Hedging under uncertainty: regret minimization meets exponentially fast convergence

Johanne Cohen[1], Amélie Héliou[2], and Panayotis Mertikopoulos[3]

[1] LRI-CNRS, Université de Paris-Sud,Université Paris-Saclay, France
johanne.cohen@lri.fr,
[2] LIX, Ecole Polytechnique, CNRS, Inria, Université Paris-Saclay, France
amelie.heliou@polytechnique.edu@lri.fr,
[3] CNRS and Univ. Grenoble Alpes, LIG, F–38000, Grenoble, France
panayotis.mertikopoulos@imag.fr,

Abstract. This paper examines the problem of multi-agent learning in N-person non-cooperative games. For concreteness, we focus on the so-called "hedge" variant of the exponential weights (EW) algorithm, one of the most widely studied algorithmic schemes for regret minimization in online learning. In this multi-agent context, we show that *a*) dominated strategies become extinct (a.s.); and *b*) in generic games, pure Nash equilibria are attracting with high probability, even in the presence of uncertainty and noise of arbitrarily high variance. Moreover, if the algorithm's step-size does not decay too fast, we show that these properties occur at a quasi-exponential rate – that is, much faster than the algorithm's $\mathcal{O}(1/\sqrt{T})$ worst-case regret guarantee would suggest.

Keywords: Dominated strategies, exponential weights, Nash equilibrium, no-regret learning.

1 Introduction

In its most basic form, the prototypical framework of online learning can be summarized as follows: at each instance $t = 1, 2, \ldots$ of a repeated decision process, a player selects an action α_t from some finite set $\mathcal{S}$, and they obtain a reward $u_t(\alpha_t)$ based on an a priori unknown payoff function $u_t \colon \mathcal{S} \to \mathbb{R}$. Subsequently, the player observes some problem-specific feedback (for instance, the resulting payoff vector or some estimate thereof), and selects a new action seeking to mazimize their reward over time. In the absence of any other considerations, this objective is usually quantified by asking that the player's *regret* $\operatorname{Reg}(T) \equiv \max_{\alpha\in\mathcal{S}} \sum_{t=1}^{T} [u_t(\alpha) - u_t(\alpha_t)]$ grow sublinearly in T, a property known as "no regret".

Game-theoretic learning is a multi-agent extension of the above in which every player's payoffs are determined by the actions of all players via a fixed mechanism – the *game*. Of course, in many applications, this mechanism may be unknown and/or opaque to the players (who, conceivably, might not even know that they are playing a game). As a result, we are led to the following key questions: if the players of a repeated game agnostically update their strategies following an algorithm that minimizes their individual regret, does the induced sequence of play converge to a Nash equilibrium (or some other rationally justifiable solution concept)? And if so, does this still hold true if the players' observations are contaminated by noise and/or uncertainty?

On the positive side, under no-regret learning, the players' empirical frequencies of play converge to the game's *Hannan set* [12], also known as the set of *coarse correlated equilibria* (CCE) [13].[4] As such, a partial answer to the first question is that coarse correlated equilibria are indeed learnable via no-regret learning. In general however, the Hannan set may contain highly non rationalizable outcomes, so the real answer to this question is "no". For instance, Viossat and Zapechelnyuk [27] recently constructed an example of a 4×4 symmetric game with a coarse correlated equilibrium that assigns positive weight *only* on strictly dominated strategies – an "equilibrium" in name only.

In view of this, our aim in this paper is to examine when no-regret learning leads to a set of rationally justifiable strategies – such as the game's Nash set or the set of undominated strategies. To that end, we focus on the widely used "hedge" variant [9] of the *exponential weights* (EW) algorithm [28, 19], where the probability of choosing an action is proportional to the exponential of its cumulative payoff over time (so better-performing actions are employed exponentially more often). As is well known, "hedging" is min-max optimal in terms of the achieved regret minimization rate [5]. Nonetheless, beyond Hannan consistency, few conclusions can be drawn from this property in a game-theoretic environment where finer convergence criteria apply.

A further complication that arises in game-theoretic learning is that the figure of merit is the convergence of the *actual* sequence of play generated by a learning process – not its time average. In online learning, this averaging comes up naturally because the focus is on the player's regret. In a game-theoretic context however, even if time averages converge, the actual sequence of play may fail to converge altogether (or may do so at a completely different rate), so the players' actual behavior (and the payoffs they obtain) could be drastically different in the two regimes. Thus, given that regret-based convergence results offer little insight on the "last iterate" of the process, the analysis of the latter requires a completely different set of tools and techniques.

1.1 Our results

In their recent paper, Viossat and Zapechelnyuk [27] showed that, in general, the set of coarse correlated equilibria may contain non rationalizable strategies supported exclusively on strictly dominated strategies. Nevertheless, by leveraging the consistent negative reinforcement of actions that perform badly in past instances of play, we show that hedging eliminates dominated strategies with probability 1. Moreover, we show that the rate of elimination is $\mathcal{O}(\exp(-c\sum_{t=1}^{T}\gamma_t))$

[4] As the second name suggests, this set contains the game's set of correlated equilibria (CE) – and hence, the game's set of Nash equilibria as well.

for some positive constant $c > 0$, where γ_t is a variable step-size parameter; in other words, dominated strategies become extinct exponentially fast if γ_t decays slower than $1/t$.

With respect to equilibrium convergence, we show that hedging in generic games converges locally to pure Nash equilibria with high probability, and the convergence is global with probability 1 if the game's equilibrium satisfies a certain variational inequality. The only added caveat for this result is that the algorithm's step-size parameter must satisfy a summability condition which precludes the use of very aggressive step-size policies. Nevertheless, the algorithm still achieves an exponential $\mathcal{O}(e^{-cT^{1-b}})$ convergence rate for step-size sequences of the form $\gamma_t \sim 1/t^b$, $b \in (1/2, 1)$.

To account for the fact that players may not have access to perfect payoff observations, we assume throughout that players can only estimate their payoff vectors up to a possibly unbounded error with arbitrarily high variance. This uncertainty is countered by means of a judicious choice of the algorithm's step-size parameter γ_t which can be used to control the weight with which new observations enter the algorithm at a given stage. This is made possible by exploiting results from martingale limit theory and the theory of stochastic approximation.

1.2 Related work

Algorithms and dynamics for learning in games have received considerable attention over the last few decades. Such procedures can be divided into two broad categories, depending on whether they evolve in continuous or discrete time: the former includes the numerous dynamics for learning and evolution (see [24] for a survey), whereas the latter focuses on learning algorithms for infinitely iterated games (such as fictitious play and its variants). In this paper, we focus exclusively on discrete-time algorithms.

In this framework, it is natural to consider agents who learn from their experience by small adjustments in their behavior based on local – and possibly imperfect – information. Several such approaches in the literature can be viewed as *decentralized no-regret dynamics* – for example the multiplicative/exponential weights algorithm and its variants [28, 9, 19], Follow the Regularized/Perturbed Leader [14], etc. Indeed, regret bounds can be used to guarantee that each player's utility approaches long-term optimality in adversarial environments, a natural first step towards long-term rational behavior. For example, it has been shown in [2, 23] that the sum of utilities approaches an approximate optimum, and there is convergence of time averages towards an equilibrium in two-player zero-sum games [3, 7, 9]. In all these examples, the players' average regret vanishes at the worst-case rate of $\mathcal{O}(1/\sqrt{T})$ where T denotes the play horizon. This convergence rate was recently improved by Syrgkanis *et al.* [26] for a wide class of N-player normal form games using a natural class of regularized learning algorithms. However, the convergence results established in [26] concerned *a*) the set of coarse correlated equilibria (which may contain highly non rationalizable strategies); and *b*) the "long-run average" $\bar{x}_T = T^{-1} \sum_{t=1}^{T} x_t$ of the actual sequence of play. By contrast, our paper focuses squarely on the algorithm's "last iterate" (which determines the players' rewards at each stage), and finer rationality properties (such as the elimination of dominated strategies or convergence to pure Nash equilibria) that cannot be deduced from coarse equilibrium convergence results.

The HEDGE algorithm received much attention in various fields such as optimization [1], multi-armed bandit problems [4], and in general algorithmic game theory. In particular, the number of payoff queries needed to compute approximate correlated equilibria has been studied in [10]: upper and lower bounds have been derived, as well as reductions between problems, such as the reduction of the problem of verifying an approximate well supported Nash equilibrium to the problem of computing a well supported Nash equilibrium under some assumptions.

Kleinberg et al. [15] studied the behavior of the dynamic of the HEDGE algorithm for some particular load balancing games (the so-called atomic load balancing games) in the "*bulletin board*" model. In this latter model, players know the actual payoff of each strategy according to the actual strategies played. They proved that if all players play according to the same mixed strategy, the dynamics of the history of play converge, and the limit is necessarily a stable distribution over states, such as a mixed Nash or a correlated equilibrium. Furthermore, the average performance of the dynamics has been analyzed in atomic load balancing games. Recently, in a similar spirit, Foster et al. [8] showed that some variants of HEDGE algorithms are such that the average of the outcome converge rapidly to an approximation of the optimum in smooth games. In parallel, Krichene *et al.* [16] extended the result to congestion games, and proved that a discounted variant of the HEDGE algorithm converges to the set of Nash equilibria in the sense of Cesàro means (time averages), while strong convergence can be guaranteed with some additional conditions. Their proof is based on the so-called *Kullback–Leibler* (KL) divergence. Finally, Coucheney *et al.* [6] also showed that a "penalty-regulated" variant of the HEDGE algorithm with bandit feedback converges to logit equilibria in congestion games, but their techniques do not extend to actual Nash equilibria. In the current paper, we do not restrict ourselves to congestion games; instead, we consider generic games that admit a Nash equilibrium in pure strategies, of which congestion and potential games are a special case.

2 Preliminaries

Throughout the paper, we focus on games that are played by a (finite) set $\mathcal{N} = \{1, \dots, N\}$ of N *players* (or *agents*). Each player $i \in \mathcal{N}$ is assumed to have a finite set of *actions* (or *pure strategies*) $\mathcal{S}_i$, and the players' preferences for one action over another are represented by each action's *utility* (or *payoff*). Specifically, as players interact with each other, the individual payoff of each player is given by a function $u_i \colon \mathcal{S} \equiv \prod_i \mathcal{S}_i \to \mathbb{R}$ of all players' actions, and each agent seeks to maximize the utility $u_i(\alpha_i; \alpha_{-i})$ of their chosen action $\alpha_i \in \mathcal{S}_i$

against the action profile α_{-i} of his opponents.[5] A game is then called (weakly) *generic* if there are no unilateral payoff ties, i.e. if $u_i(\alpha_i; \alpha_{-i}) \neq u_i(\beta_i; \alpha_{-i})$ for all $\alpha_i, \beta_i \in \mathcal{S}_i$, $i \in \mathcal{N}$.

Players can also use *mixed strategies* by playing probability distributions $x_i = (x_{i\alpha_i})_{\alpha_i \in \mathcal{S}_i} \in \Delta(\mathcal{S}_i)$ over their action sets $\mathcal{S}_i$. The resulting probability vector x_i is called the *mixed strategy* of the i-th player and the set $\mathcal{X}_i = \Delta(\mathcal{S}_i)$ is the corresponding mixed strategy space of player i; aggregating over players, we also write $\mathcal{X} = \prod_i \mathcal{X}_i$ for the game's *strategy space*, i.e. the space of all mixed strategy profiles $x = (x_i)_{i \in \mathcal{N}}$.

In this context (and in a slight abuse of notation), the expected payoff of the i-th player in the mixed strategy profile $x = (x_1, \dots, x_N)$ is

$$u_i(x) = \sum_{\alpha_1 \in \mathcal{S}_1} \cdots \sum_{\alpha_N \in \mathcal{S}_N} u_i(\alpha_1, \dots, \alpha_N)\, x_{1\alpha_1} \cdots x_{N\alpha_N}. \tag{1}$$

Accordingly, if player i plays the pure strategy $\alpha_i \in \mathcal{S}_i$, we will write

$$v_{i\alpha_i}(x) = u_i(\alpha_i; x_{-i}) = u_i(x_1, \dots, \alpha_i, \dots, x_N) \tag{2}$$

for the payoff corresponding to α_i, and $v_i(x) = (v_{i\alpha_i}(x))_{\alpha_i \in \mathcal{S}_i}$ for the resulting *payoff vector* of player i. A player's expected payoff may thus be written as

$$u_i(x) = \sum_{\alpha_i \in \mathcal{S}_i} x_{i\alpha_i} v_{i\alpha_i}(x) = \langle v_i(x) | x_i \rangle, \tag{3}$$

where $\langle v_i | x_i \rangle$ denotes the canonical bilinear pairing between v_i and x_i.

A fundamental rationality principle in game theory is that, assuming full knowledge of the game, a player would have no incentive to play an action that always yields suboptimal payoffs with respect to another. To formalize this, $\alpha_i \in \mathcal{S}_i$ is called (*strictly*) *dominated* by β_i (and we write $\alpha_i \prec \beta_i$) if

$$u_i(\alpha_i; \alpha_{-i}) < u_i(\beta_i; \alpha_{-i}) \quad \text{for all } \alpha_{-i} \in \mathcal{S}_{-i} \equiv \textstyle\prod_{j \neq i} \mathcal{S}_j,\ i \in \mathcal{N}. \tag{4}$$

Extending the notion of strategic dominance, the most widely used solution concept in game theory is that of a *Nash equilibrium* (NE), i.e. a state $x^* \in \mathcal{X}$ which is unilaterally stable in the sense that

$$u_i(x_i^*; x_{-i}^*) \geq u_i(x_i; x_{-i}^*) \quad \text{for all } x_i \in \mathcal{X}_i,\ i \in \mathcal{N}, \tag{NE}$$

or, equivalently, writing $\operatorname{supp}(x)$ for the support of x:

$$v_{i\alpha_i}(x^*) \geq v_{i\beta_i}(x^*) \quad \text{for all } \alpha_i \in \operatorname{supp}(x_i^*) \text{ and all } \beta_i \in \mathcal{S}_i,\ i \in \mathcal{N}. \tag{5}$$

If equilibrium x^* is *pure* (i.e. $\operatorname{supp}(x_i^*) = \{\alpha_i^*\}$ for some $\alpha_i^* \in \mathcal{S}_i$ and all $i \in \mathcal{N}$), then it is called a *pure equilibrium*. In generic games, a pure equilibrium satisfies (5) as a strict inequality for all $\beta_i \notin \operatorname{supp}(x_i^*)$, $i \in \mathcal{N}$, so we sometimes refer to pure equilibria in generic games as *strict*. Such equilibria will play a key role in our analysis, so we provide a convenient variational characterization below:

[5] In the above $(\alpha_i; \alpha_{-i})$ is shorthand for $(\alpha_1, \dots, \alpha_i, \dots, \alpha_N)$, used here to highlight the action of player i against that of all other players.

Proposition 1. *In generic games, x^* is a pure equilibrium if and only if*

$$\langle v(x) | x - x^* \rangle \leq -\tfrac{1}{2}\mu \|x - x^*\| \quad \textit{for some } \mu > 0 \textit{ and for all } x \textit{ near } x^*, \tag{6}$$

[6] *where* $\|x\| = \sum_i \sum_{\alpha \in \mathcal{S}_i} |x_{i\alpha_i}|$ *denotes the* L^1*-norm of* x.

Clearly, if (6) holds for all $x \in \mathcal{X}$, then x^* is the unique Nash equilibrium of the game. In particular, it is easy to verify that this is the case in the Prisoner's Dilemma and its variants, generic competition games, potential games with a unique equilibrium, etc. At a more fundamental level, (6) is closely related to the notion of a globally evolutionarily stable state (GESS) in evolutionary game theory [24]; for more details, we refer the reader to the variational stability analysis of [20].

3 Hedging under uncertainty

The algorithm that we examine is the so-called "hedge" variant of the exponential weights algorithm [9]. In a nutshell, the main idea of the algorithm is as follows: At each stage $t = 1, 2, \ldots$ of the process, players maintain and update a "performance score" for each of their actions (pure strategies) based on each action's cumulative payoff up to stage t. These scores are then converted to mixed strategies by assigning exponentially higher probability to actions with higher scores, and a new action is drawn based on these mixed strategies.

More precisely, this iterative process can be encoded as follows:

Algorithm 1.1 HEDGE with variable step-size γ_t

```
Each player $i \in \mathcal{N}$ chooses an initial score vector $y_i(1)$
for each round $t$
    Each player $i \in \mathcal{N}$ plays $x_i(t) = \Lambda_i(y_i(t))$ where the logit map $\Lambda_i$ is defined as
        $\Lambda_i(y_i) = \frac{1}{\sum_{\alpha \in \mathcal{S}_i} \exp(y_{i\alpha})} (\exp(y_{i\alpha}))_{\alpha \in \mathcal{S}_i}.$    (7)
    Each player $i \in \mathcal{N}$ draws a pure strategy $\alpha_i(t)$ according to $x_i(t)$
    Each player $i \in \mathcal{N}$ gets an estimate $\hat{v}_i(t)$ of their payoff vector $v_i(\alpha(t))$
    Each player $i \in \mathcal{N}$ updates their score vectors as
        $y_i(t+1) = y_i(t) + \gamma_t \hat{v}_i(t),$    (8)
end for
```

Mathematically, the above algorithm can be expressed as

$$\begin{aligned} x_i(t) &= \Lambda_i(y_i(t)), \\ y_i(t+1) &= y_i(t) + \gamma_t \hat{v}(t), \end{aligned} \tag{HEDGE}$$

with $y_i(1)$ initialized arbitrarily. Motivated by practical implementation issues (especially in large networks and telecommunication systems), this formulation

[6] x is near x^* if there exists an ε such that $\|x - x^*\| \leq \varepsilon$

further assumes that players have *imperfect knowledge* of their payoff vectors $v_i(x(t))$ at each iteration of the algorithm – for instance, contaminated by measurement errors or other uncertainty factors. To formalize this, we will consider a general feedback model of the form

$$\hat{v}_i(t) = v_i(\alpha(t)) + \xi_i(t), \tag{9}$$

where the error process $\xi = (\xi_i)_{i \in \mathcal{N}}$ is a L^2-bounded martingale difference sequence with respect to the history $\mathcal{F}_t$ of the process $(y(t), x(t), \alpha(t), \hat{v}(t))$. In other words, we assume that $\xi(t)$ satisfies the statistical hypotheses

1. *Zero-mean:*

$$\mathbb{E}[\xi(t) \mid \mathcal{F}_{t-1}] = 0 \qquad \text{for all } t = 1, 2, \dots \text{ (a.s.)}; \tag{H1}$$

2. *Finite mean squared error:* there exists some $\sigma > 0$ such that

$$\mathbb{E}[\|\xi(t)\|_\infty^2 \mid \mathcal{F}_{t-1}] \leq \sigma^2 \quad \text{for all } t = 1, 2, \dots \text{ (a.s.)}. \tag{H2}$$

Put differently, Hypotheses (H1) and (H2) simply mean that the payoff vector estimates $\hat{v}_i$ are *conditionally unbiased and bounded in mean square*, i.e.

$$\mathbb{E}[\hat{v}(t) \mid \mathcal{F}_{t-1}] = v(x(t)), \tag{10a}$$

$$\mathbb{E}[\|\hat{v}(t)\|_\infty^2 \mid \mathcal{F}_{t-1}] \leq V^2, \tag{10b}$$

where $V > 0$ is a finite positive constant (in the noiseless case $\xi = 0$ the constant V is simply a bound on the players' maximum absolute payoff).[7] Thus, Hypotheses (H1) and (H2) allow for a broad range of noise distributions, including all compactly supported, (sub-)Gaussian, (sub-)exponential and log-normal distributions.

4 Rationality analysis and results

In this section, we present our main convergence results for (HEDGE). We begin with the fact that hedging bypasses the negative result of Viossat and Zapachelnyuk [27], and does not lead to non-rationalizable, strategically dominated outcomes:

PM: Stopped here.

[7] Note here that (10a) is phrased in terms of the players' *mixed* strategy profile $x(t)$, not the action profile $\alpha(t) = (\alpha_i(t); \alpha_{-i}(t))$ which is chosen based on $x(t)$ at stage t. To see that (H1) indeed implies (10a), simply recall that $x_i(t) = \Lambda_i(y_i(t))$ so

$$\begin{aligned}\mathbb{E}[\hat{v}_{i\alpha_i}(t) \mid \mathcal{F}_{t-1}] &= \sum_{\alpha_{-i} \in \mathcal{S}_{-i}} \big[u_i(\alpha_i; \alpha_{-i}) x_{\alpha_{-i}}(t) + \mathbb{E}[\xi_{i\alpha_i}(t) \mid \mathcal{F}_{t-1}]\big] \\ &= u_i(\alpha_i; x_{-i}(t)) = v_{i\alpha_i}(x(t)).\end{aligned} \tag{11}$$

Theorem 1. *Suppose that* (HEDGE) *is run with a step-size sequence of the form* $\gamma_t \propto 1/t^b$ *for some* $b \leq 1$ *(not necessarily positive), and noisy payoff observations satisfying Hypotheses* (H1) *and* (H2)*. If* $\alpha_i \in \mathcal{S}_i$ *is dominated, there exists some* $c > 0$ *such that*

$$x_{i\alpha_i}(T) = \mathcal{O}(\exp(-c\textstyle\sum_{t=1}^T \gamma_t)) \quad \textit{with probability } 1. \tag{12}$$

In particular, if $b < 1$*,* α_i *becomes extinct exponentially fast (a.s.).*

Proof. Suppose that $\alpha_i \prec \beta_i$ for some $\beta_i \in \mathcal{S}_i$. Then, suppressing the player index i for simplicity, we get

$$\begin{aligned} y_\beta(T) - y_\alpha(T) &= c_{\beta\alpha} + \sum_{t=1}^T \gamma_t \left[\hat{v}_\beta(t) - \hat{v}_\alpha(t)\right] \\ &= c_{\beta\alpha} + \sum_{t=1}^T \gamma_t \left[v_\beta(x(t)) - v_\alpha(x(t))\right] + \sum_{t=1}^T \gamma_t \zeta_t, \end{aligned} \tag{13}$$

where we set $c_{\beta\alpha} = y_\beta(0) - y_\alpha(0)$ and $\zeta_t = \hat{v}_\beta(t) - v_\beta(x(t)) - [\hat{v}_\alpha(t) - v_\alpha(x(t))]$. Since $\alpha \prec \beta$, there exists some $\mu > 0$ such that $v_\beta(x) - v_\alpha(x) \geq \mu$ for all $x \in \mathcal{X}$. Then, (13) yields

$$y_\beta(T) - y_\alpha(T) \geq c_{\beta\alpha} + \theta_T \left[\mu + \frac{\sum_{t=1}^T \gamma_t \zeta_t}{\theta_T}\right], \tag{14}$$

with $\theta_T = \sum_{t=1}^T \gamma_t$.

Since $\mathbb{E}[\zeta_t \mid \mathcal{F}_{t-1}] = 0$ and $\sup_t \mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] < \infty$ by the reformulation (10a) and (10b) of Hypotheses (H1) and (H2) respectively, the law of large numbers for martingale difference sequences [11, Theorem 2.18] gives $\theta_T^{-1} \sum_{t=1}^T \gamma_t \zeta_t \to 0$ (a.s.), provided that $\sum_{t=1}^\infty (\gamma_t/\theta_t)^2 < \infty$ and $\sum_{t=1}^\infty \gamma_t = \infty$. Given that this last assumption is satisfied for all $b \leq 1$, we readily get $\theta_T^{-1} \sum_{t=1}^T \gamma_t \zeta_t \to 0$ (a.s.), so for all $c \in (0, \mu)$, there exists some (a.s. finite) T_0 such that if $T \geq T_0$, then $y_\beta(T) - y_\alpha(T) \geq c\theta_T$. Recall that $x_\alpha(T) = \Lambda_\alpha(y(T))$. We thus get

$$x_\alpha(T) = \frac{e^{y_\alpha(T)}}{\sum_\kappa e^{y_\kappa(T)}} \leq \frac{e^{y_\alpha(T)}}{e^{y_\beta(T)}} = e^{y_\alpha(T) - y_\beta(T)} \leq e^{-c\sum_{t=1}^T \gamma_t} \quad \text{(a.s.)}, \tag{15}$$

and our proof is complete. □

Remark 1. It should be noted here that the elimination of dominated strategies with imperfect knowledge of the game's payoffs is by no means a given. For instance, if players play a greedy best response scheme at each round and the payoff observation errors are not supported on a small, compact set, dominated strategies will be played infinitely often (simply because at each round, any strategy could be erroneously perceived as a best response).[8] With this in mind,

[8] Recall also the counterexample of [27] discussed in the introduction.

the fact that the rate of elimination (12) *improves* with more aggressive – even *increasing* – step-size sequences γ_t is somewhat surprising because it suggests that players can employ (HEDGE) in a very greedy fashion and achieve fast dominated strategy extinction rates, even in the presence of arbitrarily high estimation errors.

Remark 2. Regarding the number of players and actions per player, our proof reveals that c depends only on the player's payoffs – specifically, we can take $c = \frac{1}{2}\min_{\alpha_{-i}\in\mathcal{S}_{-i}}[u_i(\beta_i;\alpha_{-i}) - u_i(\alpha_i;\alpha_{-i})] > 0$. In other words, the algorithm's half-life is asymptotically *independent* of the size of the game, and only depends on the players' relative payoff differences.

Remark 3. Finally, we should note that Theorem 1 also extends to the case of iteratively dominated strategies.[9] As such, Theorem 1 implies that the sequence of play induced by (HEDGE) actually converges to the set of iteratively undominated strategies of the game.

We now turn to the convergence properties of (HEDGE) in generic games that admit pure Nash equilibria:

Theorem 2. *Fix a confidence level $\varepsilon > 0$ and suppose that* (HEDGE) *is run with a small enough (depending on ε) step-size γ_t satisfying $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$ and $\sum_{t=1}^{\infty}\gamma_t = \infty$, and imperfect payoff observations satisfying Hypotheses* (H1) *and* (H2)*. If x^* is a pure equilibrium of a generic game and* (HEDGE) *is initialized not too far from x^*, then we have*

$$\mathbb{P}\Big(\|x(T) - x^*\| \leq C' e^{-c\sum_{t=1}^{T}\gamma_t} \textit{ for all } t\Big) \geq 1 - \varepsilon, \tag{16}$$

where $c > 0$ is a constant that only depends on the game and $C' > 0$ is a (*random*) *constant that depends on the initialization of* (HEDGE)*. In particular, under the stated assumptions, $x(t) \to x^*$ with probability at least $1 - \varepsilon$.*

Corollary 1. *With assumptions as above, if* (HEDGE) *is run with a step-size of the form $\gamma_t = \gamma/t^b$ for some sufficiently small $\gamma > 0$ and $b \in (1/2, 1)$, we have*

$$\mathbb{P}\Big(\|x(t) - x^*\| = \mathcal{O}\left(e^{-\frac{c\gamma}{1-b}t^{1-b}}\right)\Big) \geq 1 - \varepsilon. \tag{17}$$

Relegating the (somewhat involved) proof of Theorem 2 to Appendix B, we note here that, in contrast to Theorem 1, the "$L^2 - L^1$" summability requirement $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$ and $\sum_{t=1}^{\infty}\gamma_t = \infty$ constrains the admissible step-size policies that lead to pure equilibrium (for instance, constant step-size policies are no longer admissible). In particular, the most aggressive step-size that satisfies the assumptions of Theorem 2 is $\gamma_t \propto t^{-b}$ for some b close (but not equal) to $1/2$, leading to a convergence rate of $\lambda^{t^{1-b}}$ for some real $\lambda < 1$ (cf. Corollary 1).

[9] This can be shown by an induction argument on the rounds of elimination of dominated strategies as in [18].

This bound on b is due to the second moment growth bound required by Doob's maximal inequality; if there is finer control on the moments of the noise process ξ (for instance, if the noise is sub-exponential), the lower bound $b > 1/2$ can be pushed all the way down to $b > 0$, implying a quasi-linear convergence rate.

As was hinted above, the main idea behind the proof of Theorem 2 is to use Doob's maximal inequality for martingales to show that the probability of $x(t)$ escaping the basin of attraction of a pure Nash equilibrium x^* can be made arbitrarily small if the algorithm's step-size is chosen appropriately. Building on this, if x^* satisfies the variational inequality (6) throughout $\mathcal{X}$, we have the stronger result:

Theorem 3. *Suppose that* (HEDGE) *is run with a step-size sequence* γ_t *such that* $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$ *and* $\sum_{t=1}^{\infty} \gamma_t = \infty$ *and imperfect payoff observations satisfying Hypotheses* (H1) *and* (H2)*. If* x^* *satisfies* (6) *for all* $x \in \mathcal{X}$*, then:*

1. $\lim_{T\to\infty} x(T) = x^*$ *(a.s.).*
2. *There exists a (deterministic) constant* $c > 0$ *such that*

$$\|x(T) - x^*\| = \mathcal{O}\left(e^{-c\sum_{t=1}^{T}\gamma_t}\right). \tag{18}$$

Corollary 2. *With assumptions as above, if* (HEDGE) *is run with a step-size of the form* $\gamma_t = \gamma/t^b$ *some* $b \in (1/2, 1)$*, we have*

$$\|x(T) - x^*\| = \mathcal{O}(e^{-\frac{\mu\gamma}{1-b}T^{1-b}}). \tag{19}$$

In contrary to Theorems 1 and 2, the proof of Theorem 3 relies heavily on the so-called *Kullback–Leibler* (KL) divergence [17], defined here as

$$D_{\mathrm{KL}}(x^*, x) = \sum_{i\in\mathcal{N}} \sum_{\alpha_i\in\mathcal{S}_i} x^*_{i\alpha_i} \log\frac{x^*_{i\alpha_i}}{x_{i\alpha_i}} \quad \text{for all } x \in \mathcal{X}^\circ. \tag{20}$$

The KL divergence is a positive-definite, asymmetric distance measure that is tailored to the analysis of the replicator dynamics [29, 24, 18]. By using this divergence as a discrete-time Lyapunov function, we show that x^* is a *recurrent point* of the process $x(t)$, i.e. $x(t)$ visits any neighborhood of x^* infinitely many times. We then use a stochastic approximation argument to show that the process actually converges to x^* at an asymptotic rate of $\mathcal{O}(e^{-c\sum_{t=1}^{T}\gamma_t})$.

The step-size assumption in the statement of Theorem 3 is key in achieving this, but it is important to note it can be relaxed to $\sum_{t=1}^{T}\gamma_t^2 \big/ \sum_{t=1}^{T}\gamma_t \to 0$ if the players' feedback noise is bounded (for instance, if players have access to their actual pure payoff information). When this is the case, it is possible to achieve an $\mathcal{O}(e^{-cT^{1-b}})$ convergence rate for any $b > 0$ by using a step-size sequence of the form $\gamma_t \propto 1/t^b$. Finally, regarding the constant $c > 0$, our proof shows that it can be chosen in roughly the same way as the respective coefficient of Theorem 1, implying that the rate of extinction of dominated strategies is of the same order as that of convergence to pure equilibria.

5 Conclusions

Our main goal is to analyse the convergence properties of the "hedge" variant of the exponential weights algorithm [9] in generic N-player games that admit a Nash equilibrium in pure strategies. Motivated by the applications of game theory to data networks, we assumed throughout that players only have acces to imperfect observations of their pure payoff vector. In this setting, using techniques drawn from the theory of stochastic approximation and martingale limit theory, we showed that (i) dominated strategies become extinct (a.s.); (ii) pure equilibria are locally attracting with high probability; and (iii) pure equilibria that satisfy a certain variational stability condition are globally attracting with probability 1. Moreover, despite the uncertainty in the players' payoff observations, we showed that the elimination of dominated strategies is exponentially fast – in stark contrast to more unrefined best response schemes which, under uncertainty, may lead to playing dominated strategies infinitely often. Using the theory of stochastic approximation and discrete-time martingale processes, we showed that the algorithm's convergence (local or global, depending on the context) to a pure Nash equilibrium is also exponentially fast, even in the presence of uncertainty and noise of arbitrary magnitude. These results apply to all generic games that admit a Nash equilibrium in pure strategies, and not only to the class of coordination and anti-coordination (or congestion) games that have been the traditional focus of the game-theoretic learning literature.

Our results can be extended to a significantly broader class of no-regret learning algorithms based on "following the regularized leader" [14], of which "hedge" can be seen as a special case; however, we do not undertake this analysis here due to lack of space. A further important extension of this work would be the so-called "bandit feedback" environment where players are only able to observe the payoff of the action that they actually played and can only estimate the payoff of their other actions via the game's history – possibly by introducing a regularized sampling process as in the very recent work [10]. Another important issue is that of asynchronicity, namely when players update at different times and there is a delay between playing and receiving feedback. We intend to explore these directions in future work.

References

1. S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.
2. A. Blum, M. T. Hajiaghayi, K. Ligett, and A. Roth. Regret minimization and the price of total anarchy. In *STOC '08*, pages 373–382. ACM, 2008.
3. A. Blum and Y. Mansour. Learning, regret minimization, and equilibria. In Noam Nisan, Tim Roughgarden, Eva Tardos, and V. V. Vazirani, editors, *Algorithmic Game Theory*, chapter 4. Cambridge University Press, 2007.
4. S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

5. N. Cesa-Bianchi and G.r Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.
6. P. Coucheney, B. Gaujal, and P. Mertikopoulos. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*, 40(3):611–633, 2015.
7. D. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997.
8. D. J Foster, T. Lykouris, K. Sridharan, E. Tardos, et al. Learning in games: Robustness of fast convergence. In *NIPS*, pages 4727–4735, 2016.
9. Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.
10. P. W. Goldberg and A. Roth. Bounds for the query complexity of approximate equilibria. *ACM Transactions on Economics and Computation*, 24:1–24:25, 2016.
11. P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application.* Probability and Mathematical Statistics. Academic Press, New York, 1980.
12. J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games, Volume III*, volume 39 of *Annals of Mathematics Studies*, pages 97–139. Princeton University Press, Princeton, NJ, 1957.
13. S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, September 2000.
14. A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.
15. R. Kleinberg, G. Piliouras, and E. Tardos. Load balancing without regret in the bulletin board model. *Distributed Computing*, 24(1):21–29, 2011.
16. W. Krichene, B. Drighès, and A. M Bayen. Learning nash equilibria in congestion games. *arXiv preprint arXiv:1408.0017*, 2014.
17. S. Kullback and R. A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
18. R. Laraki and P. Mertikopoulos. Higher order game dynamics. *Journal of Economic Theory*, 148(6):2666–2695, November 2013.
19. N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.
20. P. Mertikopoulos, E. Veronica Belmega, R. Negrel, and L. Sanguinetti. Distributed stochastic optimization via matrix exponential learning. http://arxiv.org/abs/1606.01190, 2016.
21. P. Mertikopoulos and W. H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 2016.
22. R. T. Rockafellar. *Convex Analysis.* Princeton University Press, NJ, 1970.
23. T. Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):32, 2015.
24. W. H. Sandholm. *Population Games and Evolutionary Dynamics.* Economic learning and social evolution. MIT Press, Cambridge, MA, 2010.
25. S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.
26. V. Syrgkanis, A. Agarwal, H. Luo, and R. E Schapire. Fast convergence of regularized learning in games. In *NIPS*, pages 2989–2997, 2015.
27. Y. Viossat and A. Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013.
28. V. G. Vovk. Aggregating strategies. In *COLT '90*, pages 371–383, 1990.
29. J. W. Weibull. *Evolutionary Game Theory.* MIT Press, Cambridge, MA, 1995.

A Auxiliary results

In this appendix, we collect some auxiliary results needed for our convergence analysis. We begin with the variational characterization (6) of strict Nash equilibria:

Proof of Proposition 1. Assume that x^* is a pure Nash equilibrium. Then, for all $i \in \mathcal{N}$, we have

$$\begin{aligned}\langle v_i(x)|x_i - x_i^*\rangle &= u_i(x_i; x_{-i}) - u_i(x_i^*; x_{-i}) \\ &= \sum_{\alpha_i \in \mathcal{S}_i} (x_{i\alpha_i} - x^*_{i\alpha_i})\, u_i(\alpha_i; x_{-i}) \\ &= \sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i} u_i(\alpha_i; x_{-i}) + x_{i\alpha_i^*}\, u_i(\alpha_i^*; x_{-i}) - u_i(\alpha_i^*; x_{-i}) \\ &= \sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i} \left[u_i(\alpha_i; x_{-i}) - u_i(\alpha_i^*; x_{-i})\right], \end{aligned} \tag{A.21}$$

where the first line is a consequence of (3) while the last one follows by noting that $\sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i} = 1 - x_{i\alpha_i^*}$ and rearranging. Now, by continuity of u_i – and the genericity of the game – there exists a real number $\mu > 0$ and a neighborhood U of x^* in $\mathcal{X}$ such that for all $\alpha_i \in \mathcal{S}_i \setminus \{\alpha_i^*\}$, $u_i(\alpha_i^*; x_{-i}) - u_i(\alpha_i; x_{-i}) \geq \mu > 0$. Therefore:

$$\langle v_i(x)|x_i - x_i^*\rangle \leq -\mu \sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i} \tag{A.22}$$

Hence, combining Eqs. (A.21) and (A.22), we get the bound

$$\langle v(x)|x - x^*\rangle = \sum_{i \in \mathcal{N}} \langle v_i(x)|x_i - x_i^*\rangle \leq -\mu \sum_{i \in \mathcal{N}} \sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i} \leq -\frac{\mu}{2}\|x - x^*\|, \tag{A.23}$$

where the last inequality follows from the fact that $x^*_{i\alpha_i} = 0$ if $\alpha_i \neq \alpha_i^*$ so $\|x_i - x_i^*\| = 1 - x_{i\alpha_i^*} + \sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i} = 2\sum_{\alpha_i \neq \alpha_i^*} x_{i\alpha_i}$.

Conversely, assume now that x^* satisfies (6) but is not a strict Nash equilibrium, so $v_{i\alpha_i}(x^*) \leq v_{i\beta_i}(x^*)$ for some $\alpha_i \in \operatorname{supp}(x_i^*)$, $\beta_i \in \mathcal{S}_i \setminus \{\alpha_i\}$, $i \in \mathcal{N}$. If we take $x_i = x_i^* + \lambda(e_{i\beta_i} - e_{i\alpha_i})$ and $x_{-i} = x^*_{-i}$ with $\lambda > 0$ small enough, then we get

$$\langle v(x)|x - x^*\rangle = \langle v_i(x)|x_i - x_i^*\rangle = \lambda v_{i\beta_i}(x^*) - \lambda v_{i\alpha_i}(x^*) \geq 0, \tag{A.24}$$

in contradiction to (6) which yields $\langle v(x)|x - x^*\rangle < 0$. □

We now prove two important properties of the logit map and the Kullback–Leibler divergence:

Proposition A.2. *Let $\mathcal{S} = \{1, \dots, S\}$ be a finite set and let $\Delta \equiv \Delta(\mathcal{S})$ denote the $(S-1)$-dimensional simplex spanned by $\mathcal{S}$. Let M be a positive real. Then:*

1. *If $x^* \in \Delta$ is pure (i.e.* $\operatorname{supp}(x^*) = \{\alpha^*\}$ *for some $\alpha^* \in \mathcal{S}$), then the set $U_M = \{x = \Lambda(y) : y_\alpha - y_{\alpha^*} \leq -M, \text{ for all } \alpha \neq \alpha^*\}$ is a neighborhood of x^* in Δ°. Furthermore, if M is sufficiently large, then U_M is contained in a $\|\cdot\|$-ball centered at x^*;*
2. *Let $x = \Lambda(y)$, $x' = \Lambda(y')$ for some $y, y' \in \mathbb{R}^n$. Then, we have*

$$D_{\mathrm{KL}}(x^*, x') \leq D_{\mathrm{KL}}(x^*, x) + \langle y' - y | x - x^* \rangle + \frac{1}{2}\|y' - y\|_\infty^2. \tag{A.25}$$

Proof. For our first claim, assume to the contrary that U_M is not a neighborhood of x^* in $\mathcal{X}^\circ$, so there exists a sequence $x_k = \Lambda(y_k)$ in $\mathcal{X}^\circ$ that converges to x^*, but $x_k \notin U_M$ for all k. By passing to a subsequence if necessary, there exists some $\alpha \in \mathcal{S}$ such that $y_{k,\alpha} - y_{k,\alpha^*} > -M$ and $y_{k,\alpha} \geq y_{k,\beta}$ for all β (simply pick a constant subsequence of $\arg\max\{y_{k,\beta} : y_{k,\beta} - y_{k,\alpha^*} > -M\}$ if needed). We then get

$$x_{k,\alpha} = \frac{e^{y_{k,\alpha}}}{\sum_{\beta \in \mathcal{S}} e^{y_{k,\beta}}} = \frac{1}{e^{y_{k,\alpha^*} - y_{k,\alpha}} + \sum_{\beta \neq \alpha^*} e^{y_{k,\beta} - y_{k,\alpha}}} \geq \frac{1}{e^M + S - 1}, \tag{A.26}$$

contradicting the original assumption $x_{k,\alpha} \to 0$ (since $x_k \to x^*$).

For the converse implication (namely that U_M is contained in a ball around x^*), fix some $\delta > 0$ and let $z_\alpha = y_\alpha - y_{\alpha^*}$, $\alpha \in \mathcal{S} \setminus \{\alpha^*\}$. Then, letting $x = \Lambda(y)$ for some $y \in U_M$, we have

$$x_{\alpha^*} = \frac{e^{y_{\alpha^*}}}{\sum_{\alpha \in \mathcal{S}} e^{y_\alpha}} = \frac{1}{1 + \sum_{\alpha \neq \alpha^*} e^{z_\alpha}} \geq 1 - \sum_{\alpha \neq \alpha^*} e^{z_\alpha} \geq 1 - (S-1)e^{-M}. \tag{A.27}$$

Thus, for $M > \frac{|\log \delta|}{2(S-1)}$, we obtain

$$\|x - x^*\| = 2(1 - x_{\alpha^*}) \leq 2(S-1)e^{-M} \leq \delta, \tag{A.28}$$

implying that U_M is contained in the ball $B_\delta = \{x : \|x - x^*\| \leq \delta\}$.

Finally, for our second claim, let $h(x) = \sum_{\alpha \in \mathcal{S}} x_\alpha \log x_\alpha$, $x \in \Delta(\mathcal{S})$, and let $h^*(y) = \max_{x \in \Delta}\{\langle y | x \rangle - h(x)\} = \log\left(\sum_{\alpha \in \mathcal{S}} e^{y_\alpha}\right)$ denote the convex conjugate of h [22, 25]. Then, a straightforward derivation yields

$$\frac{\partial h^*}{\partial y_\alpha} = \frac{\exp(y_\alpha)}{\sum_\beta \exp(y_\beta)} = \Lambda_\alpha(y), \tag{A.29}$$

so, by the basic properties of Legendre transformations [22], we get $\Lambda(y) = \arg\max\{\langle y | x \rangle - h(x)\}$. Therefore, taking $x = \Lambda(y)$, the KL divergence becomes

$$\begin{aligned} D_{\mathrm{KL}}(x^*, x) &= h(x^*) - h(x) - \langle \nabla h(x) | x - x^* \rangle \\ &= h(x^*) + \langle y | x \rangle - h(x) - \langle y | x \rangle - \langle \nabla h(x) | x - x^* \rangle \\ &= h(x^*) + h^*(y) - \langle y | x^* \rangle =: F(x^*, y), \end{aligned} \tag{A.30}$$

where $F(x^*, y)$ is the so-called Fenchel coupling [21], and we used the fact that $y = \nabla h(x)$.[10]

[10] Recall here that $x = \Lambda(y)$ is defined as the maximizer of the quantity $\langle y | x \rangle - h(x)$).

With this in mind, it suffices to show that

$$F(x^*, y') \leq F(x^*, y) + \langle y' - y | \Lambda(y) - x^* \rangle + \frac{1}{2}\|y' - y\|_\infty^2. \tag{A.31}$$

However, since h is 1-strongly convex with respect to the L^1-norm [25, p. 135], it follows that its convex conjugate h^* is 1-strongly smooth with respect to the L^∞-norm (the dual of the L^1-norm) [25, p. 148]. Specifically, this implies that

$$h^*(y') \leq h^*(y) + \langle y' - y | \nabla h^*(y) \rangle + \frac{1}{2}\|y' - y\|_\infty^2 \tag{A.32}$$

Eq. (A.31) then follows by writing out the definition of $F(x^*, y')$ and then using (A.32) and (A.29).

B Technical proofs

We begin with our proof of the local convergence properties of (HEDGE):

Proof of Theorem 2. Write $x^* = (\alpha_1^*, \ldots, \alpha_N^*)$ for the pure equilibrium under study, set $z_{i\alpha_i} = y_{i\alpha_i} - y_{i\alpha_i^*}$ and let $M > 0$ be such that $u_{i\alpha_i^*}(x) - u_{i\alpha_i}(x) \geq \mu$, for some μ, for all $\alpha_i \in \mathcal{S}_i \setminus \{\alpha_i^*\}$, $i \in \mathcal{N}$, whenever $x \in U_M$ (defined in Appendix A, Proposition A.2). Then, we have

$$z_{i\alpha_i}(t) = z_{i\alpha_i}(t-1) + \gamma_t \left[v_{i\alpha_i}(x(t)) - v_{i\alpha_i^*}(x(t)) \right] + \gamma_t \eta_{i\alpha_i}(t), \tag{B.1}$$

where $\eta_{i\alpha_i}(t) = \hat{v}_{i\alpha_i}(t) - v_{i\alpha_i}(x(t)) - (\hat{v}_{i\alpha_i^*}(t) - v_{i\alpha_i^*}(x(t)))$. Thus, assuming that (HEDGE) is initialized in U_{2M} and telescoping, we get

$$z_{i\alpha_i}(t) \leq -2M + \sum_{t=1}^{T} \gamma_t \left[v_{i\alpha_i}(x(t)) - v_{i\alpha_i^*}(x(t)) \right] + \sum_{t=1}^{T} \gamma_t \eta_{i\alpha_i}(t). \tag{B.2}$$

We now claim that, if γ_t is chosen appropriately, we have

$$\mathbb{P}\Big(\sup_t \textstyle\sum_{t=1}^T \gamma_t \eta_{i\alpha_i}(t) \leq M \Big) \geq 1 - \varepsilon / (N(S_i - 1)), \tag{B.3}$$

where $S_i = |\mathcal{S}_i|$. Indeed, let $X_{i\alpha_i}(t) = \sum_{k=1}^t \gamma_k \eta_{i\alpha_i}(k)$ and let $E_i(t)$ denote the event $\sup_{1 \leq k \leq t} |X_{i\alpha_i}(k)| \geq M$. By Hypothesis (H1), $X_{i\alpha_i}(t)$ is a martingale so Doob's maximal inequality [11, Theorem 2.1] yields

$$\mathbb{P}(E_i(T)) \leq \frac{\mathbb{E}[X_{i\alpha_i}(T)^2]}{M^2} \leq \frac{(8\sigma^2 + 32D^2)\sum_{t=1}^T \gamma_t^2}{M^2}, \tag{B.4}$$

where $D = \max_{x \in \mathcal{X}} \|v(x)\|_\infty < \infty$ and we used the noise variance estimate (see H2).

$$\begin{aligned}\mathbb{E}[\eta_{i\alpha_i}^2(t)] &= \mathbb{E}[\mathbb{E}[\eta_{i\alpha_i}^2(t) \mid \mathcal{F}_{t-1}]] \\ &= \mathbb{E}[\mathbb{E}[\big(\hat{v}_{i\alpha_i}(t) - v_{i\alpha_i}(x(t)) - \big(\hat{v}_{i\alpha_i^*}(t) - v_{i\alpha_i^*}(x(t))\big)\big)^2 \mid \mathcal{F}_{t-1}]] \\ &\le 4\max\nolimits_{\alpha_i\in\mathcal{S}_i} \mathbb{E}[\mathbb{E}[(\hat{v}_{i\alpha_i}(t) - v_{i\alpha_i}(x(t)))^2 \mid \mathcal{F}_{t-1}]] \\ &\le 4\max\nolimits_{\alpha_i\in\mathcal{S}_i} \mathbb{E}[\mathbb{E}[(\xi_{i\alpha_i}(t) + v_{i\alpha_i}(\alpha(t)) - v_{i\alpha_i}(x(t)))^2 \mid \mathcal{F}_{t-1}]] \\ &\le 8\sigma^2 + 8\max\nolimits_{\alpha_i\in\mathcal{S}_i} \mathbb{E}[\mathbb{E}[(v_{i\alpha_i}(\alpha(t)) - v_{i\alpha_i}(x(t)))^2 \mid \mathcal{F}_{t-1}]] \\ &\le 8\sigma^2 + 32\max\nolimits_{\alpha_i\in\mathcal{S}_i} \mathbb{E}[\mathbb{E}[v_{i\alpha_i}(x(t))^2 \mid \mathcal{F}_{t-1}]] \\ &\le 8\sigma^2 + 32D^2,\end{aligned} \tag{B.5}$$

The third line is obtained using relation $(\ell - f)^2 \le 2(\ell^2 + f^2)$ with $\ell = (\hat{v}_{i\alpha_i}(t) - v_{i\alpha_i}(x(t)))$ and $f = \big(\hat{v}_{i\alpha_i^*}(t) - v_{i\alpha_i^*}(x(t))\big)$. We use the same argument for the five line. The last line is obtained by the fact that $\mathbb{E}[\eta_{i\alpha_i}(t)\eta_{i\alpha_i}(t')] = 0$ if $t \neq t'$. Since $E_i(t) \subseteq E_i(t-1) \subseteq \dots$, it follows that the event $E_i = \bigcap_{t=1}^{\infty} E_i(t)$ occurs with probability $\mathbb{P}(E_i) \le (8\sigma^2 + 32D^2)\Gamma_2/M^2$ where $\Gamma_2 = \sum_{t=1}^{\infty} \gamma_t^2 < \infty$. Thus, if γ_t is chosen so that $\Gamma_2 \le \varepsilon M^2/(N(S_i-1)(8\sigma^2 + 32D^2))$, we get $\mathbb{P}(X_{i\alpha_i}(t) \ge M \text{ for all } t) \le \varepsilon/(N(S_i-1))$.

Assume therefore that $\Gamma_2 \le \varepsilon M^2/(N(S_i-1)(8\sigma^2+32D^2))$. Then, we obtain

$$\mathbb{P}(\max\nolimits_{i\in\mathcal{N},\alpha_i\in\mathcal{S}_i} \sup\nolimits_t X_{i\alpha_i}(t) \ge M) \le \sum\nolimits_{i\in\mathcal{N}} \sum\nolimits_{\alpha_i\neq\alpha_i^*} \tfrac{\varepsilon}{N(S_i-1)} \le \varepsilon. \tag{B.6}$$

Hence, going back to (B.2), a straightforward induction shows that $x(t) \in U_M$ for all t with probability at least $1-\varepsilon$. When this occurs, we also have

$$\mathbb{P}(z_{i\alpha_i}(T) \le -M - \mu \sum\nolimits_{t=1}^{T} \gamma_t \text{ for all } n) \le 1 - \varepsilon. \tag{B.7}$$

Thus, for all $c \in (0, \mu)$ and all sufficiently large t, the definition of Λ yields

$$\begin{aligned}x_{i\alpha_i^*}(T) &= \frac{\exp(y_{i\alpha_i^*}(T-1))}{\sum_{\alpha_i\in\mathcal{S}_i} \exp(y_{i\alpha_i}(T-1))} \\ &= \frac{1}{1 + \sum_{\alpha_i\neq\alpha_i^*} \exp(z_{i\alpha_i}(T-1))} \\ &\ge 1 - \sum_{\alpha_i\neq\alpha_i^*} \exp(z_{i\alpha_i}(T-1)) \ge 1 - \sum_{\alpha_i\neq\alpha_i^*} e^{-M} e^{-c\sum_{t=1}^{T-1}\gamma_t},\end{aligned} \tag{B.8}$$

with probability at least $1-\varepsilon$. Therefore, since $\|x_i - x_i^*\| = 1 - x_{i\alpha_i^*} + \sum_{\alpha_i\neq\alpha_i^*} x_{i\alpha_i} = 2(1 - x_{i\alpha_i^*})$, rearranging (B.8) yields

$$\mathbb{P}\Big(\|x(T) - x^*\| \le 2\sum_{i\in\mathcal{N}} \sum_{\alpha_i\neq\alpha_i^*} e^{-M} e^{-c\sum_{t=1}^{T-1}\gamma_t}\Big) \ge 1 - \varepsilon, \tag{B.9}$$

and our assertion follows. □

We now turn to the proof of the global convergence properties of (HEDGE):

Proof of Theorem 3. Since $\sum_{t=1}^{\infty}\gamma_t = \infty$, we have $\lim_{T\to\infty} e^{-c\sum_{t=1}^{T-1}\gamma_t} = 0$, and it clearly suffices to prove (18). Thus, given that $x^* = (\alpha_1^*, \dots, \alpha_N^*)$ is pure, an easy calculation yields

$$\begin{aligned} D_{\mathrm{KL}}(x^*, x) &= -\sum_{i\in\mathcal{N}} \log x_{i\alpha_i^*} = -\sum_{i\in\mathcal{N}} \log(1-(1-x_{i\alpha_i^*})) \\ &= -\sum_{i\in\mathcal{N}} \log(1-\|x_i - x_i^*\|/2) \geq \frac{1}{2}\|x-x^*\|, \end{aligned} \tag{B.10}$$

so it suffices to show that $D_{\mathrm{KL}}(x^*, x(t)) \to 0$. With this in mind, let $F_t = F(x^*, y(t)) = D_{\mathrm{KL}}(x^*, x(t+1))$. Then, with $x(t) = \Lambda(y(t-1))$, Proposition A.2 in Appendix A yields

$$F_t \leq F_{t-1} + \gamma_t\langle v(x(t))|x(t)-x^*\rangle + \gamma_t\psi_t + \frac{1}{2}\gamma_t^2\|\hat{v}(t)\|_\infty^2, \tag{B.11}$$

where we have set $\psi_t = \langle \xi(t) + v(\alpha(t)) - v(x(t))|x(t)-x^*\rangle$. Using this bound, we will show that $x(t)$ visits any neighborhood U of x^* infinitely many times.

Indeed, assume on the contrary that this is not so. It exists $\delta > 0$ such that $\|x(t) - x^*\| > \delta$ for all sufficiently large t. Then, by Proposition 1, there exists some $\delta > 0$ such that $\langle v(x(t))|x(t)-x^*\rangle \leq -\mu\delta$ for all sufficiently large t. Hence, telescoping (B.11) yields

$$\begin{aligned} F_T &\leq F_0 - \mu\delta\sum_{t=1}^{T}\gamma_t + \sum_{t=1}^{T}\gamma_t\psi_t + \frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\|\hat{v}(t)\|_\infty^2 \\ &\leq F_0 - \theta_t\left[\mu\delta - \frac{\sum_{t=1}^{T}\gamma_t\psi_t}{\theta_t} - \frac{\sum_{t=1}^{T}\gamma_t^2\|\hat{v}(t)\|_\infty^2}{2\theta_t}.\right] \end{aligned} \tag{B.12}$$

where $\theta_T = \sum_{t=1}^{T}\gamma_t$. Given that (by linearity)

$$\begin{aligned} \mathbb{E}[\psi_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}[\langle \xi(t) + v(\alpha(t)) - v(x(t))|x(t)-x^*\rangle \mid \mathcal{F}_{t-1}] \\ &= \langle \mathbb{E}[\xi(t)\mid\mathcal{F}_{t-1}]|x(t)-x^*\rangle + \langle \mathbb{E}[v(\alpha(t))\mid\mathcal{F}_{t-1}] - v(x(t))|x(t)-x^*\rangle \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[|\psi_t|^2] &\leq 2\,\mathbb{E}[\|\xi(t)\|_\infty^2 \mid \mathcal{F}_{t-1}] + 2\,\mathbb{E}[\|v(\alpha(t)) - v(x(t))\|_\infty^2 \mid \mathcal{F}_{t-1}] \\ &\leq 2\sigma^2 + 4\,\mathbb{E}[\|v(\alpha(t))\|_\infty^2 \mid \mathcal{F}_{t-1}] + 4\,\mathbb{E}[\|v(x(t))\|_\infty^2 \mid \mathcal{F}_{t-1}] \\ &\leq 2\sigma^2 + 8D^2, \end{aligned}$$

where $D = \max_{x\in\mathcal{X}}\|v(x)\|_\infty < \infty$ it follows that $R_T = \sum_{t=1}^{T}\gamma_t\psi_t$ is an L^2-bounded martinagle [11]. Hence, by the law of large numbers for martingale differences [11, Theorem 2.18], it follows that $\theta_T^{-1}\sum_{t=1}^{T}\gamma_t\psi_t \to 0$ (a.s.). Likewise, if we let $S_T = \sum_{t=1}^{T}\gamma_t^2\|\hat{v}(t)\|_\infty^2$, we get

$$\mathbb{E}[S_T] = \mathbb{E}[\mathbb{E}[S_T \mid \mathcal{F}_{T-1}]] \leq V^2\sum_{t=1}^{T}\gamma_t^2 \leq \varGamma_2 V^2, \tag{B.13}$$

where $\Gamma_2 = \sum_{t=1}^{\infty} \gamma_t^2$ (Recall that V is defined in 10b). Hence, by Doob's martingale convergence theorem [11, Theorem 2.5], S_T converges to some (random) finite value (a.s.). Combining the above, we conclude that the term in the brackets of (B.12) converges to $\mu\delta$ (a.s.). In turn, this implies that $F_T \to -\infty$, a contradiction as $F_t = F(x^*, y(t)) = D_{\mathrm{KL}}(x^*, x(t+1)) \geq 0$.

We have thus shown that $x(t)$ visits infinitely many times every neighborhood U of x^* – and hence, in particular, the neighborhood U_{2M} defined in the proof of Theorem 2. Since $x(t)$ remains in U_{2M} with positive probability, it follows that the probability that $x(t)$ exits U_{2M} infinitely many times is zero. We thus get $x(t) \in U_{2M}$ for all t greater than some random (but a.s. finite) T_0; hence, telescoping (B.1) from T_0 to T yields

$$z_{i\alpha_i}(T) \leq -2M - \theta_T \left[\mu - \theta_T^{-1} \sum\nolimits_{t=T_0}^{T} \gamma_t \eta_{i\alpha_i}(t) \right]. \tag{B.14}$$

As before, the law of large numbers [11, Theorem 2.18] shows that the term in the brackets of (B.14) converges to μ. Our claim then follows as in the proof of Theorem 2. □