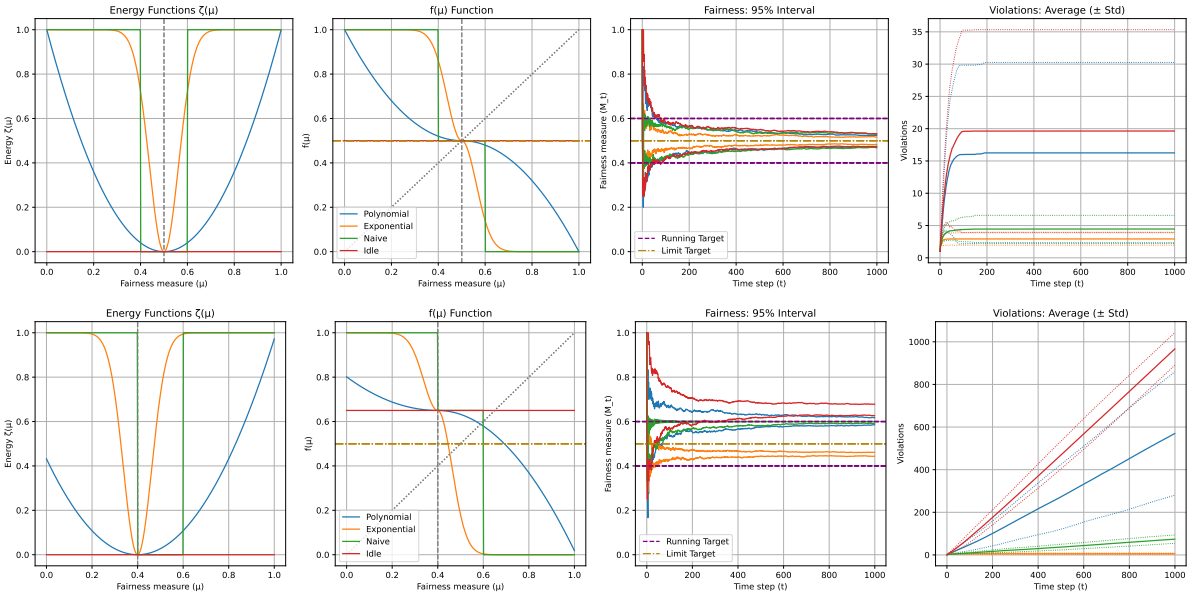


Fig. 1. Simulated impact of the energy functions ζ for fairness target $\varphi = (0, [0.4, 0.6], \{0.5\})$, decision probability p. **Rows (R):** (R1) $p = 0.5$ and $\zeta \in \{\zeta^{\text{Naive}}, \zeta^{\text{Idle}}, \zeta^{\text{Pol}}_{0.5,4,2}, \zeta^{\text{Exp}}_{0.5,1,128}\}$; (R2) considers $p = 0.65$ and $\zeta \in \{\zeta^{\text{Naive}}, \zeta^{\text{Idle}}, \zeta^{\text{Pol}}_{0.4,2.7,2}, \zeta^{\text{Exp}}_{0.4,1,128}\}$. **Columns (C):** (C1) the behavior of the energy function for each fairness value. (C2) illustration of the characteristic functions f and their respective fixpoints $f(\mu) = \mu$. (C3) 95% confidence interval of fairness values at each time. (C4) Average point fairness violations with standard deviation at each time. The simulation results are averaged over 1000 simulations for each p and ζ.

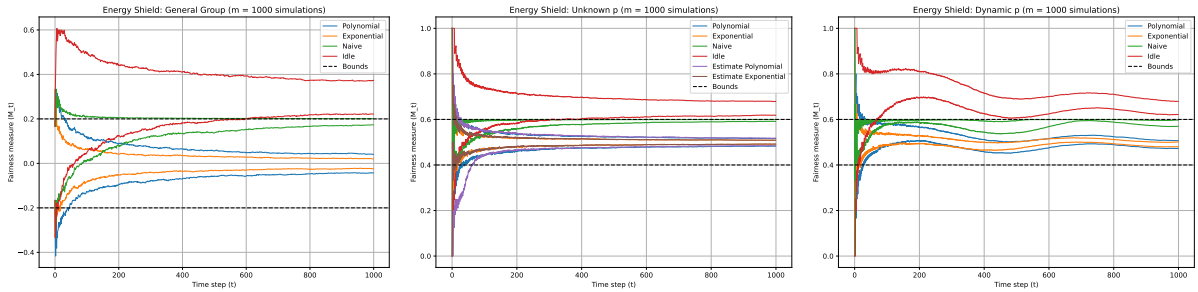


Fig. 2. 95% confidence interval of fairness values at each time for the generalized settings. **Columns (C):** (C1) two-group group-fairness setting with $\boldsymbol{p} = (r_A = 0.8, p_A = 0.7, p_B = 0.4)$ and κ chosen such that $\mu^* = 0.5$. (C2) unknown fixed decision probability ($p = 0.65$) using an online estimate $\hat{p}_t$ compared to known p. (C3) unknown time-varying decision probability p_t (sinusoidal around 0.65).

and standard deviation. We observe in C3 and C4 that if the fixpoint is contained within the running target, the number of violations decays quickly; if the decision maker is biased, a pivot within a given target interval does not guarantee convergence of the fairness value to a point in the target; and if the decision maker is fair, our shields remain useful by further reducing violations.

Generalizations. In Fig. 2 we present one experiment for each generalization for the energy functions $\zeta \in \{\zeta^{\text{Naive}}, \zeta^{\text{Idle}}, \zeta^{\text{Pol}}_{0.5,4,2}, \zeta^{\text{Exp}}_{0.5,1,128}\}$ and depict the 95% interval of the simulated runs (similar to C3 in Fig. 1). (C1) We consider the general group fairness setting with $\boldsymbol{p} = (r_A = 0.8, p_A = 0.7, p_B = 0.4)$, and choose κ such that $\mu^* = 0.5$. We observe that both the polynomial and exponential energy shield satisfy the running and limit fairness guarantees in the general group setting, while the naive shield only satisfies the running fairness guarantees. (C2) We consider the unknown decision probability setting with $p = 0.65$. Here the κ adapts to the estimated decision probability. We observe a slightly slower convergence for shields using the estimated decision probability compared to shields having access to the ground truth. (C3) We consider the unknown and dynamic decision probability setting with p_t following a sine curve centered around 0.65 and κ is set such that $\mu^* = 0.5$ w.r.t. 0.65. In this setting, our energy shields empirically aid in the satisfaction of the running and limit fairness goals.

10.2 Shield Synthesis

Here we investigate the accuracy and efficiency of the dynamic-programming algorithm and the tail-bound approximation used in synthesis?

Approximation precision. The synthesis Algorithm 1 performs a violation probability approximation (VPA) by running dynamic programming (DP) up to a horizon T_{DP} and then utilizing the infinite horizon tail-bound from Sec. 5. Because the tail-bound is conservative this induces a precision-time trade-off, which is investigated in Fig. 3. In C1, we increase the DP threshold T_{DP} from 0 to 15000 and record the precision of the obtained tail bound. In C2, we demonstrate the increase in computation time as the DP threshold T_{DP} increases. In C3, we plot the error that is obtained between a direct computation and the tail bound. In C4, we demonstrate the time-precision trade-off directly, i.e., as we decrease the precision how much does the computation time decrease. We observe that an extension of the DP horizon T_{DP} leads to an exponential gain in VPA precision, at the cost of a quadratic increase in computation time. This impacts the runtime of Alg. 1, as higher precision requirements demand longer DP horizons. The culprit is the looseness of the tail-bounds over the short horizon, as demonstrated by the gain in precision, if larger burn-ins are considered.

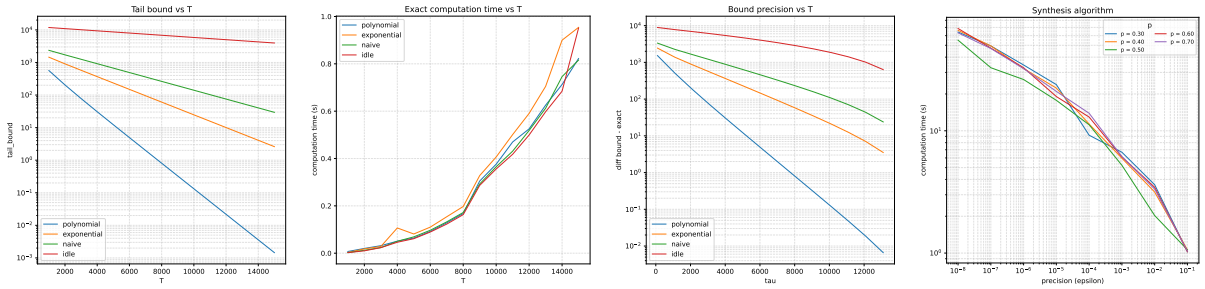


Fig. 3. Time-precision trade-off in the violation probability approximation (VPA) of Alg. 1 with fairness target $\varphi = (100, [0.3, 0.7], \{0.45, 0.55\})$, decision probability $p = 0.65$, and energy function $\zeta \in \{\zeta^{\text{Naive}}, \zeta^{\text{Idle}}, \zeta^{\text{Pol}}_{0.4,2.7,2}, \zeta^{\text{Exp}}_{0.4,1,128}\}$. **Columns (C):** (C1) VPA with increasing dynamic programming (DP) threshold T_{DP}. (C2) Computation time as T_{DP} increases. (C3) VPA precision, i.e., VPA with $T_{\text{DP}} = 0$ compared to $T_{\text{DP}} = 15000$, for increasing burn-ins τ. (C4) Runtime of Alg. 1 as precision ε increases, for different p and $\delta = 0.1$.

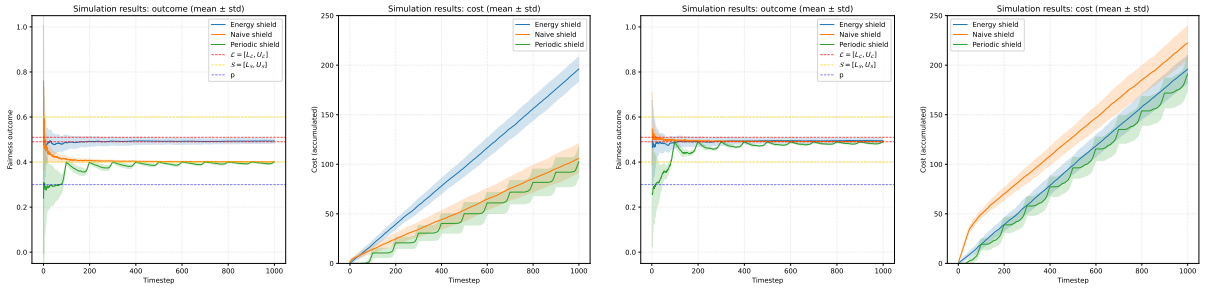
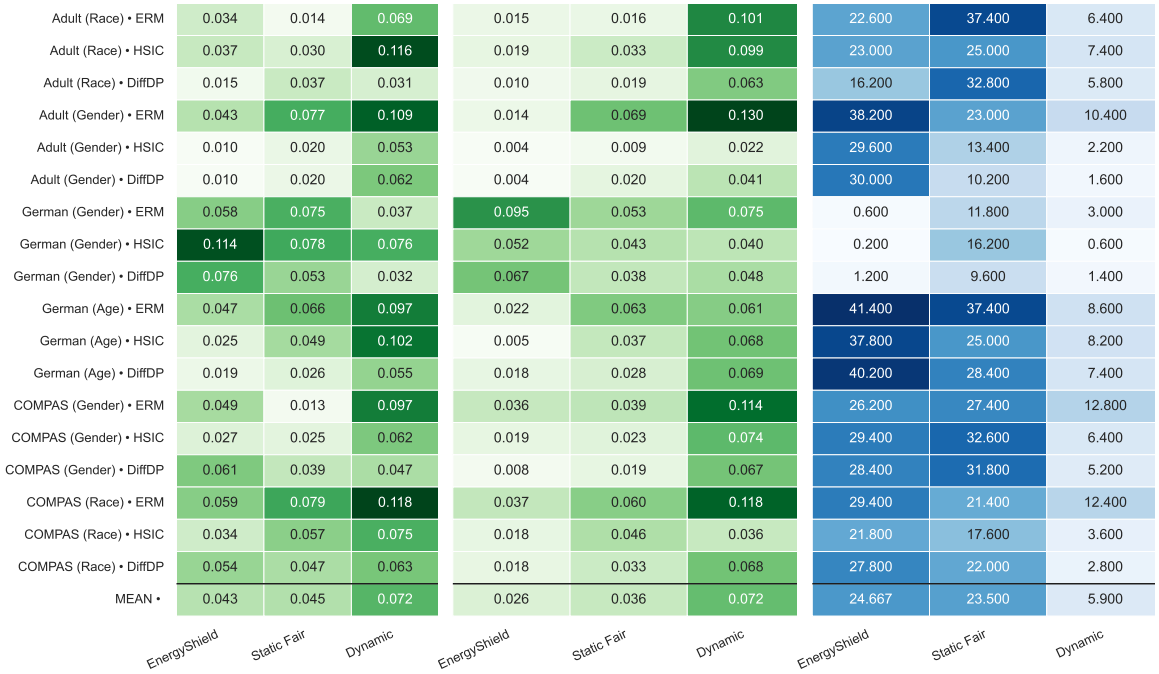


Fig. 4. Energy shields compared against naive and periodic shields, for fairness target $\varphi = (\tau = 100, \mathcal{S} = [0.4, 0.6], \mathcal{L} = [0.49, 0.51])$ and decision probability $p = 0.3$. As energy function, we use $\zeta^{\text{Mon}}_{0.1;p,\mathcal{S},\mathcal{L}}$. **Columns (C):** The naive and periodic shields are tuned for the: (C1&2) running target $[0.4, 0.6]$; (C3&4) limit target $[0.49, 0.51]$. (C1&C3) depict the fairness measure and (C2&C4) the accumulated number of interventions.

10.3 Comparison with existing Fairness Shields

Here we investigate how energy shields compare with existing fairness shields in the simple one group and the general two group setting?

Comparison: Simple. In Fig. 4, we benchmark our energy shields against the naive shield from Ex. 3.2 and the periodic shields from [14] in the simple one-group setting of Section 3. The naive shield enforces running fairness by intervening whenever the next decision would leave the running target, while the periodic shield enforces point fairness at fixed periodic times with optimal expected cost, but requires heavier runtime computation. We consider the fairness target $\varphi = (\tau = 100, \mathcal{S} = [0.4, 0.6], \mathcal{L} = [0.49, 0.51])$ and decision probability $p = 0.3$. As energy function, we use $\zeta^{\text{Mon}}_{0.1;p,\mathcal{S},\mathcal{L}}$, tuned to push the fairness value toward $\mu^* = 0.5$. The existing shields do not natively distinguish between running and limit fairness. Therefore, we evaluate them in two configurations: once with respect to the running target $\mathcal{S} = [0.4, 0.6]$, and once with respect to the stricter limit target $\mathcal{L} = [0.49, 0.51]$. We observe that, when tuned to the running target, the baseline shields keep the fairness value inside $\mathcal{S}$, but tend to remain close to the boundary and do not converge to the limit target. When tuned to the limit target, the baselines satisfy the stricter fairness requirement only by intervening much more aggressively; in particular, the



(a) Fairness at 1/3 of the run (b) Fairness at the end of the run (c) Number of Interventions

Table 1. Group-fairness experiments on COMPAS, Adult Income, and German Credit comparing energy shields against `StaticFair` and `Dynamic` fairness shields [14] under the target $\mathcal{S} = [-0.15, 0.15]$ and $\mathcal{L} = [-0.075, 0.075]$. **Columns (C):** (C1) demographic parity at one-third of the run, (C2) demographic parity at the end of the run, and (C3) total number of interventions. Darker values indicate worse performance, in terms of either demographic parity or number of interventions.

naive shield becomes almost maximally invasive. By contrast, the energy shield balances the two requirements: it keeps the fairness value within the running target while gradually steering it toward the limit target. This is reflected in the intervention costs: running-target baselines intervene less, limit-target baselines intervene more, and the energy shield lies between these two extremes while achieving the desired long-term behaviour.

Comparison: General. In Table 1, we benchmark our energy shields against the fairness shields from [14] in the general two-group setting of Section 7. We consider standard benchmarks from the fairness literature, namely COMPAS [34], Adult Income [11], and German Credit [30]. For each dataset, we evaluate decision makers trained with empirical risk minimization (ERM), HSIC [41], and Differential Demographic Parity [17]. At runtime, individuals are sampled sequentially from the test dataset, the trained classifier produces the original decision, and the shield may intervene by flipping the decision. We compare against the `StaticFair` and `Dynamic` shields from [14]. Following the same scale as in [14], we use a time horizon of 50 and 10 repetitions, so each value in Table 1 represents the result after processing 500 individuals. The previous shields are configured with a demographic parity target of 0.15. For our energy shields, we use the same value as the running specification, i.e., $\mathcal{S} = [-0.15, 0.15]$, and impose the stricter limit specification $\mathcal{L} = [-0.075, +0.075]$. Thus, the running target specifies the tolerated short-term disparity, while the limit target specifies the desired long-term disparity. Table 1 reports demographic parity after one third of the run, demographic parity at the end of the run, and the total

number of interventions. We observe that energy shields achieve better final demographic parity than their counterparts, with a number of interventions comparable to `StaticFair`. The `Dynamic` shield often intervenes less, but it does so by keeping the process close to the boundary of the running target. By contrast, energy shields continue to push the fairness value toward the stricter limit target, explaining their final fairness performance.

11 Discussion

Safety and liveness. Formal verification classifies properties over infinite traces as safety or liveness [29, 35]. A safety property asserts that "bad" things never occur, e.g., the car never crashes, thus its violation can be observed on a finite prefix. A liveness property asserts that "good" things will eventually happen, e.g., the car reaches its goal, thus its satisfaction can only be determined in the limit, and every finite trace can be extended to an infinite satisfiable trace. For stochastic processes one evaluates the satisfaction probability of the safety or liveness property w.r.t. the law of the process [7]. We deliberately connected short- and long-term fairness to safety and liveness respectively. The short-term running fairness property requires the fairness measure to remain within the running target at all times, which implies that once violated it remains violated along the infinite sequence, i.e., safety. The long-term limit fairness property requires the limit of the fairness measure to lie in the limit target, which implies it can only be determined in the limit, i.e., liveness. The convergence in the limit holds because the fairness measure is an average. This normalization allows the distance to μ^* to shrink and the process to converge almost surely. This implies that almost surely the fairness measure eventually remains within the limit target. It does not imply that the fairness measure eventually remains within the limit target almost surely, i.e., $\mathbb{P}(\exists \tau \in \mathbb{N} \forall t \geq \tau \colon M_t \in \mathcal{L}) = 1$ is satisfied, and $\exists \tau \in \mathbb{N} \colon \mathbb{P}(\forall t \geq \tau \colon M_t \in \mathcal{L}) = 1$ is not.

Burn-in. The choice of burn-in is both a value and a technical decision: from a value perspective, τ should be chosen to obtain a meaningful representation system behavior, e.g., a college admissions task force may introduce a fairness shield and wait two years before evaluating its impact. from a technical perspective, τ has to be large enough to have some confidence on fairness measure being inside the target interval. For example, if we are observing individuals being accepted (1) or rejected (0), and we have a target acceptance rate between 0.6 and 0.7, after only four candidates, the closest acceptance rate the shield can get is either 0.5 or 0.75. This is because after four candidates there can be 0, 1, 2, 3 or 4 accepts in total, i.e., all possible values of the fairness metric are in $\{0/4, 1/4, 2/4, 3/4, 4/4\}$. Hence, our target can only be achieved if enough decisions have been made. The burn-in time quantifies this concept of "enough": the shield is only required to satisfy the short-term fairness requirements after decisions.

12 Conclusion

We introduced *energy shields*, a physics-inspired framework for enforcing fairness at runtime. Energy shields are lightweight, probabilistic mechanisms that provide rigorous short- and long-term fairness guarantees. Utilizing a bowl-shaped energy function, enforcing fairness is reduced to an energy minimization problem, which enables gentle, adaptive interventions. We provide a synthesis algorithm based on binary search and dynamic programming to find the least-invasive shield for a given specification. The experimental validation demonstrates the practicality of our synthesis procedure and supports, in the comparison with [14], the claim that energy shields are the first shields to satisfy both short- and long-term guarantees. Future work will explore multi-group fairness, properties beyond fairness, and dynamic environments. We establish a foundational theory of energy-based fairness shielding, extending these guarantees to multi-group fairness and properties beyond fairness remains an exciting direction.

Endmatter

Generative AI Usage Statement

The authors used ChatGPT 5.2 for the following tasks. For writing, ChatGPT was used to find and correct grammar and spelling mistakes, and to assess the fluency, consistency, and clarity of the text. No substantial amount of text is AI generated. For the theoretical results, ChatGPT was used to brainstorm proof strategies and to search the literature for potentially relevant results. All proofs were then independently checked and fully formalized by the authors. No generative AI was used to develop the paper's core scientific contributions. For the experimental evaluation, ChatGPT was used to generate and refactor small portions of experimental code, specifically in the generation of plots. All code generated code was tested and validated by the authors.

Ethical Considerations Statement

This work is strongly motivated by the problem of ensuring algorithmic fairness in machine learning applications. Although this is a sensitive topic, the ethical considerations for this paper are minimal. Our contributions are largely theoretical, and the datasets we employ are standard benchmarks widely used in the literature already.

Acknowledgements

This work has been supported by the European Research Council under Grant No.: ERC-2020-AdG 101020093.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*. PMLR, 60–69.

[2] Parand A. Alamdari, Toryn Q. Klassen, Elliot Creager, and Sheila A. Mcilraith. 2024. Remembering to Be Fair: Non-Markovian Fairness in Sequential Decision Making. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 235. 906–920.

[3] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V Nori. 2017. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017), 1–30.

[4] Aws Albarghouthi and Samuel Vinitsky. 2019. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 211–219.

[5] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*.

[6] C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. 2003. Model-checking algorithms for continuous-time Markov chains. *IEEE Transactions on Software Engineering* 29, 6 (2003), 524–541.

[7] Christel Baier and Joost-Pieter Katoen. 2008. *Principles of model checking*. MIT press.

[8] Ezio Bartocci, Jyotirmoy Deshmukh, Alexandre Donzé, Georgios Fainekos, Oded Maler, Dejan Ničković, and Sriram Sankaranarayanan. 2018. Specification-based monitoring of cyber-physical systems: a survey on theory, tools and applications. In *Lectures on Runtime Verification*. Springer, 135–175.

[9] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. 2019. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages* 3, OOPSLA (2019), 1–27.

[10] Jan Baumeister, Bernd Finkbeiner, Frederik Scheerer, Julian Siber, and Tobias Wagenpfeil. 2025. Stream-Based Monitoring of Algorithmic Fairness. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 60–81.

[11] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[12] Vivek S. Borkar. 2008. *Stochastic approximation: a dynamical systems viewpoint*. Springer.

[13] Filip Cano, Thomas A. Henzinger, and Konstantin Kueffner. 2025. Algorithmic Fairness: A Runtime Perspective. In *Proceedings of the International Conference on Runtime Verification (RV)*. 1–21.

[14] Filip Cano, Thomas A. Henzinger, Bettina Könighofer, Konstantin Kueffner, and Kaushik Mallik. 2025. Fairness Shields: Safeguarding against Biased Decision Makers. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 15 (2025), 15659–15668.

[15] Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. 2023. Safe reinforcement learning via shielding under partial observability. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 14748–14756.

[16] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *Comput. Surveys* (2020).

[17] Ching-Yao Chuang and Youssef Mroueh. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

[18] Filip Cano Córdoba, Alexander Palmisano, Martin Fränzle, Roderick Bloem, and Bettina Könighofer. 2023. Safety shielding under delayed observation. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 33. 80–85.

[19] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*. 525–534.

[20] Alexandre Donzé and Oded Maler. 2010. Robust satisfaction of temporal logic over real-valued signals. In *International Conference on Formal Modeling and Analysis of Timed Systems*. Springer, 92–106.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[22] Peter Faymonville, Bernd Finkbeiner, Maximilian Schwenger, and Hazem Torfah. 2017. Real-time stream-based monitoring. *arXiv preprint arXiv:1711.03829*.

[23] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 259–268.

[24] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. 2021. Justicia: A stochastic SAT approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 35. 7554–7563.

[25] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning (ICML)*. PMLR, 2357–2365.

[26] Ashutosh Gupta, Thomas A. Henzinger, Konstantin Kueffner, Kaushik Mallik, and David Pape. 2025. Monitoring Robustness and Individual Fairness. In *KDD 2025*.

[27] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3315–3323.

[28] Thomas A. Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik. 2023. Runtime Monitoring of Dynamic Fairness Properties. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 604–614.

[29] Thomas A Henzinger, Nicolas Mazzocchi, and N Ege Saraç. 2023. Quantitative Safety and Liveness.. In *FoSSaCS*. 349–370.

[30] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[31] Nils Jansen, Bettina Könighofer, Sebastian Junges, Alex Serban, and Roderick Bloem. 2020. Safe Reinforcement Learning Using Probabilistic Shields (Invited Paper). In *CONCUR (LIPIcs, Vol. 171)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 3:1–3:16.

[32] Rajeeva Laxman Karandikar and Mathukumalli Vidyasagar. 2024. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *Journal of Optimization Theory and Applications* 203, 3 (2024), 2412–2450.

[33] Joost-Pieter Katoen. 2016. The probabilistic model checking landscape. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science*. 31–45.

[34] Lauren Kirchner, Surya Mattu, Jeff Larson, and Julia Angwin. 2016. Machine Bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[35] Leslie Lamport. 1977. Proving the correctness of multiprocess programs. *IEEE transactions on software engineering* 2 (1977), 125–143.

[36] Yannan Li, Jingbo Wang, and Chao Wang. 2023. Certifying the Fairness of KNN in the Presence of Dataset Bias. In *International Conference on Computer Aided Verification (CAV)*. Springer.

[37] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[38] Oded Maler and Dejan Nickovic. 2004. Monitoring temporal properties of continuous signals. In *International Symposium on Formal Techniques in Real-Time and Fault-Tolerant Systems*. Springer, 152–166.

[39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[40] Anna Meyer, Aws Albarghouthi, and Loris D'Antoni. 2021. Certifying Robustness to Programmable Data Bias in Decision Trees. *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), 26276–26288.

[41] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. 2017. Fair Kernel Learning. In *Machine Learning and Knowledge Discovery in Databases (KDD)*, Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski, Celine Vens, and Sašo Džeroski (Eds.). Springer International Publishing, Cham, 339–355.

[42] Herbert Robbins and David Siegmund. 1971. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*. Elsevier, 233–257.

[43] Scott D Stoller, Ezio Bartocci, Justin Seyster, Radu Grosu, Klaus Havelund, Scott A Smolka, and Erez Zadok. 2011. Runtime verification with state estimation. In *International conference on runtime verification*. Springer, 193–207.

[44] Bing Sun, Jun Sun, Ting Dai, and Lijun Zhang. 2021. Probabilistic verification of neural networks against group fairness. In *International Symposium on Formal Methods*. Springer, 83–102.

[45] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 1144–1152.
[46] Wen-Chi Yang, Giuseppe Marra, Gavin Rens, and Luc De Raedt. 2023. Safe Reinforcement Learning via Probabilistic Logic Shields. In *IJCAI*. ijcai.org, 5739–5749.
[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

A Detailed Proofs

Claim 1 (Shielded decision process)**.** *The shielded decision process generated by the energy shield can be written as a sequence of Bernoulli random variables with evolving bias, $Z_t \sim \text{Bernoulli}(p_t)$. The biases are defined recursively as $p_1 = p$ and $p_{t+1} = f(\mu_t)$ for a given history $z_1, \dots, z_t$, where*

$$f(\mu) = \begin{cases} p + (1-p)\zeta(\mu) & \text{if } \mu \le \kappa, \\ p \cdot (1 - \zeta(\mu)) & \text{if } \mu > \kappa. \end{cases} \tag{14}$$

Moreover, the resulting shielded fairness process can be written as

$$M_t = M_{t-1} + \frac{1}{t}(Z_t - M_{t-1}) \quad \text{(with } M_0 = 0\text{)}. \tag{15}$$

Proof. Both equations are just simple computations.

Equation 14. Suppose $\mu_t \le \kappa$, i.e., the shield favors 1's. Then a 1 can be obtained either by $X_{t+1} = 1$ (which happens with probability p), or by flipping the decision $X_{t+1} = 0$ (which happens with probability $(1-p)\zeta(\mu_t)$). Therefore, when $\mu_t \le \kappa$, Z_{t+1} behaves like a Bernoulli of bias $p + (1-p)\zeta(\mu_t)$. Analogously, when $\mu_t > \kappa$, the shield favors 0's, so to obtain a 1 we need to toss $X_{t+1} = 1$ and for the shield to fail to flip the decision, which happens with probability $(1 - \zeta(\mu_t))$.

Equation 15.

$$M_t = \frac{\sum_{i=1}^{t} Z_i}{t} = \frac{Z_t}{t} + \frac{t-1}{t} \cdot \frac{\sum_{i=1}^{t-1} Z_i}{t-1} = \frac{Z_t}{t} + \frac{t-1}{t} M_{t-1} = M_{t-1} + \frac{1}{t}(Z_t - M_{t-1}).$$

□

A.1 Long term guarantees

Lemma 6.2. *The function $f\colon [0,1] \to [0,1]$ defined as in Eq. 2 is continuously differentiable, and has a unique point $\mu^* \in [0,1]$ such that $f(\mu^*) = \mu^*$. Furthermore, μ^* sits between p and κ.*

Proof. We need to prove smoothness, existence of the fixpoint, and that it sits between p and κ.

Smoothness. The function f inherits continuous differentiability from ζ clearly at all points except maybe $\mu = \kappa$. For $\mu = \kappa$, the assumption of the energy function being flat at the pivoting point κ guarantees continuous differentiability (Def. 4.1, item 3). We first write the expression for f':

$$f'(\mu) = \begin{cases} -p \cdot \zeta'(\mu) & \text{if } \mu > \kappa \\ (1-p)\zeta'(\mu) & \text{otherwise.} \end{cases} \tag{16}$$

We can check that f is continuous at $\mu = \kappa$, as

$$p \cdot (1 - \zeta(\kappa)) = p, \text{ and } \quad p + (1-p) \cdot \zeta(\kappa) = p.$$

Similarly, we can check that f' is continuous at $f = \kappa$, as

$$-p \cdot \zeta'(\kappa) = 0, \text{ and } (1-p) \cdot \zeta'(\kappa) = 0.$$

Note that both are necessary conditions. For f to be continuous at $\mu = \kappa$, we need

$$p(1-\zeta(\kappa)) = p + (1-p)\zeta(\kappa) \iff -p\zeta(\kappa) = \zeta(\kappa) - p\zeta(\kappa) \iff \zeta(\kappa) = 0.$$

Similarly, for f' to be continuous as $\mu = \kappa$, we need

$$-p\zeta'(\kappa) = (1-p)\zeta'(\kappa) \iff \zeta'(\kappa) = 0.$$

Existence and uniqueness of a fixpoint. Consider the function $g(\mu) = f(\mu) - \mu$. We have $g(0) = 0 + (1-p)\zeta(0) - 0 > 0$ and $g(1) = 1 \cdot (1-\zeta(1)) - 1 < 1 - 1 = 0$. Since g is continuous, there must be a point $\mu^* \in [0,1]$ such that $g(\mu^*) = 0$. Suppose this point is not unique, i.e., there exist two point $x < \mu$ such that $g(x) = g(\mu) = 0$. As we can see in Eq. 16, $f'(\mu) \leq 0$ for all $\mu \in [0,1]$, therefore g is non-increasing, so if $g(x) = g(\mu) = 0$, then $g(z) = 0$ for all $z \in [x, \mu]$. The interval $[x, \mu]$ contains at least two numbers that are either larger or lower than κ. Let $z_1, z_2 \in [x, \mu] \cap [0, \kappa)$ be such numbers, with $z_1 < z_2$. Then we have

$$p + (1-p)\zeta(z_1) - z_1 = p + (1-p)\zeta(z_2) - z_2,$$

which implies that

$$z_2 - z_1 = (1-p)(\zeta(z_2) - \zeta(z_1)).$$

On the left-hand side we have a positive number (because $z_1 < z_2$). On the right-hand side we have a non-positive number (because ζ is decreasing in the range $[0, \kappa)$, which is a contradiction. Since assuming $z_1, z_2 \in [x, \mu] \cap [0, \kappa)$ leads to a contradiction, it must be that $z_1, z_2 \in [x, \mu] \cap (\kappa, 1]$. However, with a similar argument this implies that

$$p(1-\zeta(z_1)) - z_1 = p(1-\zeta(z_2)) - z_2 \iff z_2 - z_1 = p(\zeta(z_1) - \zeta(\zeta_2)).$$

Since ζ is in the decreasing range, in the right-hand side we have again a non-positive term, which is a contradiction.

Placement of the fixpoint. We write the case where $p \leq \kappa$, the case where $p > \kappa$ is analogous. If $p \leq \kappa$, we have

$$g(p) = (1-p)\zeta(p) \geq 0, \quad \text{and} \quad g(\kappa) = p - \kappa \leq 0.$$

By continuity, there exists a point $\mu^* \in [p, \kappa]$ such that $g(\mu^*) = 0$. Note that this includes the special case where $p = \kappa$, in which the unique fixpoint is $\mu^* = \kappa = p$. □

Lemma 6.3. *The process described in Eq. 3 converges almost surely to the unique fixpoint μ^* of f, as defined in Eq. 2. Furthermore, the convergence rate of the error satisfies $(M_t - \mu^*)^2 = o(1/t^\lambda)$ for all $\lambda \in (0,1)$ almost surely.*

Proof. *Error construction.* For convenience, we will prove that the sequence converges to the unique root of $g(\mu) = f(\mu) - \mu$. Let $(V_t)_{t \in \mathbb{N}}$ be the sequence of squared distances to the target, i.e., $V_t = (\mu_t - \mu^*)^2$. We will prove that $V_t \to 0$ a.s. (almost surely).

First we need to find a recurrence formula for V_t. Recall from Eq. 7 the recurrence for μ_t is

$$\mu_t = \mu_{t-1} + \gamma_t(g(\mu_{t-1}) + \xi_t),$$

where $\gamma_t = 1/t$ and ξ_t has null expectation. For V_t we have

$$V_t = (\mu_t - \mu^*)^2 = (\mu_{t-1} - \mu^* + \gamma_t(g(\mu_{t-1}) + \xi_t))^2 \tag{17}$$

$$= (\mu_{t-1} - \mu^*)^2 + 2\gamma_t(\mu_{t-1} - \mu^*)(g(\mu_{t-1}) + \xi_t) + \gamma_t^2(g(\mu_{t-1}) + \xi_t)^2. \tag{18}$$

Taking conditional expectations we have

$$\mathbb{E}[V_t \mid \mu_{t-1}] = V_{t-1} + 2\gamma_t(\mu_{t-1} - \mu^*)g(\mu_{t-1}) + \gamma_t^2\mathbb{E}[(g(\mu_{t-1}) + \xi_t)^2]. \tag{19}$$

Since $g(\mu_{t-1})$ is deterministic given μ_{t-1} and $\mathbb{E}[\xi_t \mid \mu_{t-1}] = 0$, we have

$$\mathbb{E}[(g(\mu_{t-1}) + \xi_t)^2] = g(\mu_{t-1})^2 + \sigma_t^2,$$

where we define $\sigma_t^2 = \mathbb{E}[\xi_t^2 \mid \mu_{t-1}]$. Recall that $\xi_t = Z_t - f(\mu_{t-1})$, which is always in the interval $[-1, 1]$, therefore $\sigma_t^2 \leq 1$. Plugging everything into Eq. 18, we have

$$\mathbb{E}[V_t \mid \mu_{t-1}] \leq V_{t-1} + 2\gamma_t(\mu_{t-1} - \mu^*)g(\mu_{t-1}) + \gamma_t^2(g(\mu_{t-1})^2 + \sigma_t^2). \tag{20}$$

Let $\alpha_t = -2\gamma_t \frac{g(\mu_{t-1})}{\mu_{t-1}-\mu^*}$ and $\beta_t = \gamma_t^2(g(\mu_{t-1})^2 + \sigma_t^2)$. Note that α_t is well defined, even when $\mu_{t-1} = \mu^*$ as, using L'Hôpital's rule:

$$\lim_{\mu\to\mu^*} \frac{g(\mu)}{\mu - \mu^*} = \lim_{\mu\to\mu^*} \frac{f(\mu) - \mu}{\mu - \mu^*} = \lim_{\mu\to\mu^*} \frac{g'(\mu)}{1} = g'(\mu^*).$$

Also, g is decreasing around μ^*, so $g(\mu_{t-1})$ and $\mu_{t-1} - \mu^*$ have opposite signs. Therfore $\alpha_t \geq 0$. We can then rewrite Eq. 20 as

$$\mathbb{E}[V_t \mid \mu_{t-1}] \leq (1 - \alpha_t)V_{t-1} + \beta_t. \tag{21}$$

Proof of convergence. This is a standard form for the condition in the classical Robbins-Siegmund theorem [42], which guarantees convergence of V_t almost surely whenever the sequences (α_t) and (β_t) are non-negative and satisfy the limiting properties that $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \beta_t < \infty$ almost surely. We use the recent results in [32] to guarantee convergence. In particular, we use Thm. 5.1 to guarantee $V_t \to \infty$ almost surely, and Thm. 5.2 to guarantee rates of convergence. To apply both theorems, we need to guarantee that

(1) $\sum_{t=1}^{\infty} \alpha_t = \infty$ almost surely, and
(2) $\sum_{t=1}^{\infty} \beta_t < \infty$ almost surely.

Note that both results are true in the strict sense (not only "almost surely"). In the case of α_t, note that there exists $C > 0$ such that $g(\mu_t)/(\mu^* - \mu_t) \geq C$. Otherwise, it would imply that g has a zero other than μ^* or that $g'(\mu^*) = 0$, both which we know are not true. Therefore

$$\sum_{t=1}^{\infty} \alpha_t \geq 2C \sum_{t=1}^{\infty} \gamma_t = \infty.$$

Since $\gamma_t = 1/t$, the sum $\sum_t \gamma_t$ diverges. For the case of β_t, recall by definition that $g(\mu_t)^2 \leq 1$. We have also previously established that $\sigma_t^2 \leq 1$:

$$\sum_{t=1}^{\infty} \beta_t \leq 2 \sum_{t=1}^{\infty} \gamma_t^2 < \infty.$$

Theorem 5.1 in [32] is stated in terms of a sequence h_t. We can use $h_t = V_t$ (since the identity is what [32] defines as a class $\mathcal{B}$ function), to guarantee $V_t \to 0$ almost surely.

Convergence rates. From Theorem 5.2 in [32], we have that $V_t = o(1/t^\lambda)$ for any λ such that $\alpha_t \geq \lambda/t$ for sufficiently large t. In our case, we can take $\lambda = 2|g'(\mu^*)| - \epsilon$ for any $\epsilon > 0$. Therefore, $V_t = o(1/t^\lambda)$ for any $\lambda \in (0, 1) \cap (0, 2|g'(\mu^*)|) = (0, \min\{1, 2|g'(\mu^*)|\})$.

However, we can be more fine-grained and show that actually $\min\{1, 2|g(\mu^*)|\} = 1$. Recall from the definition of g that $g'(\mu) = f'(\mu) - 1$, and $f'(\mu) \leq 0$ for all $\mu \in [0, 1]$. Therefore $|g'(\mu)| = 1 + |f'(\mu)| \geq 1$. Altogether, we obtain the expected result of $V_t = o(1/t^\lambda)$ for all $\lambda \in (0, 1)$. □

Lemma 6.4. *For the process described in Eq. 3 , the corresponding sequence of average interference $(v_t)_{t\in\mathbb{N}}$ converges to $h(\mu^*)$, where μ^* is the fixpoint of f (Eq. 2) and h is as defined in Eq. 9.*

Proof. By definition $N_t = \frac{1}{t}\sum_{i=1}^{t} Y_i$.

Since $\mathbb{E}[Y_t \mid \mu_{t-1}] = h(\mu_{t-1})$, and $\mu_t \xrightarrow{t\to\infty} \mu^*$ almost surely. Since h is a continuous function,

$$\lim_{t\to\infty} N_t = h\left(\lim_{t\to\infty} M_t\right) = h(\mu^*).$$

□

Theorem 6.1. *Let $\mu^* \in [0,1]$. Given the shielded decision process as described in Eq. 3, with bias parameter p and energy function ζ, the shielded fairness process $(M_t)_{t\in\mathbb{N}}$ converges almost surely to μ^* if and only if*

$$\zeta(\mu^*) = \begin{cases} |\mu^* - p|/(1-p) & \text{if } p < \mu^*, \\ |\mu^* - p|/p & \text{otherwise.} \end{cases} \tag{22}$$

Furthermore, the expected interference cost $(\mathbb{E}[v_t])_{t\in\mathbb{N}}$ converges almost surely to $|\mu^* - p|$.

Proof. We know from Lemma 6.3 that the fairness outcome converges to a fixpoint that is always between p and κ. If we want the process to converge to a desired outcome μ^*, then we need to find κ and ζ such that $f(\mu^*) = \mu^*$.

The function $f(\mu)$ is defined by parts depending on whether μ is larger or smaller than κ. If $p \le \mu^*$, then we need $\kappa \ge \mu^*$, so we will be using the expression $f(\mu) = p + (1-p)\zeta(\mu)$ for $\mu = \mu^*$. Imposing μ^* to be a fixpoint, we have

$$\mu^* = p + (1-p)\zeta(\mu^*) \iff \zeta(\mu^*) = \frac{\mu^* - p}{1-p}.$$

Analogously, if $p \ge \kappa$, then we need $\kappa \le \mu^*$, so we will use the expression $f(\mu) = p(1-\zeta(\mu))$. If we set μ^* to be a fixpoint, we have

$$\mu^* = p(1-\zeta(\mu^*)) \iff \zeta(\mu^*) = 1 - \frac{\mu^*}{p}.$$

To prove the expectation of cost, just recall from Lemma 6.4 that the expected cost converges to $h(\mu^*)$. If $p \le \mu^*$, then $\kappa \ge \mu^*$, so we have

$$h(\mu^*) = (1-p)\zeta(\mu^*), \quad \text{and} \quad \zeta(\mu^*) = \frac{\mu^* - p}{1-p} \implies h(\mu^*) = \mu^* - p.$$

Analogously, if $p \ge \mu^*$, then $\kappa \le \mu^*$, so we have

$$h(\mu^*) = p\zeta(\mu^*), \quad \text{and} \quad \zeta(\mu^*) = 1 - \frac{\mu^*}{p} \implies h(\mu^*) = p - \mu^*.$$

This completes the proof, since in both cases $h(\mu^*) = |\mu^* - p|$. □

A.2 Short-term guarantees: Upper Bounds

Lemma A.1. *Let $\Delta_t = \mu_t - \mu^*$. Recall $\xi_t = Z_t - f(M_{t-1})$ from Equation 7 The following holds:*

$$\Delta_t = A_{1,t}\,\Delta_1 + \sum_{i=1}^{t-1} w_{i+1,t}\,\xi_{i+1}, \qquad \text{where} \tag{23}$$

$$A_{i,t} = \prod_{j=i}^{t-1}\left(1 - \frac{\alpha_j}{j+1}\right), \qquad w_{i,t} = \frac{A_{i,t}}{i}, \qquad \text{and} \qquad \alpha_t := 1 - \frac{f(\mu_t) - \mu^*}{\mu_t - \mu^*} \quad \text{with} \quad \alpha_t = 1, \ \text{if } \mu_t = \mu^*. \tag{24}$$

Furthermore, the following inequality holds:

$$\sum_{i=1}^{t-1} w_{i+1,t}^2 \le \frac{4^{1-\beta}}{t+1} \qquad \text{with} \qquad \beta = \sup_{r\in[0,1]} f'(r). \tag{25}$$

Proof. From the definition of Δ_t and the recursion for μ_t expressed in Equation 3, we have

$$\Delta_{t+1} \;=\; \Big(1-\frac{\alpha_t}{t+1}\Big)\Delta_t \;+\; \frac{\xi_{t+1}}{t+1}. \tag{26}$$

Equation 23 follows by induction on t from Eq. 26. The base case $t=1$ is trivial, since product defining $A_{i,t}$ is empty, and so is the sum of $w_{i+1,t}\xi_{i+1}$. For the induction step, can apply the induction hypothesis to Equation 26 to obtain

$$\begin{aligned}\Delta_{t+1} &= \Big(1-\frac{\alpha_t}{t+1}\Big)\Delta_t \;+\; \frac{\xi_{t+1}}{t+1}\\ &= \Big(1-\frac{\alpha_t}{t+1}\Big)\left(A_{1,t}\,\Delta_1 \;+\sum_{i=1}^{t-1} w_{i+1,t}\;\xi_{i+1}\right) + \frac{\xi_{t+1}}{t+1}.A_{1,t}\Big(1-\frac{\alpha_t}{t+1}\Big)\Delta_1 +\end{aligned}$$

Note that, from the definition of $A_{i,t}$, we have

$$\Big(1-\frac{\alpha_t}{t+1}\Big)A_{i,t} = A_{i,t+1}, \qquad \text{and} \qquad \Big(1-\frac{\alpha_t}{t+1}\Big)w_{i,t} = w_{i,t+1}.$$

Using the previous identities, plus the case that $w_{t+1,t+1} = \frac{1}{t+1}$ we get

$$\begin{aligned}\Delta_{t+1} &= A_{1,t+1}\Delta_1 + \sum_{i=1}^{t-1} w_{i+1,t+1}\xi_{i+1} + \frac{\xi_{t+1}}{t+1}\\ &= A_{1,t+1}\Delta_1 + \sum_{i=1}^{t} w_{i+1,t+1}\xi_{i+1},\end{aligned}$$

which finishes the induction step.

To prove the lower bound expressed in Equation 25, consider $\beta = \sup_{r\in[0,1]} f'(r)$. Since $f(\mu^*)=\mu^*$, and f is differentiable on the compact interval $[0,1]$, by the mean-value theorem

$$\frac{f(\mu)-\mu^*}{\mu-\mu^*} = \frac{f(\mu)-f(\mu^*)}{\mu-\mu^*} \in \{f'(z):\ z \text{ between } \mu \text{ and } \mu^*\} \subseteq \big(-\infty,\ \beta\big].$$

Therefore, we have the following lower bound for α_t that holds for every $t\in\mathbb{N}$:

$$\alpha_t = 1-\frac{f(\mu)-\mu^*}{\mu-\mu^*} \geq 1-\beta. \tag{27}$$

For $u\in(0,1)$, we can use the bound $\log(1-u)\leq -u$. We can apply this to the definition of $A_{i,t}$

$$\log A_{i,t} = \sum_{j=i}^{t-1}\log\left(1-\frac{\alpha_j}{j+1}\right) \leq -\sum_{j=i}^{t-1}\frac{\alpha_j}{j+1} \leq (\beta-1)\log\frac{t+1}{i+1} = \log\left(\frac{i+1}{t+1}\right)^{1-\beta}. \tag{28}$$

The last inequality stems from using twice that $\sum_{i=1}^{t} 1/t \leq \log(t+1)$. Using $(i+1)/i\leq 2$, we have

$$w_{i,t}^2 = \left(\frac{A_{i,t}}{i}\right)^2 \leq \frac{1}{i^2}\left(\frac{i+1}{t+1}\right)^{2-2\beta} \leq \left(\frac{2}{t+1}\right)^{2-2\beta} i^{-2\beta}.$$

Using this bound, we have

$$\sum_{i=1}^{t-1} w_{i+1,t}^2 \;\leq\; \frac{4^{1-\beta}}{(t+1)^{2-2\beta}}\sum_{i=1}^{t-1}(i+1)^{-2\beta} \leq_{(*)} \frac{4^{1-\beta}}{(t+1)^{2-2\beta}}(t-1)t^{-2\beta} \leq \frac{4^{1-\beta}}{t+1},$$

where in the $(*)$ inequality we are bounding each element of the sum by tha largest one, which is $t^{-2\beta}$, and that there are $t-1$ summands. This is valid because, given our definition of f as always non-increasing, $\beta\leq 0$. □

Lemma A.2. *Let $\delta > 0$. Let $\tau = \frac{2^{1+\frac{1}{1-\beta}}}{\delta^{1-\beta}} - 1$, and $K = (1/32) \cdot 4^{\beta}$. For all $t \geq \tau$, we have*

$$\mathbb{P}[\Delta_t > \delta] \leq \exp\left(-K\delta^2 t\right), \qquad \mathbb{P}[\Delta_t < -\delta] \leq \exp\left(-K\delta^2 t\right). \tag{29}$$

Proof. The idea is to use the expression in Equation 23 to bound Δ_t. Of the two summands, $A_{1,t}\Delta_1$ will be bounded by bounding $A_{1,t}$ for large enough t. For the second summand, we will show it is a martingale and use a concentration inequality to bound its value. We show the case for the first inequality. The second inequality follows the same argument, with a symmetric use of the Azuma-Hoeffding's inequality.

Fix $t \in \mathbb{N}$ and consider $E_{i,t} = \sum_{j=1}^{i} w_{j+1,t}\xi_{j+1}$, for $i < t$. When conditioning over the decisions up to time i, we have $\mathbb{E}[\xi_{i+1} \mid \mu_i] = 0$, and therefore

$$\mathbb{E}[E_{i,t} \mid \mu_i] = E_{i-1,t} + w_{i+1,t}\mathbb{E}[\xi_{i+1} \mid \mu_i] = E_{i-1,t}.$$

So $(E_{i,t})_{i=1}^{t-1}$ is a martingale. Its increments are bounded by $|E_{i,t} - E_{i-1,t}| = |w_{i+1,t}\xi_{i+1}| \leq |w_{i+1,t}|$. Applying Azuma-Hoeffding's inequality to $E_{t-1,t}$ we have, for any $\delta > 0$

$$\mathbb{P}\left[E_{t-1,t} > \delta\right] \leq \exp\left(-\frac{\delta^2}{2\sum_{i=1}^{t-1} w_{i+1,t}^2}\right) \leq \exp\left(-\frac{(t+1)\delta^2}{2 \cdot 4^{1-\beta}}\right) \tag{30}$$

From Equation 28, applied to $i = 1$, we have

$$A_{1,t} \leq \left(\frac{2}{t+1}\right)^{1-\beta}.$$

We can apply the triangle inequality to Equation 23 to get

$$|\Delta_t| \leq |A_{1,t}\Delta_1| + |E_{t-1,t}|. \tag{31}$$

So we can apply the bound in Equation 30 with $\delta/2$ to bound $\mathbb{P}[\Delta_t > \delta]$ as long as t is large enough to guarantee $|A_{1,t}\Delta_1| \leq \delta/2$. In such case

$$\mathbb{P}[\Delta_t > \delta] \leq \mathbb{P}[E_{t-1,t} > \delta/2] \leq \exp\left(-\frac{(t+1)(\delta/2)^2}{2 \cdot 4^{1-\beta}}\right) = \exp\left(-K(t+1)\delta^2\right), \tag{32}$$

with $K = (8 \cdot 4^{1-\beta})^{-1} = (1/32) \cdot 4^{\beta}$. For the previous bound to hold, we need t large enough so that $|A_{1,t}\Delta_1| \leq \delta/2$. Using $|\Delta_1| \leq 1$ and Equation 31, we have

$$|A_{1,t}\Delta_1| \leq \left(\frac{2}{t+1}\right)^{1-\beta}, \tag{33}$$

which is smaller than $\delta/2$ for $t \geq \frac{2^{1+\frac{1}{1-\beta}}}{\delta^{1-\beta}} - 1$. □

Theorem 5.2. *Let $I = [L, U]$ such that $\kappa, p \in I$. Let $\tau = 4/\min(|L - \mu^*|, |U - \mu^*|)$. Then for every $t \geq \tau$ we have*

$$\mathbb{P}(M_t \notin I) \leq \exp\left(-Kt|L - \mu^*|^2\right) + \exp\left(-Kt|U - \mu^*|^2\right), \tag{34}$$

where K is a positive constant defined as $K = (1/32) \cdot 4^{\beta}$ and $\beta = \sup_{r \in [0,1]} f'(r)$.

Proof. This is just a matter of unpacking the results from Lemma A.2 into the case of having a concrete interval. First note that

$$\mathbb{P}[M_t \notin I] = \mathbb{P}[M_t < L] + \mathbb{P}[M_t > U] = \mathbb{P}\Big[\Delta_t < -|L - \mu^*|\Big] + \mathbb{P}\Big[\Delta_t > |U - \mu^*|\Big].$$

We bound each of the summands using Lemma A.2, which is guaranteed to hold as long as $t \geq \frac{2^{1+\frac{1}{1-\beta}}}{\delta^{1-\beta}} - 1$. Since f is always a decreasing function with $f'(\kappa) = 0$, the value of β is always 0, so we have that

$$\frac{2^{1+\frac{1}{1-\beta}}}{\delta^{1-\beta}} - 1 \leq \frac{4}{\delta} - 1 \leq 4/\delta,$$

so if $t \geq 4/\delta$, Lemma A.2 holds. This is exactly the condition that $t \geq \tau$, for $\tau = 4/\min(|L - \mu^*|, |U - \mu^*|)$. □

Corollary 5.3. *Let $r_- = \exp(-K|L - \mu^*|^2)$, $r_+ = \exp(-K|U - \mu^*|^2)$. For every $T, T' \in \mathbb{N}$ such that $\tau \leq T < T'$ we have:*

$$\mathcal{E}_{\mathcal{S}}(M_{T:T'}) \leq \sum_{t=T}^{T'} (r_-^t + r_+^t) \qquad \text{and} \qquad \mathcal{E}_{\mathcal{S}}(M_{T:\infty}) \leq \frac{r_-^T}{1 - r_-} + \frac{r_+^T}{1 - r_+}.$$

Moreover, this provides the upper bound $\mathcal{P}_{\mathcal{S}}(M_{T:T'}) \leq \mathcal{E}_{\mathcal{S}}(M_{T:T'})$ for $T' \in \mathbb{N} \cup \{\infty\}$ s.t. $T' \geq T$.

Proof. This corollary is just an application of Boole's inequality (also known as union bound) on the results from Theorem 5.2. □

A.3 Short-term guarantees: Monotonicity

To prove Theorem 5.5, we will use an inductive argument and the tower property of the expectation.

Definition A.3 (Tower operator). Let $\mathcal{U} = \{u \colon [0,1] \to \mathbb{R}\}$ be the space of all measurable functions from $[0,1]$ to $\mathbb{R}$. Let $t \in \mathbb{N}$, and ζ be an energy function. The *tower operator* is the function $T_{\zeta,t} \colon \mathcal{U} \to \mathcal{U}$, that takes a function u and returns $T_{\zeta,t}u$ defined as

$$(T_{\zeta,t}u)(y) \;:=\; f_\zeta(y)\, u\big(y_t^+(y)\big) + \big(1 - f_\zeta(y)\big)\, u\big(y_t^-(y)\big),$$

where $y_t^+(y) := y + \frac{1-y}{t+1}$ and $y_t^-(y) := y - \frac{y}{t+1}$.

The *iterated tower operator* is $T_\zeta^{(t)} = T_{\zeta,1} \circ T_{\zeta,2} \circ \cdots \circ T_{\zeta,t}$

Lemma A.4. *Let $u, v \colon [0,1] \to \mathbb{R}$, $y \in [0,1]$, $t \in \mathbb{N}$, ζ an energy function and (y_k) be a stochastic process generated following Eq. 3. The following properties hold:*

- Expectation: $\mathbb{E}_\zeta[u(y_t)] = \mathbb{E}_\zeta[T_{\zeta,t-1}u(y_{t-1})]$
- Iterated expectation: $\mathbb{E}_\zeta[u(y_t)] = \mathbb{E}\left[T_\zeta^{(t-1)}u(y_1)\right]$.
- Monotonicity: *If $u(y) \leq v(y)$, then $(T_{\zeta,t}u)(y) \leq (T_{\zeta,t}v)(y)$.*

Proof. The first property follows from a simple computation of the expectation, using that $\mathbb{P}[y_{t+1} = y_t^+ \mid y_t] = f_\zeta(y_t)$. In fact, the tower operator is defined with the expectation property in mind. Applying the expectation property consecutively leads to the next property

$$\mathbb{E}_\zeta[u(y_t)] = \mathbb{E}_\zeta[T_\zeta^{(t-1)}u(y_1)].$$

Monotonicity follows directly from the definition, noting that both $f_\zeta(y^+)$ and $1 - f_\zeta(y^-)$ are always non-negative. □

The iterated expectation property is useful because it lets us write the expectation of $u(y_m)$, which depends on the distribution of y_m after m steps of the process, in terms of the expectation of a function depending only on the much simpler distribution of y_1. In particular, the distribution over y_1 does not depend on ζ. If we

want to prove a result of the form $\mathbb{E}_{\zeta_1}[u(y_t)] \le \mathbb{E}_{\zeta_2}[u(y_t)]$ for some u, ζ_1, and ζ_2, it is equivalent to prove that $\mathbb{E}[T_{\zeta_1}^{(t-1)} u(y_1)] \le \mathbb{E}[T_{\zeta_2}^{(t-1)} u(y_1)]$, so it suffices to prove that

$$T_{\zeta_1}^{(t-1)} u(y) \le T_{\zeta_2}^{(t-1)} u(y), \qquad \text{for} \quad y \in \{0,1\}. \tag{35}$$

This observation lets us go from a local comparison to a global result. We study two properties on the short term: the probability of violating point fairness, and the expected number of point fairness violations. For the first one, we can take $u(y) = \mathbf{1}\{y \notin (L,U\}$. For the second one, we can take $u(y) = (L - y_k)_+ + (y_k - U)_+$.

Lemma A.5. *Consider the functions y_t^+, y_t^- as defined in Def. A.3. Then*

(1) *If $y \le \kappa$ (start below):*
   (a) *If $y_t^+(y) \le \kappa$ (no overshoot), then $|y_t^+(y) - \kappa| < |y_t^-(y) - \kappa|$.*
   (b) *If $y_t^+(y) > \kappa$ (overshoot), then*

$$|y_t^+(y) - \kappa| \le \frac{1-\kappa}{t+1}, \qquad |y_t^-(y) - \kappa| \le \frac{1}{t+1}. \tag{36}$$

(2) *If $y > \kappa$ (start above):*
   (a) *If $y_t^-(y) \ge \kappa$ (no overshoot), then $|y_t^+(y) - \kappa| > |y_t^-(y) - \kappa|$,*
   (b) *If $y_t^-(y) < \kappa$ (overshoot), then*

$$|y_t^-(y) - \kappa| \le \frac{\kappa}{t+1}, \qquad |y_t^+(y) - \kappa| \le \frac{1}{t+1}. \tag{37}$$

Proof. We follow the proof case by case.

- *Case 1(a).* $|y_t^-(y) - \kappa| - |y_t^+(y) - \kappa| = y_t^+(y) - y_t^-(y) = 1/(t+1) > 0$.
- *Case 1(b)(left).*
$$|y_t^+(y) - \kappa| = \frac{1-y}{t+1} - (\kappa - y) = \frac{1 - \kappa t - \kappa + yt}{t+1}.$$
  Using $y \le \kappa$ in the previous equation, we can cancel $-\kappa t$ with κt, and get the result.
- *Case 1(b)(right).* Using $y_t^-(y) \ge 0$, we have $|y_t^-(y) - \kappa| = \kappa - y_t^-(y) < \kappa$.
- *Case 2(a).* $|y_t^+(y) - \kappa| - |y_t^-(y) - \kappa| = y_t^+(y) - y_t^-(y) = 1/(t+1) > 0$.
- *Case 2(b)(left).*
$$|y_t^-(y) - \kappa| = \frac{y}{t+1} - (y - \kappa) = \frac{y - yt - y + \kappa t + \kappa}{t+1}.$$
  We can cancel y and $-y$, and using $\kappa \le y$, we can cancel yt with $-yt$, and get the result.
- *Case 2(b)(right).* Simply note that $|y_t^+(y) - \kappa| \le |y_t^+(y) - y_t^-(y)| = 1/(t+1)$.

□

Definition A.6 (Unimodal function). Let $D \subseteq \mathbb{R}$ be some domain. We say that a function $u\colon D \to \mathbb{R}$ is κ-unimodal if $\kappa \in D$, it is non-increasing on $\{x : x \le \kappa\}$, and non-decreasing on $\{x : \kappa \le x\}$.

Note that if a function u is κ-unimodal, then $u(\kappa)$ is its absolute minimum.

Lemma A.7. *Let $\zeta_1 \succeq \zeta_2$. Let $T \in \mathbb{N}$. Let u be a κ-unimodal function, and with $u(y) = 0$ for $y \in [a,b]$, where*

$$a \le \kappa - \frac{1}{T+1}, \qquad b \ge \kappa + \frac{1}{T+1}.$$

Then for all $t \ge T$, we have for $y \in \{0,1\}$ that

$$T_{\zeta_1}^{(t)} u(y) \le T_{\zeta_2}^{(t)} u(y).$$

Proof. *Step 1: a pointwise one-step inequality.* We first prove that for any $t \in \mathbb{N}$, $y \in [0,1]$, and any κ-unimodal function v that is null on $[a,b]$, we have

$$(T_{\zeta_1,t}v)(y) \;\le\; (T_{\zeta_2,t}v)(y). \tag{38}$$

Using the definition of the tower operator, we have

$$\begin{aligned}(T_{\zeta_1,t}v)(y) - (T_{\zeta_2,t}v)(y) &= \Big(f_{\zeta_1}(y)v(y_t^+) + (1-f_{\zeta_1}(y))v(y_t^-)\Big) - \Big(f_{\zeta_2}(y)v(y_t^+) + (1-f_{\zeta_2}(y))v(y_t^-)\Big)\\ &= f_{\zeta_1}(y)\Big(v(y_t^+) - v(y_t^-)\Big) - f_{\zeta_2}(y)\Big(v(y_t^+) - v(y_t^-)\Big)\\ &= \Big(f_{\zeta_1}(y) - f_{\zeta_2}(y)\Big)\cdot\Big(v(y_t^+) - v(y_t^-)\Big).\end{aligned} \tag{39}$$

Therefore, to check $(T_{\zeta_1,t}v)(y) - (T_{\zeta_2,t}v)(y) \le 0$ we just need to prove that both factors of the previous equation are of different sign, or one of them is zero.

We check both sides of κ separatedly.

Case $y \le \kappa$. Since $\zeta_1 \succeq \zeta_2$, we have $f_{\zeta_1}(y) \ge f_{\zeta_2}(y)$. If we are in the no-overshoot regime ($y_t^+(y) \le \kappa$), we have v in its descending mode in the whole interval $[y_t^-(y), y_t^+(y)]$, so $v(y_t^+(y)) \le v(y_t^-(y))$. Therefore $f_{\zeta_1}(y) - f_{\zeta_2}(y)$ and $v(y_t^+) - v(y_t^-)$ have opposed signs.

If we are in the overshoot regime ($y_t^+(y) \le \kappa$), we want to make sure that $v(y_t^+) = v(y_t^-) = 0$. Using Eq. 36, we can guarantee it with

$$a \le \kappa - \frac{1}{t+1}, \qquad \text{and} \qquad b \ge \kappa + \frac{1-\kappa}{t+1},$$

which is guaranteed for $t \ge T$ in the hypothesis of the lemma.

Case $y > \kappa$. This case is analogous, following the same argument as in $y \le \kappa$, taking into account the switch in signs.

Step 2: iterate the one-step inequality. For $r = 0, 1, \dots, t$ define

$$\Phi_r := B_1 B_2 \cdots B_r \, A_{r+1} A_{r+2} \cdots A_t \, u,$$

with the conventions

$$A_s := T_{\zeta_1,s}, \qquad B_s := T_{\zeta_2,s}, \qquad \Phi_0 = A_1 A_2 \cdots A_t u = T_{\zeta_1}^{(t)} u, \qquad \Phi_t = B_1 B_2 \cdots B_t u = T_{\zeta_2}^{(t)} u.$$

Our goal is to show the chain of inequalities

$$\Phi_0(y) \;\le\; \Phi_1(y) \;\le\; \cdots \;\le\; \Phi_t(y), \qquad \text{for } y \in \{0,1\}, \tag{40}$$

which implies $T_{\zeta_1}^{(t)} u(y) \le T_{\zeta_2}^{(t)} u(y)$ at the endpoints.

Telescoping principle. For each $r = 1, \dots, t$, set

$$v^{(r)} := A_{r+1} A_{r+2} \cdots A_t u.$$

Then

$$\Phi_{r-1} = B_1 \cdots B_{r-1}\,(A_r v^{(r)}), \qquad \Phi_r = B_1 \cdots B_{r-1}\,(B_r v^{(r)}).$$

Thus, if we can show

$$A_r v^{(r)}(y) \;\le\; B_r v^{(r)}(y), \qquad y \in \{0,1\}, \tag{41}$$

then applying the prefix operator $B_1 \cdots B_{r-1}$ (which is order-preserving because of the monotonicity property in Lemmas A.4) yields

$$\Phi_{r-1}(y) \;\le\; \Phi_r(y) \qquad \text{for } y \in \{0,1\}.$$

Chaining over $r = 1, \dots, t$ establishes the desired inequality (Eq.40).

Verification of 41. Fix r and write $y^+ := y_r^+(y)$, $y^- := y_r^-(y)$ for brevity. By the algebraic identity from Step 1 (Eq.39),

$$(A_r v^{(r)} - B_r v^{(r)})(y) = (f_{\zeta_1}(y) - f_{\zeta_2}(y))\,(v^{(r)}(y^+) - v^{(r)}(y^-)).$$

At the endpoints $y \in \{0,1\}$, the sign of $f_{\zeta_1} - f_{\zeta_2}$ is fixed by the assumption $\zeta_1 \succeq \zeta_2$, while the sign of $v^{(r)}(y^+) - v^{(r)}(y^-)$ is determined by the geometry of the successors and the plateau assumption:

- For $y = 0$: we have $f_{\zeta_1}(0) \geq f_{\zeta_2}(0)$. If $y^+ \leq \kappa$ (no overshoot), then $|y^+ - \kappa| < |y^- - \kappa|$, hence $v^{(r)}(y^+) \leq v^{(r)}(y^-)$. If $y^+ > \kappa$ (overshoot), then by the plateau condition both $y^\pm \in [a,b]$, so $v^{(r)}(y^\pm) = 0$. In both cases, $v^{(r)}(y^+) - v^{(r)}(y^-) \leq 0$, so the product is ≤ 0.

- For $y = 1$: we have $f_{\zeta_1}(1) \leq f_{\zeta_2}(1)$. If $y^- \geq \kappa$ (no overshoot), then $|y^- - \kappa| < |y^+ - \kappa|$, hence $v^{(r)}(y^-) \leq v^{(r)}(y^+)$, i.e. $v^{(r)}(y^+) - v^{(r)}(y^-) \geq 0$. If $y^- < \kappa$ (overshoot), then $y^\pm \in [a,b]$, so $v^{(r)}(y^\pm) = 0$. In both cases, $v^{(r)}(y^+) - v^{(r)}(y^-) \geq 0$, so the product is ≤ 0.

Thus $(A_r v^{(r)} - B_r v^{(r)})(y) \leq 0$ for $y \in \{0,1\}$, proving 41, which in turn proves Eq. 40, as we wanted. □

Theorem 5.5. *Let $\zeta_1 \succeq \zeta_2$ be two energy functions, with a common minimum at κ. Let $\mathcal{S} = [L,U]$. Let M^{ζ_1} and M^{ζ_2} be the shielded fairness process generated by enforcing the decision process of $p \in [0,1]$ with ζ_1 and ζ_2, respectively. Let $\tau = \lceil \max\{1/|\kappa - L|, 1/|\kappa - U|\} \rceil$, for all $T \in \mathbb{N}$ and $T' \in \mathbb{N} \cup \{\infty\}$ such that $T < T'$ we have*

$$\mathcal{E}_{\mathcal{S}}(M^{\zeta_1}_{T:T'}) \leq \mathcal{E}_{\mathcal{S}}(M^{\zeta_2}_{T:T'}), \qquad \text{and} \qquad \mathcal{P}_{\mathcal{S}}(M^{\zeta_1}_{T:T'}) \leq \mathcal{P}_{\mathcal{S}}(M^{\zeta_2}_{T:T'}).$$

Proof. This is a direct consequence of Lemma A.7. The condition on τ follows from it, and the condition on the values of either probability or expectation of point fairness violation come from the characterization of both of the safety measures as the expectation with certain functions u. In particular, for the first one, we can take $u(y) = \mathbf{1}\{y \notin (L,U)\}$, and for the second one, we can take $u(y) = (L-y)_+ + (y-U)_+$. The expectation of the outcome at a certain timestep corresponds to the expectation of the iterated tower operator at the first timestep because of Lemma A.4. □

A.4 Extension to Group Fairness

For $g \in \{A,B\}$ define the group counts and shielded acceptance sums

$$N_{g,t} := \sum_{i=1}^{t} \mathbf{1}[G_i = g], \qquad S_{g,t} := \sum_{i=1}^{t} Z_i\,\mathbf{1}[G_i = g],$$

and the empirical group acceptance rates (when $N_{g,t} \geq 1$)

$$M_{g,t} := \frac{S_{g,t}}{N_{g,t}}, \qquad M_t := M_{A,t} - M_{B,t} \in [-1,1].$$

Write $r_A = \mathbb{P}[G_t = A]$, $r_B = 1 - r_A$, and $r_{\min} = \min\{r_A, r_B\}$.

Energy-shielded decision rule in the 2-group setting. Fix pivot $\kappa \in [-1,1]$ and energy function $\zeta : [-1,1] \to [0,1]$ with minimum at κ (same conditions as Def. 4.1, with domain $[-1,1]$). At time $t+1$ the shield observes (G_{t+1}, X_{t+1}) and the current fairness M_t.

- If $M_t \leq \kappa$, the shield attempts to *increase* M_t: it favors $Z_{t+1} = 1$ when $G_{t+1} = A$, and favors $Z_{t+1} = 0$ when $G_{t+1} = B$, flipping the unfavorable outcome with probability $\zeta(M_t)$.

- If $M_t > \kappa$, the shield attempts to *decrease* M_t: it favors $Z_{t+1} = 0$ when $G_{t+1} = A$, and favors $Z_{t+1} = 1$ when $G_{t+1} = B$, flipping the unfavorable outcome with probability $\zeta(M_t)$.

Induced conditional biases. Define the baseline disparity $d := p_A - p_B \in (-1, 1)$ and the two coefficients

$$c_- := (1 - p_A) + p_B = 1 - d, \qquad c_+ := p_A + (1 - p_B) = 1 + d.$$

Conditioned on $M_t = \mu$, the shielded decision has the following conditional acceptance probabilities:

$$\mathbb{P}[Z_{t+1} = 1 \mid M_t = \mu, G_{t+1} = A] = \begin{cases} p_A + (1 - p_A)\zeta(\mu) & \mu \le \kappa, \\ p_A(1 - \zeta(\mu)) & \mu > \kappa, \end{cases}$$

$$\mathbb{P}[Z_{t+1} = 1 \mid M_t = \mu, G_{t+1} = B] = \begin{cases} p_B(1 - \zeta(\mu)) & \mu \le \kappa, \\ p_B + (1 - p_B)\zeta(\mu) & \mu > \kappa. \end{cases}$$

Consequently, the *one-dimensional drift map* for the fairness difference is

$$f(\mu) := \begin{cases} d + c_- \, \zeta(\mu) & \mu \le \kappa, \\ d - c_+ \, \zeta(\mu) & \mu > \kappa. \end{cases} \tag{42}$$

Fairness recursion (random step sizes). If $G_{t+1} = A$ then $M_{B,t+1} = M_{B,t}$ and

$$M_{A,t+1} = M_{A,t} + \frac{1}{N_{A,t+1}}\big(Z_{t+1} - M_{A,t}\big),$$

hence

$$M_{t+1} = M_t + \frac{1}{N_{A,t+1}}\big(Z_{t+1} - M_{A,t}\big) \qquad \text{if } G_{t+1} = A. \tag{43}$$

If $G_{t+1} = B$ then $M_{A,t+1} = M_{A,t}$ and

$$M_{B,t+1} = M_{B,t} + \frac{1}{N_{B,t+1}}\big(Z_{t+1} - M_{B,t}\big),$$

hence

$$M_{t+1} = M_t - \frac{1}{N_{B,t+1}}\big(Z_{t+1} - M_{B,t}\big) \qquad \text{if } G_{t+1} = B. \tag{44}$$

A.4.1 Long-term guarantees in the 2-group setting.

Lemma A.8 (Unique fixpoint of the 2-group drift map). *Let $f : [-1, 1] \to [-1, 1]$ be as in Eq. 42. Then f is continuously differentiable and has a unique fixpoint $\mu^* \in [-1, 1]$ satisfying $f(\mu^*) = \mu^*$. Moreover, μ^* lies between $d = p_A - p_B$ and the pivot κ.*

Proof. *Smoothness.* Away from κ, f inherits differentiability from ζ. At κ, the conditions $\zeta(\kappa) = 0$ and $\zeta'(\kappa) = 0$ imply

$$\lim_{\mu \uparrow \kappa} f(\mu) = d + c_- \, \zeta(\kappa) = d = d - c_+ \, \zeta(\kappa) = \lim_{\mu \downarrow \kappa} f(\mu),$$

and similarly for the derivative since both side-derivatives equal $c_- \, \zeta'(\kappa) = -c_+ \, \zeta'(\kappa) = 0$. Hence f is continuously differentiable.

Existence. Let $g(\mu) := f(\mu) - \mu$. We have $g(-1) = f(-1) + 1 \ge -1 + 1 = 0$ and $g(1) = f(1) - 1 \le 1 - 1 = 0$, so by continuity there exists $\mu^* \in [-1, 1]$ with $g(\mu^*) = 0$.

Uniqueness. By the energy-function shape, ζ is non-increasing on $(-1, \kappa]$ and non-decreasing on $[\kappa, 1)$. Therefore f is non-increasing on all of $[-1, 1]$: on $\mu \le \kappa$ we have $f'(\mu) = c_- \, \zeta'(\mu) \le 0$, and on $\mu > \kappa$ we have $f'(\mu) = -c_+ \, \zeta'(\mu) \le 0$. Hence $g(\mu) = f(\mu) - \mu$ is strictly decreasing wherever $f'(\mu) > -1$, and in any case is decreasing overall. A decreasing continuous function can have at most one zero, so the fixpoint is unique.

Placement between d and κ. Assume $d \le \kappa$ (the other case is symmetric). Then $g(d) = f(d) - d = c_- \, \zeta(d) \ge 0$ and $g(\kappa) = f(\kappa) - \kappa = d - \kappa \le 0$. By continuity, the unique root μ^* lies in $[d, \kappa]$. □

Lemma A.9 (Almost sure convergence of M_t to the fixpoint). *Let (M_t) be generated by the 2-group energy shield and assume $r_A, r_B \in (0,1)$. Then $M_t \to \mu^*$ almost surely, where μ^* is the unique fixpoint from Lemma A.8. Moreover, $(M_t - \mu^*)^2 = o(1/t^\lambda)$ for all $\lambda \in (0,1)$ almost surely.*

Proof. We follow the same Robbins–Monro / Robbins–Siegmund template as in Lemma 6.3, noting that the only difference is that the step size is random but of order $1/t$.

Define $g(\mu) := f(\mu) - \mu$ and $\Delta_t := M_t - \mu^*$. Using the update rules 43–44 we can rewrite

$$M_{t+1} = M_t + \gamma_{t+1}\big(g(M_t) + \xi_{t+1}\big),$$

where

$$\gamma_{t+1} := \begin{cases} 1/N_{A,t+1} & \text{if } G_{t+1} = A, \\ 1/N_{B,t+1} & \text{if } G_{t+1} = B, \end{cases} \qquad \xi_{t+1} := \big(M_{t+1} - M_t\big)/\gamma_{t+1} - g(M_t).$$

By construction, $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0$ and $|\xi_{t+1}| \le 2$ (since all involved quantities lie in $[-1,1]$). Moreover, standard law of large numbers gives $N_{g,t}/t \to r_g$ a.s., hence $\gamma_t \asymp 1/t$ a.s. and

$$\sum_{t\ge 1} \gamma_t = \infty, \qquad \sum_{t\ge 1} \gamma_t^2 < \infty \qquad \text{a.s.}$$

Since f is decreasing and has a unique fixpoint, $g(\mu)$ has opposite sign to $(\mu - \mu^*)$, implying the same "negative drift in squared error" inequality as in Eq. 20 (with γ_t instead of $1/t$ and bounded noise). Thus Robbins–Siegmund applies and yields $\Delta_t \to 0$ almost surely. Rates follow by the same argument as in Lemma 6.3, using that $\gamma_t \asymp 1/t$ a.s. □

Convergence to a desired target. Exactly as in the 1D case, the process converges to the (unique) fixpoint of f. Therefore, to enforce a desired long-term target $\mu^* \in [-1,1]$, it is necessary and sufficient to choose ζ so that $f(\mu^*) = \mu^*$.

Theorem 7.1 *Let $\mu^* \in [-1,1]$ and $d = p_A - p_B$. The 2-group shielded fairness process (M_t) converges almost surely to μ^* if and only if*

$$\zeta(\mu^*) = \begin{cases} (\mu^* - d)/(1-d) & \text{if } d \le \mu^*, \\ (d - \mu^*)/(1+d) & \text{if } d > \mu^*. \end{cases} \tag{45}$$

Proof. By Lemma A.9, M_t converges to the unique fixpoint of f. Hence $M_t \to \mu^*$ a.s. iff $f(\mu^*) = \mu^*$.

If $\mu^* \le \kappa$ then $f(\mu^*) = d + (1-d)\zeta(\mu^*)$ (since $c_- = 1 - d$), giving

$$\mu^* = d + (1-d)\zeta(\mu^*) \iff \zeta(\mu^*) = \frac{\mu^* - d}{1-d}.$$

This corresponds exactly to the case $d < \mu^*$ (because the fixpoint lies between d and κ).

If $\mu^* > \kappa$ then $f(\mu^*) = d - (1+d)\zeta(\mu^*)$ (since $c_+ = 1 + d$), giving

$$\mu^* = d - (1+d)\zeta(\mu^*) \iff \zeta(\mu^*) = \frac{d - \mu^*}{1+d},$$

corresponding to $d > \mu^*$. The case $d = \mu^*$ forces $\zeta(\mu^*) = 0$. □

Interference cost. Let $v_t = \frac{1}{t}\sum_{i=1}^t Y_i$ be the average intervention rate. Conditioned on $M_t = \mu$, an intervention happens iff the raw decision is unfavorable and the shield flips it with probability $\zeta(\mu)$. Thus define

$$h(\mu) := \begin{cases} r_A(1-p_A)\zeta(\mu) + r_B\, p_B\, \zeta(\mu) & \mu \le \kappa, \\ r_A\, p_A\, \zeta(\mu) + r_B(1-p_B)\zeta(\mu) & \mu > \kappa. \end{cases} \tag{46}$$

Lemma A.10 (2-group analogue of Lemma 6.4). *We have $\nu_t \to h(\mu^*)$ almost surely, where μ^* is the fixpoint of f. In particular, if the shield is tuned so that $\mu^* = \mu^*$, then $\nu_t \to h(\mu^*)$ almost surely.*

Proof. By definition, $\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] = h(M_t)$ and h is continuous. Since $M_t \to \mu^*$ a.s. (Lemma A.9), we obtain $h(M_t) \to h(\mu^*)$ a.s. A standard martingale SLLN argument for bounded increments yields $\nu_t - \frac{1}{t}\sum_{i=1}^{t} h(M_{i-1}) \to 0$ a.s., hence $\nu_t \to h(\mu^*)$ a.s. □

Lemma 7.2 *For every $\eta \in (0,1)$, let $\tau_G(\eta) := \frac{8}{r_{\min}} \log \frac{4}{\eta}$. Then for all $T \geq \tau_G(\eta)$, it holds that $\mathbb{P}(E_T) \geq 1 - \eta$.*

Proof. In particular, we prove that

$$\tau_G(\eta) := \frac{8}{r_{\min}} \log \frac{4}{\eta}.$$

For each fixed t, Chernoff bounds for binomials give

$$\mathbb{P}\big[N_{A,t} < \tfrac{r_A}{2} t\big] \leq \exp\big(-\tfrac{r_A}{8} t\big), \qquad \mathbb{P}\big[N_{B,t} < \tfrac{r_B}{2} t\big] \leq \exp\big(-\tfrac{r_B}{8} t\big).$$

Union bound over $t \geq T$ and over both groups yields

$$\mathbb{P}(E_T^c) \leq 2 \sum_{t \geq T} \exp\big(-\tfrac{r_{\min}}{8} t\big) \leq 2 \cdot \frac{\exp(-\frac{r_{\min}}{8} T)}{1 - \exp(-\frac{r_{\min}}{8})}.$$

Choosing $T = \tau_G(\eta)$ makes the RHS at most η (up to constant slack absorbed by the chosen $8 \log(4/\eta)$ form). □

Theorem 7.3. *Fix an interval $I = [L, U] \subseteq [-1, 1]$ such that $\kappa, d \in I$ and let $\eta \in (0,1)$. Let $K = \min\{r_A, r_B\}/32$, then for all $t \geq \tau_G(\eta)$*

$$\mathbb{P}(M_t \notin I) \;\leq\; \eta \;+\; \exp\big(-K\, t\, |L - \mu^*|^2\big) \;+\; \exp\big(-K\, t\, |U - \mu^*|^2\big). \tag{47}$$

Proof. Define

$$\varepsilon_L := |\mu^* - L|, \qquad \varepsilon_U := |U - \mu^*|,$$

and

$$K_2 := \frac{r_{\min}}{32} \cdot 4^{\beta_2} \qquad \text{with} \qquad \beta_2 := \sup_{\mu \in [-1,1]} f'(\mu) \leq 0.$$

On the event E_τ, for all $t \geq \tau$ we have

$$\gamma_{t+1} \leq \max\Big\{\frac{2}{r_A(t+1)}, \frac{2}{r_B(t+1)}\Big\} \leq \frac{2}{r_{\min}(t+1)}.$$

Thus, conditioning on E_τ, the recursion for Δ_t becomes a Robbins–Monro recursion with deterministic upper bounds on step sizes, and the same "unrolled martingale + Azuma" argument as in Lemmas A.2–5.2 applies verbatim with the substitution

$$\frac{1}{t+1} \;\rightsquigarrow\; \frac{2}{r_{\min}(t+1)}.$$

This scales the variance proxy by a factor $4/r_{\min}^2$, yielding an exponential tail with constant $K_2 = \Theta(r_{\min})$ (as stated). Finally, decondition:

$$\mathbb{P}(M_t \notin I) \leq \mathbb{P}(E_\tau^c) + \mathbb{P}(M_t \notin I \mid E_\tau) \leq \eta + (\text{Azuma tails}).$$

□

Theorem 7.4. *Let $\zeta_1 \succeq \zeta_2$ be two energy functions on $[-1,1]$ with common minimum at κ. Fix $I = [L, U] \subseteq [-1, 1]$ and let M^{ζ_1}, M^{ζ_2} be the corresponding 2-group shielded fairness processes. Then for every $\eta \in (0,1)$, for all $T < T'$ with $T \geq \tau_G(\eta)$,*

$$\mathcal{E}_I\Big(M^{\zeta_1}_{T:T'} \mid E_T\Big) \;\leq\; \mathcal{E}_I\Big(M^{\zeta_2}_{T:T'} \mid E_T\Big), \quad \textit{and} \quad \mathcal{P}_I\Big(M^{\zeta_1}_{T:T'} \mid E_T\Big) \;\leq\; \mathcal{P}_I\Big(M^{\zeta_2}_{T:T'} \mid E_T\Big).$$

Proof. Condition on E_T so that for all $t \geq T$, both denominators satisfy $N_{g,t} \geq (r_g/2)t$, hence the one-step updates 43–44 move the state by at most $O(1/t)$. On each side of κ, the drift map f from Eq. 42 is affine in ζ with a *nonnegative* coefficient: for $\mu \leq \kappa$, $f(\mu) = d+(1-d)\zeta(\mu)$ is increasing in $\zeta(\mu)$, and for $\mu > \kappa$, $f(\mu) = d-(1+d)\zeta(\mu)$ is decreasing in $\zeta(\mu)$. Thus, exactly as in Eq. 39, the one-step tower difference factors as

$$(T_{\zeta_1,t}v - T_{\zeta_2,t}v)(\mu) = \big(f_{\zeta_1}(\mu) - f_{\zeta_2}(\mu)\big) \cdot \big(v(\mu^+) - v(\mu^-)\big),$$

where $\mu^\pm$ denote the two possible next fairness values (accept/reject), and where the sign of $f_{\zeta_1} - f_{\zeta_2}$ is fixed by $\zeta_1 \succeq \zeta_2$ and the side of κ. For any κ-unimodal test function v that is 0 on an interval containing $\kappa \pm O(1/t)$ (the same plateau condition as in Lemma A.7), the term $v(\mu^+)-v(\mu^-)$ has the opposite sign, implying $(T_{\zeta_1,t}v)(\mu) \leq (T_{\zeta_2,t}v)(\mu)$. Iterating this inequality exactly as in Lemma A.7 yields the claimed ordering of expectations for the indicator test function defining $\mathcal{P}$ and for the additive test function defining $\mathcal{E}$. □

A.5 Shield Synthesis

Markov chain construction. We construct the acyclic Markov chain $\mathcal{M} = (S, P, s_0)$ consisting of a set of states S, a Markov transition kernel $P\colon S \times S \to [0,1]$, and an initial state $s_0 = (1,0)$. Intuitively, the set of states is stratified into time steps $S = \bigcup_{t\in[0;T]} S_t$ where $S_t = \{(t,m) \mid m \in [t]\}$ for all $t \in [1;T]$. Intuitively, a state $s = (t,c) \in S$ is labeled with a point in time t, the positive decision counter c clearly upper bound by the point in time t. The Markov transition kernel P is defined as follows: for every state $s = (t,c)$ with $t \in [T-1]$ we transition to state $s = (t+1, c+1)$ with probability $f(c/t)$, and to state $s' = (t+1, c)$ with probability $1 - f(c/t)$.

Dynamic programming. We compute the value of both the probabilistic and expected violation measure for the interval $[1;T]$ using dynamic programming on the Markov chain (S, P, s_0). Let $\gamma(s) = \mathbf{1}[c/t \notin \mathcal{S}]$, then for every $s = (t,c) \in S$ we define its value $V_{\mathcal{E}}$ for the expected violation measure as

$$\begin{aligned} V_{\mathcal{E}}(s) &= \gamma(s) + \big(f(c/t)V_{\mathcal{E}}(t+1,c+1) + (1-f(c/t))V_{\mathcal{E}}(t+1,c)\big) && \text{if } t \in [T-1] \\ V_{\mathcal{E}}(s) &= \gamma(s) && \text{if } t = T \end{aligned}$$

Moreover, we define its value $V_{\mathcal{P}}$ for the expected violation measure as

$$\begin{aligned} V_{\mathcal{P}}(s) &= \max\Big(\gamma(s), f(c/t)V_{\mathcal{P}}(t+1,c+1) + (1-f(c/t))V_{\mathcal{P}}(t+1,c)\Big) && \text{if } t \in [T-1] \\ V_{\mathcal{P}}(s) &= \gamma(s) && \text{if } t = T \end{aligned}$$

Notice that if $\gamma(c,t) = 1$ no further computation recursion is required.

Intuition. The value function $V_{\mathcal{E}}(s)$ represents the expected number of violations incurred from state $s = (t,c)$ onward, including a possible violation at time t. Formally, our process M takes on the value $M_t = c/t$ at time t, then

$$V_{\mathcal{E}}(t,c) \;=\; \mathbf{1}[M_t \notin \mathcal{S}] + \mathbb{E}\left[\sum_{i=t+1}^{T} \mathbf{1}[M_i \notin \mathcal{S}] \;\middle|\; M_t\right].$$

Hence, the value at the initial state $s_0 = (0,0)$ equals the total expected number of violations up to time T:

$$V_{\mathcal{E}}(s_0) \;=\; \mathbb{E}\left[\sum_{i=1}^{T} \mathbf{1}[M_i \notin \rho(i)]\right].$$

The value function $V_{\mathcal{P}}(s)$ instead captures the probability of encountering at least one violation from state s onward. This probability equals 1 if $M_t \notin \mathcal{S}$, and otherwise it coincides with the probability of observing a violation at some later time:

$$V_{\mathcal{P}}(t,c) \;=\; \mathbb{P}\left[\max_{i\in[t;T]} \mathbf{1}[M_i \notin \mathcal{S}] = 1 \;\middle|\; M_t\right].$$

In particular, $V_{\mathcal{P}}(s_0)$ gives the probability of observing any violation within the time horizon $[1, T]$.

B Families of energy functions of interest

In this section we propose two families of general-purpose energy shields, and one family of energy shields specifically designed to fit a given specification. Note that some of these functions have isolated non-smooth points, where the function is continuous but not differentiable. These do not pose a theoretical issue, since we can always "glue" the non-smooth endpoints of the two smooth pieces using infinitely differentiable bump functions[1].

B.1 Polynomial

$$\zeta^{Pol}_{\kappa,\alpha,\beta}(x) = \alpha|x-\kappa|^{\beta},$$

where $\beta \in (1,\infty)$, $\kappa \in (0,1)$ and $\alpha \in \left(0, \frac{1}{\max\{\kappa,1-\kappa\}^{\beta}}\right)$. In this family, κ marks the pivoting point, and α and β control the shape, with larger α and β producing steeper functions.

B.2 Exponential

$$\zeta^{Exp}_{\kappa,\alpha,\beta}(x) = \alpha\left(1 - e^{-\beta(x-\kappa)^2}\right),$$

where $\beta \in (0,\infty)$, $\kappa \in (0,1)$, and $\alpha \in \left(0, \frac{1}{1-e^{-\beta(\min\{\kappa,1-\kappa\})^2}}\right)$.

B.3 Construction of a monotonic family of energy functions

Let $p \in (0,1)$, $\varphi = (\tau, \mathcal{S}, \mathcal{L})$ be a specification, with $\mathcal{S} = [L_{\mathcal{S}}, U_{\mathcal{S}}]$ and $\mathcal{L} = [L_{\mathcal{L}}, U_{\mathcal{L}}]$, $\mathcal{L} \subset \mathcal{S}$. We give the construction for $\zeta^{\text{Mon}}_{r;p,\mathcal{S},\mathcal{L}}$. For ease of notation, within this section we will call it just ζ_r. The family of monotonic function is built differently depending on the relative position of p and $\mathcal{L}$.

- **If $p < L_{\mathcal{L}}$.** The family of energy functions is $(\zeta_r)_{r\in R}$, with $R = (0,1)$, and is built as follows. Let $\kappa = (U_{\mathcal{L}} + U_{\mathcal{S}})/2$, $a_r = (1-r)L_{\mathcal{L}} + rU_{\mathcal{L}}$, $C_r = (a_r - p)/(1-p)$, $\alpha_r = (1-r)/r$.

$$\zeta_r(x) = \begin{cases} \zeta_r^1(x) & \text{if } x < a_r, \\ \zeta_r^2(x) & \text{if } a_r \le x \le \kappa, \\ \zeta_r^3(x) & \text{otherwise.} \end{cases} \tag{48}$$

  With the following definitions:

$$\zeta_r^1(x) = C_r + (1-C_r)\left(1 - e^{\frac{x-a_r}{\alpha_r}}\right),$$
$$\zeta_r^2(x) = C_r\left(1 - \frac{x-a_r}{\kappa - a_r}\right)^{\alpha_r},$$
$$\zeta_r^3(x) = 1 - \exp\left(-\left(\tfrac{x-\kappa}{\alpha_r}\right)^m\right).$$

  where m is a fixed integer $m \ge 2$, we choose $m = 2$.
- **If $p > U_{\mathcal{L}}$.** We follow a symmetric construction as the previous case. The family of energy functions is $(\zeta_r)_{r\in R}$, with $R = (0,1)$, and is built as follows. Let $\kappa = (L_{\mathcal{S}} + L_{\mathcal{L}})/2$, $a_r = rL_{\mathcal{L}} + (1-r)U_{\mathcal{L}}$, $C_r = (p - a_r)/p$,

[1]This follows the same rationale as why most theoretical results in convergence of machine learning algorithms work with the non-smooth ReLU activation function.

$\alpha_r = (1-r)/r$.

$$\zeta_r(x) = \begin{cases} \zeta_r^1(x) & \text{if } x < \kappa, \\ \zeta_r^2(x) & \text{if } \kappa \le x \le a_r, \\ \zeta_r^3(x) & \text{otherwise.} \end{cases} \tag{49}$$

With the following definitions:

$$\zeta_r^1(x) = 1 - e^{-\left(\frac{x-\kappa}{\alpha_r}\right)^m}$$

$$\zeta_r^2(x) = C_r \left(1 - \frac{a_r - x}{a_r - \kappa}\right)^{\alpha_r},$$

$$\zeta_r^3(x) = C_r + (1 - C_r)\left(1 - e^{\frac{a_r - x}{\alpha_r}}\right).$$

- **If** $p \in [L_{\mathcal{L}}, U_{\mathcal{L}}]$**.** In this case, the natural bias of the process already aligns with the short-term requirements, so we can choose $\kappa = p$ and use a family of either polynomial or exponential functions. We show here a family of modified polynomial functions that satisfy the monotonicity requirements. The family of energy functions is $(\zeta_r)_{r \in R}$, with $R = (0, 1)$, built as follows. Let $\alpha_r = r/(1-r)$, and $m \ge 2$ fixed, $l_r = \kappa - \frac{1}{\alpha_r^{1/m}}$, $u_r = \kappa + \frac{1}{\alpha_r^{1/m}}$. Then

$$\zeta_r(x) = \begin{cases} \alpha_r |x - \kappa|^m & \text{if } x \in [l_r, u_r] \\ 1 & \text{otherwise.} \end{cases}$$

Claim 2. *The family of functions $(\zeta_r)_{r \in R}$ satisfies the following properties:*

- *It is monotonic in r, i.e., for each $x \in [0, 1]$ and each pair $s, r \in (0, 1)$, if $s \le r$, then $\zeta_s(x) \le \zeta_r(x)$.*
- *The corresponding characteristic function f_r (as defined in Eq. 2) has a fixpoint at $\mu = a_r$.*
- *If $p < L_{\mathcal{L}}$, for all $x \neq \kappa$:*

$$\lim_{r \to 0} \zeta_r(x) = \begin{cases} (L_{\mathcal{L}} - p)/(1-p) & \text{if } x \le L_{\mathcal{L}} \\ 0 & \text{otherwise.} \end{cases}, \quad \lim_{r \to 1} \zeta_r(x) = \begin{cases} (L_{\mathcal{L}} - p)/(1-p) & \text{if } x \in [U_{\mathcal{L}}, \kappa] \\ 1 & \text{otherwise.} \end{cases}$$

- *If $p > U_{\mathcal{L}}$, for all $x \neq \kappa$:*

$$\lim_{r \to 0} \zeta_r(x) = \begin{cases} (p - L_{\mathcal{L}})/p & \text{if } x \ge U_{\mathcal{L}} \\ 0 & \text{otherwise.} \end{cases}, \quad \lim_{r \to 1} \zeta_r(x) = \begin{cases} (p - L_{\mathcal{L}})/p & \text{if } x \in [\kappa, L_{\mathcal{L}}] \\ 1 & \text{otherwise.} \end{cases}$$

- *If $p \in \mathcal{L}$, for all $x \neq \kappa$ and all $l \in \{0, 1\}$, $\lim_{r \to l} \zeta_r(x) = l$.*

All the properties are satisfied by construction. Fig. 5 illustrates them with $m = 2$.

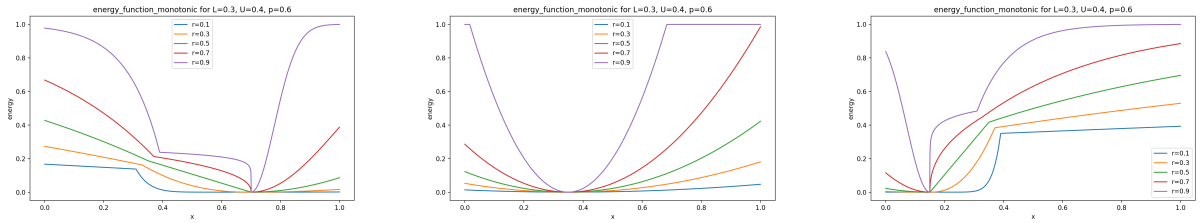


Fig. 5. Monotonic families of functions