# LINEAR LAST-ITERATE CONVERGENCE RATES FOR OPTIMISTIC LEARNING IN HARMONIC GAMES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We prove that the last iterate of Optimistic Follow-The-Regularized-Leader (OFTRL) converges at a *linear rate* to a Nash equilibrium in multiplayer harmonic games. Our result holds under a mild game-regularity condition and for a large family of regularizers, including entropy, Tsallis, and log-barrier. Our analysis is based on a dual perspective of OFTRL: we prove that the interaction between optimism and the global structure of harmonic games leads to the dissipation of an energy function along the dual iterates, for which we establish tight quantitative bounds. Our result also holds for the Extragradient FTRL variant.

## 1 INTRODUCTION

The last-iterate behavior of online learning algorithms in strategic multiagent settings has attracted growing attention in the study of learning in games. Unlike regret or time-average results, a last-iterate guarantee concerns the actual strategies played by agents over time, and the distinction is often consequential: while standard no-regret algorithms that are run for $T$ iterations may enjoy quantitative convergence guarantees for *time-averaged* play, the *$T$-th iterate itself* may fail to stabilize and can even be repelled from equilibrium and spiral out towards the boundary of the simplex (see, e.g., Bailey & Piliouras (2018); Daskalakis et al. (2018); Cheung & Piliouras (2019)).

Partly motivated by this gap, the past decade has seen substantial work on the design and analysis of *optimistic* online learning algorithms (Rakhlin & Sridharan, 2013). Broadly speaking, optimistic methods use recent feedback to predict future losses. Within this line of work, *Optimistic Follow-the-Regularized-Leader* (OFTRL) is a prominent family of methods which augments the classical FTRL framework with an optimistic prediction and naturally accommodates constrained, non-Euclidean geometries (Shalev-Shwartz, 2012). With entropic regularization, OFTRL recovers the popular *Optimistic Multiplicative Weights Update* (OMWU) algorithm. In the narrow, yet canonical setting of two-player zero-sum games, the last-iterate behavior of OMWU is now comparatively well understood: asymptotic convergence in the bimatrix case was shown by Daskalakis & Panageas (2019), subsequently strengthened to a *linear* rate by Wei et al. (2021) and extended to the extensive-form setting by Lee et al. (2021). More recent work has also given a finer geometric characterization of the last-iterate rates in the same bimatrix setting (Lazarsfeld et al., 2026).

Despite this detailed picture in two-player zero-sum games, the last-iterate behavior of OFTRL beyond zero-sum settings, and beyond just OMWU, remains much less understood. This is in contrast to the wealth of regret guarantees for OMWU in two-player zero-sum (Tsuchiya, 2026) and multiplayer general-sum games (Syrgkanis et al., 2015; Daskalakis et al., 2021; Anagnostides et al., 2022; Soleymani et al., 2025). This gap raises a broader question: which structural properties of general multiplayer games enable fast *last-iterate* convergence of FTRL-type dynamics to equilibrium?

Toward this direction, Legacci et al. (2024a) recently showed that some of the dynamical properties of OFTRL in two-player zero-sum games extend to the broader class of *multiplayer harmonic games*. These games capture strategic interactions between agents with conflicting interests, and they form one of the fundamental components in the decomposition of normal-form games into their potential and harmonic parts (Candogan et al., 2011; Abdou et al., 2022). While two-player zero-sum games with an interior equilibrium form a subclass, harmonic games need not be constant-sum, may admit non-convex equilibrium sets, and can induce nonlinear, non-monotone game operators. Nevertheless, Legacci et al. (2024a) proved in this broader setting that the last iterate of both OFTRL and the related *Extragradient FTRL* (EGFTRL) variant *converges asymptotically* to the game's Nash equilibrium set. However, establishing *convergence rates* remains open. We thus ask:

*How fast do optimistic FTRL-type dynamics converge in harmonic games, and what structural mechanisms govern their convergence?*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

### 1.1 Our Contributions and Techniques

The main contributions of our paper give quantitative answers to the above question.

**Last-iterate convergence rate for OFTRL, EGFTRL.** We prove that both OFTRL and EGFTRL *converge at a linear last-iterate rate to a Nash equilibrium in multiplayer harmonic games* (Theorem 4.1). Specifically, when instantiated with a suitable regularizer $Q$, and constant stepsize $\eta$, the iterates $\{w_t\}$ of the algorithms converge to an equilibrium $w^\star$ and satisfy

**Theorem 4.1 (Informal):** $\mathsf{D}_Q(w^\star, w_{t+1}) \leq 2\kappa \mathsf{D}_Q(w^\star, w_0) \cdot \exp(-\gamma\eta^2 \cdot t)$ for all $t \geq 0$ .

Here, $\mathsf{D}_Q$ is the Bregman divergence induced by $Q$, and $\gamma > 0$ and $\kappa > 0$ are instance-dependent constants. Note that convergence rates in $\mathsf{D}_Q$ also imply rates in the $\mathsf{Gap}$ function (Remark 4.2). This result establishes the first global, quantitative last-iterate guarantees for FTRL variants in harmonic games. Moreover, in contrast to prior work specific to two-player zero-sum games (Daskalakis & Panageas, 2019; Wei et al., 2021; Lazarsfeld et al., 2026), we do not assume existence of a *unique* Nash equilibrium, and we additionally give an exact characterization of the equilibrium selected by the dynamics as a function of the initialization, strengthening the asymptotic convergence result of Legacci et al. (2024a). Our result holds for a broad class of Legendre regularizers (including negative Shannon entropy, Tsallis entropies of order $q \in (0, 1)$, and the log-barrier function), and under a mild non-degeneracy condition on the game. This condition always holds for the linear polymatrix setting, as well as in other non-linear settings.

**Dual and variational perspective.** Our analysis is based on a dual perspective on OFTRL and EGFTRL. We prove that the convex conjugate of the regularizer plays the role of an *energy function* that dissipates along the dual iterates (Lemma 5.4). This energy dissipation corresponds to a contraction of the Bregman divergence from equilibrium along the primal iterates (Proposition 5.3). Dissipation stems from the action of optimism on the global structure of harmonic games. Structurally, the *loss operators* of these games (defined in (2)) satisfy a *weighted* Minty variational inequality (Minty, 1967) that holds globally with equality, a property shown by Legacci et al. (2024a) and restated in Proposition 2.3. In continuous time, due to this global harmonic structure, energy is conserved along the FTRL flow in a scaled coordinate system.

However, in discrete time, we show that an idealized *Implicit FTRL* (IFTRL) discretization of the FTRL flow *dissipates* energy. By coupling their iterates with IFTRL, we prove that dissipation persists under both OFTRL and EGFTRL with a sufficiently small constant stepsize. Our dual viewpoint and proof techniques are inspired from the recent work of Lazarsfeld et al. (2026) who studied OMWU in two-player zero-sum games. However, the special skew-symmetric structure of zero-sum game operators is not available in harmonic games, and we develop a different coupling analysis technique with the implicit method. Moreover, our analysis and convergence rates apply to *multiplayer* games with possibly *multiple* equilibria, and they hold for both Optimistic *and* Extragradient FTRL with regularizers beyond negative Shannon entropy.

**Quantitative separation for vanilla FTRL.** Our techniques additionally establish quantitative guarantees for vanilla (non-optimistic) FTRL in polymatrix harmonic games: here, we prove that the FTRL iterates *move away* from equilibrium at a geometric rate within any fixed Bregman divergence sublevel set (Theorem 4.5). This extends and makes quantitative the asymptotic result of Bailey & Piliouras (2018) that was proven for two-player zero-sum games. Taken together, our results give a unified quantitative picture of the last-iterate behavior of FTRL-type dynamics under the global harmonic structure. We illustrate our results further in Figure 1.

**Other related work.** We briefly mention several additional threads of related work here and defer a more complete discussion to Section A. Optimistic and extragradient methods have a long history in optimization.[1] Compared to the FTRL family, the last-iterate behavior of optimistic and extragradient *gradient-descent-ascent* methods is better understood (see, e.g., Mokhtari et al. (2020); Golowich et al. (2020a); Cai et al. (2022b); Gorbunov et al. (2022b)). In constrained settings, however, these Euclidean schemes are generally distinct from FTRL dynamics (see, e.g., Cai et al. (2024)). For FTRL and their closely related variants, the iterate-based guarantees most relevant to our setting are either asymptotic (Hsieh et al., 2021), concern only the *best iterate* (Anagnostides et al., 2022), or provide *local* convergence rates (Lei et al., 2021; Azizian et al., 2024). We provide global linear last-iterate rates in the non-Euclidean setting of harmonic games.

[1] Optimistic and extragradient methods date back at least to Korpelevich (1976), Popov (1980), Nemirovski (2004). Throughout, we colloquially say *optimism* to describe both types of variants. See Section A.

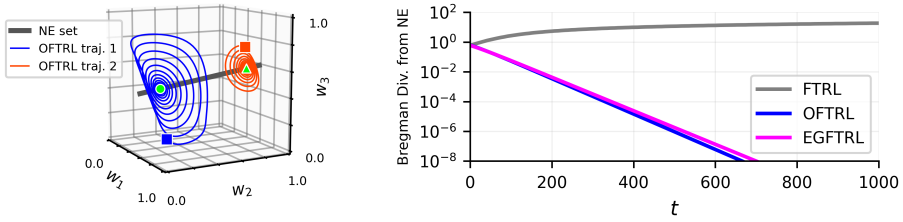


Figure 1: Variants of FTRL (with entropy regularizer) on the 3-player harmonic game in B.4.1. Left: two trajectories of OFTRL converging to their distinct analytic limits (in green) from Theorem 4.1 (i). Right: Bregman divergence from NE for the $t$-th iterate of OFTRL, EGFTRL, and FTRL from a fixed initialization. OFTRL and EGFTRL converge linearly (Theorem 4.1(ii)), while FTRL diverges at a geometric rate (Theorem 4.5).

## 2 Preliminaries and Setting

**Notation.** For any integer $k \geq 1$, let $[k] = \{1, \dots, k\}$. Let $\Delta_k \subset \mathbb{R}^k$ denote the probability simplex, and let $\mathrm{ri}(\Delta_k) = \{q \in \Delta_k : q(j) > 0, \forall j \in [k]\}$ denote its relative interior. We say that $q$ is *interior* when $q \in \mathrm{ri}(\Delta_k)$. Let $\mathbf{1}_k \in \mathbb{R}^k$ and $I_k \in \mathbb{R}^{k \times k}$ denote the $k$-dimensional all-ones vector and identity matrix, respectively. Let $\langle \cdot, \cdot \rangle$ be the standard $\ell_2$ inner product and $\|v\|_2 = \sqrt{\langle v, v \rangle}$. If $M$ is positive semi-definite, we write $M \succeq 0$. For a linear subspace $\mathcal{U} \subseteq \mathbb{R}^k$, let $\Pi_{\mathcal{U}} : \mathbb{R}^k \to \mathcal{U}$ denote the orthogonal projection onto $\mathcal{U}$. For $g : \mathcal{U} \to \mathcal{U}$ and $u \in \mathcal{U}$, let $J_g(u) : \mathcal{U} \to \mathcal{U}$ denote the Jacobian of $g$ at $u$ when it exists. For a convex set $\mathcal{C} \subseteq \mathbb{R}^k$ and a differentiable convex $f : \mathcal{C} \to \mathbb{R}$, the *Bregman divergence* $\mathsf{D}_f : \mathcal{C} \times \mathrm{ri}(\mathcal{C}) \to \mathbb{R}$ is given by $\mathsf{D}_f(y', y) = f(y') - f(y) - \langle \nabla f(y), y' - y \rangle$.

### 2.1 Multiplayer Normal-Form Games

Finite normal-form games consist of $N \geq 2$ players. Each player $i \in [N]$ has a set $\mathcal{A}_i := \{\alpha_j\}_{j \in [d_i]}$ of $d_i$ actions, and under a joint action profile $a = (a_1, \dots, a_N) \in \mathcal{A} := \prod_{i \in [N]} \mathcal{A}_i$, player $i$ incurs a loss $\ell_i(a)$, where $\ell_i : \mathcal{A} \to \mathbb{R}$ is the player's loss function. Each player $i$ can choose a mixed strategy $w_i \in \mathcal{W}_i := \Delta_{d_i}$ over $\mathcal{A}_i$. Using the joint space $\mathcal{W} := \prod_{i \in [N]} \mathcal{W}_i$, we overload notation and let $\ell_i : \mathcal{W} \to \mathbb{R}$ denote player $i$'s expected loss defined as $\ell_i(w) = \mathbf{E}_{a \sim w}[\ell_i(a)]$ for any joint strategy profile $w = (w_1, \dots, w_N) \in \mathcal{W}$. We set $D := d_1 + \cdots + d_N$ and use the standard notation $\mathcal{A}_{-i} := \prod_{j \neq i \in [N]} \mathcal{A}_j$ and $\mathcal{W}_{-i} := \prod_{j \neq i \in [N]} \Delta_{d_j}$.

**Nash equilibrium.** Our primary solution concept will be the celebrated Nash equilibrium: For $\epsilon \geq 0$, $w^\star \in \mathcal{W}$ is an $\epsilon$-approximate Nash equilibrium ($\epsilon$-NE) if $\max_{i \in [N]} \max_{w_i \in \mathcal{W}_i} \{\ell_i(w^\star) - \ell_i(w_i, w^\star_{-i})\} \leq \epsilon$. Due to Nash (1951), every finite normal-form game has at least one ($\epsilon = 0$) NE.

### 2.2 Harmonic Games

We focus on the class of multiplayer *harmonic games* defined as follows.

**Definition 2.1** (Harmonic game)**.** A finite normal-form game is *harmonic* if for each player $i \in [N]$, there exists a positive measure $\mu_i = (\mu_i(1), \dots, \mu_i(d_i)) \in (0, \infty)^{d_i}$ such that

$$\sum_{i \in [N]} \sum_{j \in [d_i]} \mu_i(j) \cdot \big(\ell_i(a) - \ell_i(\alpha_j, a_{-i})\big) = 0 \quad \text{for all } a = (a_i, a_{-i}) \in \mathcal{A}\,. \tag{1}$$

Definition 2.1 is the *weighted* definition of harmonic games (introduced by Abdou et al. (2022) and considered in Legacci et al. (2024a)), extending the unweighted definition of Candogan et al. (2011). Intuitively, (1) imposes a global balance condition: at every action profile, the $\mu_i$-weighted sum of the loss differences induced by all unilateral deviations is zero.

**Remark 2.2** (Examples)**.** Harmonic games (Definition 2.1) are in general multiplayer, non-constant-sum, and can have non-convex sets of Nash equilibria. We provide such families of examples in Section B.4. An important subclass is two-player zero-sum games with at least one interior NE (but not every zero-sum game is harmonic); See Proposition B.7.

### 2.3 Variational Characterization of Harmonic Games

Our analysis will rely on the variational formulation of finite games through their associated game operator (Facchinei & Pang, 2003). In this language, the harmonicity condition of Definition 2.1 admits an equivalent characterization that will be central to the analysis that follows. We introduce the relevant variational notation and discuss this characterization below.

**Joint loss operator.** For a finite normal-form game, we define its joint loss operator $G : \mathcal{W} \to \mathbb{R}^D$. For any $w \in \mathcal{W}$, the $i$-th block-coordinate $G_i(w)$ of $G(w)$ is given by $g_i : \mathcal{W}_{-i} \to \mathbb{R}^{d_i}$, where

$g_i(w_{-i}) = \left(\mathbf{E}_{a_{-i}\sim w_{-i}}[\ell_i(\alpha_j, a_{-i})]\right)_{j\in[d_i]} \in \mathbb{R}^{d_i}$. In particular, as $\ell_i(w) = \langle w_i, g_i(w_{-i})\rangle$, observe that each $g_i(w_{-i}) = \nabla_{w_i}\ell_i(w)$ is exactly the gradient of the $i$-th player's loss with respect to their own action. The map $G$ is then the concatenation of these individual loss gradients, i.e.

$$G(w) := (\nabla_{w_1}\ell_1(w), \ldots, \nabla_{w_N}\ell_N(w)) \in \mathbb{R}^D \text{ for all } w \in \mathcal{W} . \tag{2}$$

We associate each game with its operator $G$ and make standard regularity assumptions. For all $i \in [N]$, let each $\mathcal{W}_i$ be equipped with norm $\|\cdot\|_i$ and dual norm $\|\cdot\|_{i,*}$. Then we define the global product norms $\|w\| = (\sum_{i=1}^N \|w_i\|_i^2)^{1/2}$ for $w \in \mathcal{W}$ and $\|z\|_* = (\sum_{i=1}^N \|z_i\|_{i,*}^2)^{1/2}$ for $z \in \mathbb{R}^D$. We then assume that $G$ is $L$-Lipschitz and $B$-bounded with respect to these norms.

**Assumption 1** (Lipschitz and bounded operator)**.** There exist constants $L \geq 1$ and $B \geq 1$ such that (i) $\|G(w') - G(w)\|_* \leq L\|w' - w\|$ and (ii) $\|G(w)\|_* \leq B$ for all $w, w' \in \mathcal{W}$.

Assumption 1 is always satisfied when the losses $\ell_i$ are smooth, with the global constants $L$ and $B$ inherited from the corresponding regularity bounds on the losses.

**Variational characterization of NE.** The operator $G$ also provides a standard variational characterization of Nash equilibrium: $w^\star \in \mathcal{W}$ is a Nash equilibrium (for $\epsilon = 0$) if and only if

$$\langle G(w^\star), w - w^\star\rangle \geq 0 \text{ for all } w \in \mathcal{W} . \tag{3}$$

We will thus denote by $\mathrm{NE}(G) \subseteq \mathcal{W}$ the set of all $w^\star$ satisfying (3). Moreover, for any point $w \in \mathcal{W}$, we can measure its distance to a Nash equilibrium via the *gap function*: for any $w \in \mathcal{W}$, let

$$\mathsf{Gap}(w) := \max_{\widetilde{w}\in\mathcal{W}} \langle G(w), w - \widetilde{w}\rangle . \tag{4}$$

If $\mathsf{Gap}(w') \leq \epsilon$, then $w'$ is an $\epsilon$-NE.

**Variational characterization of harmonic games.** As shown by Legacci et al. (2024a), harmonic games from Definition 2.1 also admit the following variational characterization in terms of $G$:

**Proposition 2.3** (Legacci et al. (2024a), Proposition A.4)**.** *A game with loss operator $G : \mathcal{W} \to \mathbb{R}^D$ is harmonic if and only if there exist (i) a matrix $M = \mathrm{Diag}(m_1 I_{d_1}, \ldots, m_N I_{d_N})$ where $m_i > 0$ for all $i \in [N]$ and (ii) an interior point $w' \in \mathrm{ri}(\mathcal{W})$ such that:*

$$\langle MG(w), w - w'\rangle = 0 \text{ for all } w \in \mathcal{W} . \tag{5}$$

Intuitively, (5) says that, relative to a center $w'$, the $M$-weighted operator is orthogonal to all displacements $w - w'$, expressing in variational terms the weighted cancellation of unilateral deviations in (1). We give more background in Section B.2, and we highlight here several additional points:

- *Equilibrium condition*: Any $w' \in \mathcal{W}$ satisfying the proposition is an NE (see Lemma B.5), and thus (5) is simultaneously a structural property of $G$ as well as an equilibrium condition.
- *Weight matrix*: We call $M$ from the proposition a *weights matrix*, and note that each $m_i = \|\mu_i\|_1$ for $\mu_i$ from Definition 2.1. For fixed $M$, we let $m_{\max} = \max_{i\in[N]} m_i$ and $m_{\min} = \min_{i\in[N]} m_i$.
- *Operator properties:* Operators satisfying (5) are generally non-monotone and nonlinear, even for multiplayer normal-form games. The case of linear $G$ (meaning $G(w) = Pw$) corresponds to *polymatrix games* with the harmonic property; see Section B.4.

**Weighted Minty equality.** Note that (3) is *Stampacchia's variational inequality* while (5) is a weighted Minty variational *equality* for the operator $G$ (see Section B.1 for background). For this reason, it is useful to define the set of *all solutions* of (5), which we dub *weighted Minty equality* points. For a fixed weights matrix $M$, let

$$\mathrm{WMEQ}_M(G) = \{w' \in \mathcal{W} : \langle MG(w), w - w'\rangle = 0 \text{ for all } w \in \mathcal{W}\} . \tag{6}$$

By Proposition 2.3, a game is harmonic iff $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weights matrix $M$. Due to Lemma B.5, we have $\mathrm{WMEQ}_M(G) \subseteq \mathrm{NE}(G)$, but in general $\mathrm{NE}(G) \neq \mathrm{WMEQ}_M(G)$. Moreover, we note the generality of this variational perspective in terms of the underlying game.

**Remark 2.4** (Generality of variational perspective)**.** Our main results only use the characterization of Proposition 2.3 under the conditions on $G$ from Assumption 1. More generally, our results extend beyond normal-form harmonic games to the setting where the individual loss functions $\ell_i : \mathcal{W} \to \mathbb{R}$ are each convex in their $i$'th coordinate $w_i$, where $G$ is the resulting pseudo-gradient map in (2), and where $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weight matrix $M$. By Proposition 2.3, normal-form games satisfying Definition 2.1 are one such instantiation.

# 3 A GLOBAL VIEW OF OPTIMISTIC AND EXTRAGRADIENT FTRL

We now introduce the global dynamics of the family of OFTRL and EGFTRL algorithms where each player $i \in [N]$ uses a strictly convex regularizer $Q_i : \mathcal{W}_i \to \mathbb{R}$ and a fixed stepsize $\eta_i > 0$. We can then define the induced *global regularizer* $Q : \mathcal{W} \to \mathbb{R}$ and *global stepsize* $\eta > 0$, which together reduce the simultaneous playerwise updates into a single primal-dual system. We develop this global view here and defer the full details on the playerwise updates to Section C.1.

**Global stepsize and regularizer.** Let $\eta := \max_{i\in[N]} \eta_i > 0$, and let $Q : \mathcal{W} \to \mathbb{R}$ be the separable regularizer that absorbs the player-specific stepsizes via

$$Q(w) = \sum_{i=1}^{N} \frac{\eta}{\eta_i} Q_i(w_i) \quad \text{for all } w = (w_1, \dots, w_N) \in \mathcal{W} \,. \tag{7}$$

**Global primal and dual updates.** Using $\eta$ and $Q$ from (7), the global primal update for FTRL-type algorithms admits a similar structure to the standard playerwise definitions. Formally, fix initial $w_0 \in \mathrm{ri}(\mathcal{W})$. Then for $t \geq 1$, let $w_t := (w_{t,i})_{i\in[N]} \in \mathcal{W}$ be the *global primal iterate* given by

$$w_t = \mathrm{argmin}_{w\in\mathcal{W}} \{-\langle w, z_t\rangle + Q(w)\} \quad \text{for all } t \geq 1. \tag{8}$$

Here, $z_t = (z_{t,i})_{i\in[N]} \in \mathbb{R}^D$ is a variant-specific *global dual iterate* that accumulates operator vectors scaled by $\eta$. For example, the vanilla FTRL variant sets $z_t := z_0 - \eta \sum_{k=0}^{t-1} G(w_k)$ for all $t \geq 1$ and initial $z_0 \in \mathbb{R}^D$. More generally, for initial $z_0 \in \mathbb{R}^D$ and all $t \geq 0$, the dual sequences $\{z_t\}$ for the vanilla (FTRL), implicit (IFTRL), optimistic (OFTRL), and extragradient (EGFTRL) variants can each be written recursively via the update rules:

$$\begin{aligned}
z_{t+1} &= z_t - \eta\, G(w_t) && \text{(FTRL)}\\
z_{t+1} &= z_t - \eta\, G(w_{t+1}) && \text{(IFTRL)}\\
z_{t+1} &= z_t - 2\eta\, G(w_t) + \eta\, G(w_{t-1}) && \text{(OFTRL)}\\
z_{t+1/2} &= z_t - \eta\, G(w_t)\,, \quad z_{t+1} = z_t - \eta\, G(w_{t+1/2})\,. && \text{(EGFTRL)}
\end{aligned}$$

Here, we use for (OFTRL) the notational convention $w_{-1} := w_0 \in \mathrm{ri}(\mathcal{W})$, and for (EGFTRL) the convention that $w_{t+1/2}$ is the minimizer of the objective in (8) with respect to $z_{t+1/2}$.

**Remark 3.1** (Intuition for optimism)**.** Intuitively, both OFTRL and EGFTRL augment the vanilla FTRL update in an attempt to *predict* the next vector $G(w_{t+1})$. Thus, both methods can be viewed as approximations of the idealized (unimplementable) IFTRL update (see also Section 5.1).

## 3.1 ASSUMPTIONS ON REGULARIZER

We study instantiations of FTRL variants with regularizers $Q$ satisfying the following property:

**Assumption 2** (Relatively Legendre)**.** $Q$ is (i) strictly convex and twice continuously differentiable on $\mathrm{ri}(\mathcal{W})$, and (ii) $\|\nabla Q(w)\| \to \infty$ as $w$ approaches the relative boundary of $\mathcal{W}$.

Let $H := Q^* : \mathbb{R}^D \to \mathbb{R}$ denote the *convex conjugate* of $Q$, which we refer to as an *energy function* following prior work (Wibisono et al., 2022). Under Assumption 2, $H$ is continuously differentiable and admits the following key primal-dual relationship for FTRL variants (see Section C.2):

**Proposition 3.2** (Primal-dual invariant)**.** *Let $\{w_t\}$ be the iterates from (8) using the dual sequence $\{z_t\}$ of (FTRL), (IFTRL), (OFTRL), or (EGFTRL). Then $w_t = \nabla H(z_t) \in \mathrm{ri}(\mathcal{W})$ for all $t \geq 1$.*

**Remark 3.3** (OMWU)**.** The (componentwise) negative Shannon entropy is the canonical $Q$ satisfying Assumption 2. Here, the energy $H = Q^*$ is the *log-sum-exp* function, and $\nabla H(z) : \mathbb{R}^D \to \mathrm{ri}(\mathcal{W})$ is the *softmax map*. This yields the MWU algorithm and its optimistic variants. See Section C.3.

Our main convergence results also require two additional assumptions on $Q$ and $H$.

**Assumption 3** (1-strong-convexity)**.** $\forall w \in \mathrm{ri}(\mathcal{W}), w' \in \mathcal{W}$, $\mathsf{D}_Q(w', w) \geq \frac{1}{2}\|w' - w\|^2$.

**Assumption 4** (Local Hessian stability)**.** $H$ is twice continuously differentiable, and $\exists\, \omega \geq 1$ s.t.

$$\forall z, z' \in \mathbb{R}^D : \quad \exp(-\omega\|z - z'\|_*) \cdot \nabla^2 H(z') \preceq \nabla^2 H(z) \preceq \exp(\omega\|z - z'\|_*) \cdot \nabla^2 H(z') \,. \tag{LHS}$$

Assumption 3 is standard and readily generalizes to $K$-strong-convexity for $K > 0$. Assumption 4 specifies a local notion of third-order smoothness related to *generalized self-concordance* (Sun & Tran-Dinh, 2019), which intuitively controls the variation of the local dual geometry induced by $H$. Due to the separability of $Q$, Assumptions 2, 3, 4 all hold whenever each $Q_i$ and $H_i := Q_i^*$ satisfy the componentwise properties, including for the following canonical choices:

**Remark 3.4** (Examples)**.** Regularizers satisfying Assumptions 2, 3, and 4 include negative Shannon entropy, Tsallis entropy of order $q \in (0,1)$, and the log-barrier regularizer; see Section C.3.

### 3.2 Geometry of Primal and Dual Iterates

Our analysis of FTRL variants also relies on several geometric properties that we introduce here:

**Dual update space and reachable primal set.** Each dual iterate of the FTRL variants satisfies $z_t - z_0 \in \mathcal{Z} := \mathrm{Span}(G(\mathcal{W}))$, and thus we call the linear subspace $\mathcal{Z} \subseteq \mathbb{R}^D$ the *dual update space*. Similarly, for initializations with $w_0 = \nabla H(z_0)$, each $w_t \in \mathcal{R}(w_0) := \{\nabla H(z_0 + z) : z \in \mathcal{Z}\}$, and thus we refer to $\mathcal{R}(w_0) \subseteq \mathrm{ri}(\mathcal{W})$ as the *reachable primal set*.

**Effective dual space and stationarity map.** Importantly, not every dual direction in $\mathcal{Z}$ affects updates to the primal point under $\nabla H$. To make this precise, define $\mathcal{S} \subseteq \mathbb{R}^D$ and $\mathcal{S}^\perp \subseteq \mathbb{R}^D$ by

$$\mathcal{S} := \prod_{i=1}^N \mathrm{Span}(\mathbf{1}_{d_i}) \quad \text{and} \quad \mathcal{S}^\perp := \{v \in \mathbb{R}^D : \langle v, s\rangle = 0\,, \forall s \in \mathcal{S}\}\,. \tag{9}$$

The linear subspace $\mathcal{S}$ and its orthogonal complement $\mathcal{S}^\perp$ are the *normal space* and *tangent space* of the affine hull of $\mathcal{W}$, respectively (see Section B.1). Under Assumption 2, a key invariant is that $\nabla H(z_0 + z) = \nabla H(z_0 + \Pi_{\mathcal{S}^\perp} z)$ for $z \in \mathcal{Z}$ (Proposition C.4). Thus, we let $\mathcal{U} := \Pi_{\mathcal{S}^\perp}\mathcal{Z}$ denote the *effective dual space*, and for $z_0 \in \mathbb{R}^D$ and $w_0 = \nabla H(z_0)$, we define the *effective stationarity map*

$$\pi_0 : \mathcal{U} \to \mathcal{U}\,, \quad \pi_0(u) = \Pi_{\mathcal{S}^\perp} G(\nabla H(z_0 + u)) \ \text{ for all } u \in \mathcal{U}\,. \tag{10}$$

Importantly, $\pi_0(u) = 0$ if and only if $\nabla H(z_0 + u) \in \mathrm{NE}(G)$ (Proposition E.1), and thus $\|\pi_0(u)\|$ acts as a first-order measure of distance to equilibrium in effective dual coordinates. We will also assume the following regularity condition on the zeros of $\pi_0$:

**Assumption 5** (Regularity)**.** For initial $z_0 \in \mathbb{R}^D$ and $w_0 = \nabla H(z_0)$, assume $\mathcal{R}(w_0) \not\subseteq \mathrm{NE}(G)$, and that every zero of $\pi_0$ is *regular*, i.e., if $\pi_0(u) = 0$ for $u \in \mathcal{U}$, then the Jacobian $J_{\pi_0}(u)$ is nonsingular.

**Remark 3.5.** Assumption 5 is a first-order nondegeneracy condition whose main consequence is a local linear error bound around each reachable Nash equilibrium. The residual $\|\pi_0(u)\|$ controls the distance to that equilibrium in the effective dual space. For linear $G$ (polymatrix games) the assumption always holds (Proposition E.6), and it also holds for nonlinear examples (see Section B.4.2).

## 4 Main Result: Linear Convergence Rates in Harmonic Games

Our main theorem proves that, for harmonic games, OFTRL and EGFTRL converge at a linear last-iterate rate to an NE $w^\star \in \mathrm{WMEQ}_M(G)$ in Bregman divergence $\mathsf{D}_Q(w^\star, \cdot)$. To state the result, we define the *$M$-weighted regularizer* given by $Q_M(w) = \sum_{i=1}^N m_i \cdot (\frac{\eta}{\eta_i} Q_i(w_i))$.

**Theorem 4.1** (Linear last-iterate convergence rate)**.** *Let $G : \mathcal{W} \to \mathbb{R}^D$ be a loss operator such that $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weight matrix $M$, and let $\kappa_M = \frac{m_{\max}}{m_{\min}}$. Let Assumptions 1, 2, 3, 4, and 5 hold. Then for (OFTRL) or (EGFTRL) with initialization $w_0 \in \mathrm{ri}(\mathcal{W})$ and with stepsize $\eta \in (0, 1/(880LB\omega\kappa_M)]$, there exists a constant $\gamma(w_0) > 0$ such that the following hold:*

*(i) There exists a unique limit $w_\infty := \lim_{t\to\infty} w_t = \mathrm{argmin}_{w^\star \in \mathrm{WMEQ}_M(G)} \mathsf{D}_{Q_M}(w^\star, w_0)$.*

*(ii) For all $t \geq 0$: $\mathsf{D}_Q(w_\infty, w_{t+1}) \leq 2\kappa_M \cdot \mathsf{D}_Q(w_\infty, w_0) \cdot \exp(-\gamma(w_0)\eta^2 \cdot t)$.*

Theorem 4.1 also gives a linear convergence rate in $\mathsf{Gap}(w_t)$ due to the following relationships:

**Remark 4.2** (Gap vs. Bregman divergence)**.** Let $w^\star \in \mathrm{WMEQ}_M(G)$. Then for all $w \in \mathrm{ri}(\mathcal{W})$:

$$\tfrac{1}{2(L\rho)^2} \cdot \mathsf{Gap}(w)^2 \ \leq\ \tfrac{1}{2}\|w^\star - w\|^2 \ \leq\ \mathsf{D}_Q(w^\star, w)\,, \tag{11}$$

where $\rho := \max_{w', w'' \in \mathcal{W}} \|w' - w''\|$ is the diameter of $\mathcal{W}$. Thus, convergence in $\mathsf{D}_Q(w^\star, \cdot)$ implies convergence in $\mathsf{Gap}(\cdot)$. The full derivation of these relationships is in Proposition B.8.

Before giving an overview of the proof in Section 5, we make several additional remarks:

**Remark 4.3** (Equilibrium selection)**.** Part (i) of Theorem 4.1 proves that the limit points of OFTRL and EGFTRL are in $\mathrm{WMEQ}_M(G) \subseteq \mathrm{NE}(G)$, along with an exact initialization-dependent characterization in terms of the $M$-weighted regularizer $Q_M$. This strengthens the result of Legacci et al. (2024b), who showed only asymptotic convergence to the larger set $\mathrm{NE}(G)$. Note that the theorem does not assume uniqueness of NE, in contrast to prior last-iterate results for OMWU in two-player zero-sum games (Daskalakis & Panageas, 2019; Wei et al., 2021; Lazarsfeld et al., 2026).

**Remark 4.4** (Instance-dependent rate)**.** The constant $\gamma(w_0) > 0$ in part (ii) of Theorem 4.1 depends on $G$, on properties of the regularizer $Q$ and energy function $H = Q^*$, and on the initial $w_0 \in \mathrm{ri}(\mathcal{W})$. As in prior results for two-player zero-sum games (Wei et al., 2021; Lazarsfeld et al., 2026), the constant $\gamma(w_0)$ is thus instance-dependent, but uniform over all steps $t \geq 1$.

**High-level proof strategy.** The proof has two main components: (i) proving *dissipative* bounds on the change in energy along the dual iterates, and (ii), proving a structural *operator domination* inequality that relates the magnitude of dissipation back to primal Bregman divergence from equilibrium. This high-level strategy is similar to Lazarsfeld et al. (2026) for the special case of OMWU in two-player zero-sum games, with a unique, interior NE, but our result is substantially more general. We do not leverage the special skew-symmetric structure of zero-sum game operators, do not assume uniqueness of NE, and our analysis simultaneously holds for both OFTRL *and* EGFTRL for a broad class of regularizers beyond negative Shannon entropy.

**Implications of analysis for vanilla FTRL.** Our techniques for proving Theorem 4.1 simultaneously establish a geometric rate of *divergence* away from equilibrium under *vanilla* FTRL. Specifically, we prove the following result for linear operators $G$ (which capture polymatrix games):

**Theorem 4.5** (Divergence of FTRL)**.** *Let $G : \mathcal{W} \to \mathbb{R}^D$ be a linear loss operator such that $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weight matrix $M$, and let $\kappa_M = \frac{m_{\max}}{m_{\min}}$. Consider (FTRL) with $\eta \in (0, 1/(10B\omega\kappa_M)]$ and $w_0 \notin \mathrm{NE}(G)$. Then for $w^\star \in \mathrm{NE}(G) \cap \mathrm{ri}(\mathcal{W})$ and $\alpha > \mathsf{D}_{Q_M}(w^\star, w_0)$, letting $T_\alpha(w^\star) = \sup\{t \geq 0 : \mathsf{D}_{Q_M}(w^\star, w_t) \leq \alpha\}$, there exists $\beta(w_0, w^\star, \alpha) > 0$ such that*

$$\mathsf{D}_Q(w^\star, w_t) \geq \kappa_M^{-1} \cdot \mathsf{D}_Q(w^\star, w_0) \cdot (1 + \beta(w_0, w^\star, \alpha) \cdot \eta^2)^t \ \text{ for all } 0 \leq t \leq T_\alpha(w^\star) \ .$$

The proof is in Section L and is similar to that of Theorem 4.1. Moreover, we note the following:

**Remark 4.6** (Quantitative separation)**.** Theorem 4.5 shows a separation between (FTRL) and the (OFTRL) and (EGFTRL) variants. In the case of linear $G$ (which captures polymatrix games), the last-iterate of FTRL moves away from all interior Nash equilibria and toward the simplex boundary, and within any fixed Bregman sublevel set with respect to a fixed interior NE, it does so at a geometric rate. This is in contrast to last-iterate convergence rates to NE for OFTRL and EGFTRL. The theorem generalizes and makes quantitative the divergence result of Bailey & Piliouras (2018) beyond the case of OMWU in two-player zero-sum games.

## 5 Overview of Proof of Theorem 4.1

In this section, we give a technical overview of the proof of Theorem 4.1 (the full proof is in Section K). The starting point of our analysis is to introduce a change of basis that allows for studying OFTRL and EGFTRL over scaled spaces, and using a *symmetrized* operator $V$.

**Reduction to unweighted system.** Fix $G : \mathcal{W} \to \mathbb{R}^D$ with $\mathrm{WMEQ}_M(G) \neq \emptyset$ for a weight matrix $M$. Define the scaled primal domain $\mathcal{X} := M^{1/2}\mathcal{W} \subseteq \mathbb{R}^D$, and for each $w \in \mathcal{W}$ and $z \in \mathbb{R}^D$, let $x = M^{1/2}w$ and $y = M^{1/2}z$. Then let $V : \mathcal{X} \to \mathbb{R}^D$ be the $M$-symmetrized operator, where

$$V(x) = M^{1/2}G(M^{-1/2}x) \ \text{ for all } x \in \mathcal{X} \ . \tag{12}$$

The key property of $V$ is that all solutions of the *weighted* Minty variational equality of (5) under $G$ also satisfy the following *unweighted* Minty variational equality under $V$:

**Proposition 5.1** (Unweighted Minty equality)**.** *Let $V : \mathcal{X} \to \mathbb{R}^D$ be the operator from (12). For any solution $w^\star \in \mathrm{WMEQ}_M(G)$, let $x^\star := M^{1/2}w^\star \in \mathcal{X}$. Then it holds that:*

$$\langle V(x), x - x^\star \rangle = 0 \ \text{ for all } x \in \mathcal{X} \ . \tag{MEQ}$$

We write $\mathrm{MEQ}(V) \subseteq \mathcal{X}$ to denote the set of solutions $x^\star$ satisfying (MEQ). As a straightforward corollary: $w^\star \in \mathrm{WMEQ}_M(G)$ if and only if $x^\star = M^{1/2}w^\star \in \mathrm{MEQ}(V)$ (see Section D). From an algorithmic standpoint, the change of basis also yields the following equivalence:

**Proposition 5.2** (Symmetrized dynamics)**.** *Let $V : \mathcal{X} \to \mathbb{R}^D$ be the operator from (12). Then there exist a global regularizer $R : \mathcal{X} \to \mathbb{R}$ and energy function $F = R^*$ such that, if $\{(w_t, z_t)\}$ is an (OFTRL) or (EGFTRL) trajectory for $(G, H, \eta)$, then letting $x_t = M^{1/2}w_t$ and $y_t = M^{1/2}z_t$ for all $t \geq 0$, it follows that $\{(x_t, y_t)\}$ is an (OFTRL) or (EGFTRL) trajectory for $(V, F, \eta)$.*

See Section D for the proof and the exact definitions of $R$ and $F$. There, we also show for all $w^\star \in \mathrm{WMEQ}_M(G)$, $x^\star = M^{1/2}w^\star \in \mathrm{MEQ}(V)$, and all $w \in \mathrm{ri}(\mathcal{W})$, $x = M^{1/2}w \in \mathrm{ri}(\mathcal{X})$ that:

$$m_{\min} \cdot \mathsf{D}_Q(w^\star, w) \;\le\; \mathsf{D}_R(x^\star, x) \;\le\; m_{\max} \cdot \mathsf{D}_Q(w^\star, w)\,. \tag{13}$$

**Intuition for reduction.** While $G$ satisfies the $M$-*weighted* variational equality of (6), the iterates $\{z_t\}$ are updated via $G$ in an *unweighted* manner. The reduction specifies the coordinate system under which the resulting operator satisfies the unweighted equality (MEQ) while also preserving the OFTRL/EGFTRL update structure. Analyzing $\mathsf{D}_Q(w^\star, w_t)$ thus reduces to controlling $\mathsf{D}_R(x^\star, x_t)$.

### 5.1 Energy Dissipation and Continuous-Time Intuition

To prove $\mathsf{D}_R(x^\star, x_t) \to 0$ under the scaled primal iterates $\{x_t\}$, our analysis is based on establishing the dissipation of a *shifted* energy function, related to $\mathsf{D}_R$, along the scaled dual iterates $\{y_t\}$.

**Shifted energy function.** For $x^\star \in \mathcal{X}$, let $F_{x^\star} : \mathbb{R}^D \to \mathbb{R}$ be the *shifted energy function*:

$$F_{x^\star}(y) = F(y) - \langle x^\star, y\rangle \;\text{ for all } y \in \mathbb{R}^D\,. \tag{14}$$

We then have the following key relationship between $\mathsf{D}_R$ and $F_{x^\star}$ (see the proof in Section D):

**Proposition 5.3** (Change in shifted energy and Bregman divergence). *Let $y, y' \in \mathbb{R}^D$, and let $x = \nabla F(y), x' = \nabla F(y') \in \mathrm{ri}(\mathcal{X})$. Then for any $x^\star \in \mathrm{ri}(\mathcal{X})$, it holds that:*

$$\mathsf{D}_R(x^\star, x') - \mathsf{D}_R(x^\star, x) \;=\; F_{x^\star}(y') - F_{x^\star}(y)\,.$$

**Dissipation of shifted energy.** Given Proposition 5.3, it then suffices to establish *dissipative* bounds under OFTRL and EGFTRL on the one-step change in $F_{x^\star}$ for $x^\star \in \mathrm{MEQ}(V)$. This first requires the following notation. For all $y \in \mathbb{R}^D$, the Hessian $\nabla^2 F(y)$ is positive semi-definite and induces a local degenerate inner product and local semi-norm: for $u, v \in \mathbb{R}^D$ define $\langle u, v\rangle_y := \langle u, \nabla^2 F(y) v\rangle$ and $\|v\|_y := \sqrt{\langle v, \nabla^2 F(y) v\rangle}$ (see Section G.1 for further background). Then, our main technical lemma shows the dissipation of $F_{x^\star}$ along the dual iterates, quantified under the semi-norm:

**Lemma 5.4** (Dissipation of shifted energy). *Let Assumptions 1, 2, 3, 4 hold. Let $\{(x_t, y_t)\}$ be the* OFTRL *or* EGFTRL *iterates from Proposition 5.2 with $\eta \in (0, 1/(880LB\omega\kappa_M)]$. Then for $t \ge 1$:*

$$-\tfrac{7}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \;\le\; F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \;\le\; -\tfrac{1}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2\,, \quad \forall\, x^\star \in \mathrm{MEQ}(V).$$

Note that these energy dissipation bounds hold simultaneously for all points $x^\star \in \mathrm{MEQ}(V)$, and the magnitude of dissipation is quantified under the geometry of $\nabla^2 F(y_t)$ at the current dual iterate. Moreover, these bounds are tight (with the lower and upper inequality matching up to an absolute constant), and thus when combined with Proposition 5.3, the lemma implies that the magnitude of dissipation exactly controls the change in $\mathsf{D}_R(x^\star, x_t)$ under the iterates for all $x^\star \in \mathrm{MEQ}(V)$.

**Intuition from continuous time.** The starting point for understanding the dissipation in Lemma 5.4 is the continuous-time analogue of FTRL, which is the first-order ODE that we refer to as *FTRL flow*:

$$\dot{y}_t = -V(x_t) \quad \text{and} \quad x_t = \nabla F(y_t)\,. \tag{FTRL flow}$$

Importantly, along (FTRL flow), it holds for every $x^\star \in \mathrm{MEQ}(V)$ that the Bregman divergence from $x^\star$ (and equivalently the shifted energy by Proposition 5.3) is *exactly conserved*. Indeed, differentiating Proposition 5.3 along the (FTRL flow) trajectory and using the chain rule, we have for all initializations and all $x^\star \in \mathrm{MEQ}(V)$ that:

$$\begin{aligned}\tfrac{d}{dt}\mathsf{D}_R(x^\star, x_t) = \tfrac{d}{dt}F_{x^\star}(y_t) &= \langle \dot{y}_t, \nabla F_{x^\star}(y_t)\rangle \\ &= -\langle V(x_t), \nabla F(y_t) - x^\star\rangle = -\langle V(x_t), x_t - x^\star\rangle = 0\,.\end{aligned} \tag{15}$$

Here, the final equality holds exactly due to the structural (MEQ) property. This conservation was shown in Legacci et al. (2024a) and is similar to other continuous-time conservation results for two-player zero-sum games. A consequence of (15) is that the behavior of discrete-time FTRL variants can be better understood when viewed as discretizations of the (FTRL flow):

**FTRL and Implicit FTRL.** Taking a forward (resp. backward) *Euler discretization* of (FTRL flow) with stepsize $\eta > 0$ then yields the vanilla (FTRL) (resp. implicit (IFTRL)) update for the symmetrized system. A similar calculation to (15) in discrete time yields for all $x^\star \in \mathrm{MEQ}(V)$:

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) = \begin{cases} \mathsf{D}_F(y_{t+1}, y_t) \;\ge\; 0 & \text{for FTRL} \\ -\mathsf{D}_F(y_t, y_{t+1}) \;\le\; 0 & \text{for IFTRL}\,. \end{cases} \tag{16}$$

In other words, the continuous-time energy conservation under (FTRL flow) becomes energy *expansion* under the forward discretization (FTRL), and energy *dissipation* under the backward discretization (IFTRL) (see Proposition F.1 for the proof). The resulting $\mathsf{D}_F$ (resp. $-\mathsf{D}_F$) term captures the *second-order* forward (resp. backward) discretization error. Under this perspective, (OFTRL) and (EGFTRL) can both be viewed as implementable approximations of the idealized (IFTRL), which leads to the dissipation in Lemma 5.4. More concretely, the proof of the lemma follows by coupling OFTRL and EGFTRL at each $t \geq 1$ with an *implicit* step, which we sketch below:

**Proof sketch of Lemma 5.4.** To set up the coupling, we establish the following proposition:

**Proposition 5.5** (Implicit step coupling). *Let $\{(x_t, y_t)\}$ be a sequence satisfying $x_t = \nabla F(y_t)$ for all $t \geq 0$. Further, define the implicit update $\widehat{y}_{t+1} := y_t - \eta V(x_{t+1})$ for $t \geq 0$. Then for $t \geq 1$:*

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \;=\; \underbrace{-\mathsf{D}_F(y_t, y_{t+1})}_{\text{Dissipation term}} + \underbrace{\langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle}_{\text{Coupling error term } C_t}, \quad \forall\, x^\star \in \mathrm{MEQ}(V)\,.$$

Fixing any $x^\star \in \mathrm{MEQ}(V)$, Proposition 5.5 then shows that the one-step change in shifted-energy exactly decomposes into a *dissipation* term (that matches the one-step change of (IFTRL) from (16)), and a *coupling error term* $C_t$. Under the dual iterates of OFTRL and EGFTRL, we prove that both $-\mathsf{D}_F(y_t, y_{t+1}) \lesssim -\eta^2 \|V(x_t)\|_{y_t}^2$ and that $C_t \lesssim \eta^2 \|V(x_t)\|_{y_t}^2$, and for a sufficiently small constant stepsize $\eta$, the constant scaling of the negative dissipation term dominates that of $C_t$. In particular, to control the error terms, we rely on consequences of the local Hessian stability property (LHS) (Assumption 4). This establishes the upper bound of the lemma, and the matching lower bound follows similarly. The full proof is given in Section H.1. ■

### 5.2 Operator Domination Inequality

Next, we prove a structural inequality relating the local seminorms from Lemma 5.4 back to $\mathsf{D}_R$. For this, we extend the notation from Section 3.2 to the present, symmetrized coordinates: for $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0) \in \mathrm{ri}(\mathcal{X})$, let $\mathcal{R}(x_0) := \{\nabla F(y_0 + y) : y \in \mathrm{Span}(V(\mathcal{X}))\}$, and let $\mathrm{NE}(V) \subseteq \mathcal{X}$ denote the solutions of the Stampacchia variational inequality (3) for the operator $V$.

**Proposition 5.6** (Operator domination on reachable set). *Let Assumption 5 hold. Fix $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0)$ such that $\mathcal{R}(x_0) \not\subseteq \mathrm{NE}(V)$. Let $x^\star \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$, and let $\mathcal{K} \subset \mathrm{ri}(\mathcal{X})$ be any compact set such that $x^\star \in \mathcal{K}$. Then there exists a positive constant $\widetilde{\gamma}(x_0, \mathcal{K}) > 0$ such that*

$$\|V(x)\|_y^2 \;\geq\; \widetilde{\gamma}(x_0, \mathcal{K}) \cdot \mathsf{D}_R(x^\star, x)\,, \quad \forall\, y \in \mathbb{R}^D \text{ s.t. } x = \nabla F(y) \in \mathcal{K} \cap \mathcal{R}(x_0)\,.$$

Proposition 5.6 establishes a structural inequality between the dissipation term $\|V(x)\|_y^2$ and the Bregman divergence $\mathsf{D}_R(x^\star, x)$ that is analogous in spirit to a gradient-domination inequality in optimization. In general, such a relationship need not hold uniformly over the entire primal space. However, under the regularity condition of Assumption 5, we show that a local error bound around a reachable MEQ point $x^\star \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$ can be made uniform over $\mathcal{K} \cap \mathcal{R}(x_0)$ for any compact subset $\mathcal{K} \subset \mathrm{ri}(\mathcal{X})$ containing $x^\star$. This yields a positive domination constant $\widetilde{\gamma}(x_0, \mathcal{K})$ for which the inequality holds uniformly over $\mathcal{K} \cap \mathcal{R}(x_0)$, but with dependencies on the initialization $x_0$ and the set $\mathcal{K}$ (and, intrinsically, on geometric properties of the regularizer over this set). We give the full details of the proof in Section J.

**Combining the pieces.** To combine the pieces, we show that the dissipation bounds of Lemma 5.4 imply that all OFTRL and EGFTRL iterates lie in a compact set. In Lemma I.1 (see Section I), we use this compactness to then prove the asymptotic equilibrium selection result of part (i) of the theorem. Compactness then further allows us to apply the operator domination inequality of Proposition 5.6 along the OFTRL and EGFTRL iterates. Together with the upper bound of Lemma 5.4, this recursively establishes the linear convergence rate of part (ii).

## 6 Conclusion

This work establishes linear last-iterate convergence rates for OFTRL and EGFTRL in harmonic games, and more broadly for operators satisfying the global (weighted) Minty equality. Several directions remain open, including adaptive stepsize variants, and a finer understanding of the instance dependence of the linear rate. More broadly, it would be interesting to investigate whether our techniques can be used in general games through their harmonic and potential decompositions.

## AI USE STATEMENT

In this work, we used generative AI tools to proofread and suggest editing changes for the technical writing and proofs. We also used generative AI tools for verifying properties of example regularizers and for generating and verifying examples of harmonic games. Finally, we used generative AI tools to suggests edits of our manuscript to improve readability, and to run simulations of the algorithms we study. We take responsibility for the final content of this work, including all proofs.

## REFERENCES


Joseph Abdou, Nikolaos Pnevmatikos, Marco Scarsini, and Xavier Venel. Decomposition of games: some strategic considerations. *Mathematics of Operations Research*, 47(1):176–208, 2022.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2022)*, 2022.

Ioannis Anagnostides, Gabriele Farina, Tuomas Sandholm, and Brian Hu Zhang. A polynomial-time algorithm for variational inequalities under the minty condition. *arXiv preprint arXiv:2504.03432*, 2025.

Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The rate of convergence of bregman proximal methods: Local geometry versus regularity versus sharpness. *SIAM Journal on Optimization*, 34(3):2440–2471, 2024.

James Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: Vanishing regret with non-vanishing step sizes. *Advances in Neural Information Processing Systems (NeurIPS 2019)*, 2019a.

James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC 2018)*, 2018.

James P Bailey and Georgios Piliouras. Multi-agent learning in network zero-sum games is a Hamiltonian system. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)*, 2019b.

Anas Barakat, Wayne Lin, John Lazarsfeld, and Antonios Varvitsiotis. Optimistic online learning in symmetric cone games. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022a.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. *Advances in Neural Information Processing Systems*, 35:33904–33919, 2022b.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Separation of nonergodic uniform convergence rates for regularized learning in games. *Operations Research*, 2026.

Ozan Candogan, Ishai Menache, Asuman Ozdaglar, and Pablo A Parrilo. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503, 2011.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, pp. 807–834. PMLR, 2019.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science Conference, (ITCS 2019)*, 2019.

Constantinos Daskalakis and Christos H Papadimitriou. On a network generalization of the minmax theorem. In *International Colloquium on Automata, Languages, and Programming (ICALP 2009)*, 2009.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems (NeurIPS 2021)*, 2021.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Michail Fasoulakis, Evangelos Markakis, Yannis Pantazis, and Constantinos Varsos. Forward looking best-response multiplicative weights update methods for bilinear zero-sum games. In *International Conference on Artificial Intelligence and Statistics*, pp. 11096–11117. PMLR, 2022.

Michail Fasoulakis, Evangelos Markakis, Giorgos Roussakis, and Christodoulos Santorinaios. Improved last-iterate convergence properties for the flbr-mwu dynamics. *arXiv preprint arXiv:2608.01170*, 2026.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pp. 1758–1784. PMLR, 2020b.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O(1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022a.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022b.

Eduard Gorbunov, Adrien Taylor, Samuel Horváth, and Gauthier Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: the case of negative comonotonicity. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *Conference on Learning Theory*, pp. 2388–2422. PMLR, 2021.

Jonas E Katona, Xiuyuan Wang, and Andre Wibisono. A symplectic analysis of alternating mirror descent. *Journal of Machine Learning Research*, 27(44):1–61, 2026.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

John Lazarsfeld, Georgios Piliouras, Ryann Sim, and Stratis Skoulakis. Optimism without regularization: Constant regret in zero-sum games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025a.

John Lazarsfeld, Georgios Piliouras, Ryann Sim, and Andre Wibisono. Fast and furious symmetric learning in zero-sum games: Gradient descent as fictitious play. *Conference on Learning Theory (COLT 2025)*, 2025b.

John Lazarsfeld, Anas Barakat, Georgios Piliouras, Antonios Varvitsiotis, and Andre Wibisono. When and why is optimistic multiplicative weights slow? the geometry of energy dissipation. *arXiv preprint arXiv:2605.13242*, 2026.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. *Advances in Neural Information Processing Systems (NeurIPS 2021)*, 2021.

Davide Legacci, Panayotis Mertikopoulos, Christos Papadimitriou, Georgios Piliouras, and Bary Pradelski. No-regret learning in harmonic games: Extrapolation in the face of conflicting interests. *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024a.

Davide Legacci, Panayotis Mertikopoulos, and Bary Pradelski. A geometric decomposition of finite games: Convergence vs. recurrence under exponential weights. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024b.

Qi Lei, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, 2021.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the twenty-ninth annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2018)*, 2018.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

George J Minty. On the generalization of a direct method of the calculus of variations. *Bull. Amer. Math. Soc.*, 73(6):315–321, 1967.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

John Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Francesco Orabona. Last-iterate convergence of optimistic multiplicative weight update. *arXiv preprint arXiv:2606.11773*, 2026.

Thomas Pethick, Ioannis Mavrothalassitis, and Volkan Cevher. Efficient interpolation between extragradient and proximal methods for weak mvis. In *International Conference on Learning Representations*, volume 2025, pp. 50896–50923, 2025.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems (NeurIPS 2013)*, 2013.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Walter Rudin. Functional analysis 2nd ed. *International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York*, 202, 1991.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Ashkan Soleymani, Georgios Piliouras, and Gabriele Farina. Faster rates for no-regret learning in general games via cautious optimism. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC 2025)*, 2025.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems (NeurIPS 2015)*, 2015.

Taira Tsuchiya. Tight regret upper and lower bounds for optimistic hedge in two-player zero-sum games. In *The 29th International Conference on Artificial Intelligence and Statistics (AISTATS 2026)*, 2026.

Guanghui Wang, Krishna Acharya, Lokranjan Lakshmikanthan, Juba Ziani, and Vidya Muthukumar. Last-iterate convergence for symmetric, general-sum, $2 \times 2$ games under the exponential weights dynamic. In *37th International Conference on Algorithmic Learning Theory*, 2026.

Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jun-Kun Wang, Jacob Abernethy, and Kfir Y Levy. No-regret dynamics in the fenchel game: A unified framework for algorithmic convex optimization. *Mathematical Programming*, 205(1): 203–268, 2024.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

Andre Wibisono, Molei Tao, and Georgios Piliouras. Alternating mirror descent for constrained min-max games. *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Jiduan Wu, Anas Barakat, Ilyas Fatkhullin, and Niao He. Learning zero-sum linear quadratic games with improved sample complexity and last-iterate convergence. *SIAM Journal on Control and Optimization*, 63(5):3244–3271, 2025.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Brian Hu Zhang, Ioannis Anagnostides, and Tuomas Sandholm. Scale-invariant regret matching and online learning with optimal convergence: Bridging theory and practice in zero-sum games. *arXiv preprint arXiv:2510.04407*, 2025.

## TABLE OF CONTENTS


**1 Introduction** 1

**2 Preliminaries and Setting** 3

**3 A Global View of Optimistic and Extragradient FTRL** 5

**4 Main Result: Linear Convergence Rates in Harmonic Games** 6

**5 Overview of Proof of Theorem 4.1** 7

**6 Conclusion** 9

**A Additional Related Work** 16

**B Additional Preliminaries** 17

**C Details on FTRL Dynamics** 23

**D Details on Reduction to Symmetrized System** 28

**E Details on Effective Dual Space and Reachable Primal Sets** 33

**F Intuition from Continuous Time** 36

**G Properties of Energy Function Hessian** 37

**H Proof of Lemma 5.4: Energy Dissipation** 42

**I Details on Equilibrium Selection** 53

**J Proof of Proposition 5.6: Operator Domination Inequality** 56

**K Proof of Theorem 4.1: Last-Iterate Convergence Rates** 58

**L Proof of Theorem 4.5: Divergence of FTRL** 59

## A Additional Related Work

**Optimistic and extragradient methods.** Optimistic and extragradient methods have a long history in optimization (Korpelevich, 1976; Popov, 1980; Nemirovski, 2004; Nesterov, 2007). Our coupling proof strategy between OFTRL and Extragradient FTRL with IFTRL is reminiscient of the unified analysis proposed in Mokhtari et al. (2020). Showing that both OGDA and the extragradient method can be interpreted as approximate variants of the proximal point method, Mokhtari et al. (2020) prove linear last-iterate convergence rates for both optimistic and extragradient methods. In contrast to their *unconstrained* convex-concave min-max problem, our setting is *constrained*. The interpretation of the extragradient method as an approximation of the Proximal Point method was first discussed and developed in Nemirovski (2004).

**Last-iterate convergence in monotone variational inequalities and beyond.** Last iterate convergence results for optimistic gradient and extragradient methods were proved in Golowich et al. (2020b;a) and later tightened under more relaxed conditions (Gorbunov et al., 2022b;a; Cai et al., 2022a). A line of works also established convergence guarantees *beyond* monotone variational inequalities (see, e.g., Gorbunov et al. (2023); Pethick et al. (2025) and the references therein).

**Other works on optimistic online learning methods.** Optimistic variants of online learning algorithms have also been studied for their connections to accelerated methods in optimization (Wang & Abernethy, 2018; Wang et al., 2024). Recent work has also analyzed the optimistic FTRL methods in symmetric cone games (Barakat et al., 2026).

**Dual and energy-based perspectives on learning in games.** There has been a growing body of work in the past decade that study learning dynamics in games using dual and physics-inspired perspectives, in particular for FTRL-type dynamics; see, e.g., Mertikopoulos et al. (2018), Bailey & Piliouras (2018), Bailey & Piliouras (2019a), Bailey & Piliouras (2019b), Wibisono et al. (2022), Azizian et al. (2024), Lazarsfeld et al. (2025a;b), Katona et al. (2026), and the references therein.

**Learning in harmonic games.** We refer the reader to Legacci et al. (2024b) for a discussion on the dynamics of harmonic games within a dynamics-driven decomposition of games, where it is further shown that the continuous-time exponential weight dynamics are Poincaré recurrent in harmonic games. Besides Legacci et al. (2024a) which is discussed in more details in the main body of this work, several recent works have developed convergence results in harmonic games. Under a weighted Minty condition extending the standard Minty condition in variational inequalities, Zhang et al. (2025) provide *best-iterate* guarantees for parameter-free variants of predictive regret matching. In particular, this result applies to harmonic games which satisfy the aforementioned weighted Minty condition. Taking a computational perspective, Anagnostides et al. (2025) provide a polynomial-time algorithm for computing approximate Nash equilibria in harmonic games.

**Last-iterate convergence results for FTRL-type dynamics.** There have been a series of last-iterate results obtained for optimistic and extragradient FTRL variants in several continuous settings, but only with asymptotic guarantees; see, e.g., Mertikopoulos et al. (2019), Lei et al. (2021), Hsieh et al. (2021), Fasoulakis et al. (2022). Recent works of Cai et al. (2024; 2026) have also established a quantitative separation in the last-iterate, best-iterate, and random-iterate guarantees of OFTRL dynamics in two-player zero-sum games. Recently, Fasoulakis et al. (2026) also obtained a linear last-iterate convergence rate in two-player zero-sum games for a certain extragradient variant of MWU. Wang et al. (2026) proved last-iterate guarantees for the *vanilla* variant of MWU in two-player, two-action *symmetric games*, which comes in contrast to the divergence result for vanilla FTRL that we prove for harmonic games.

**Further results on last-iterate convergence in min-max optimization.** Recently, beyond existing results in two-player zero-sum games, Orabona (2026) proved asymptotic last-iterate convergence of OMWU with a small enough constant stepsize in convex-concave saddle point problems. We refer the reader to Section 2 in Orabona (2026) for a more detailed related work discussion on existing last-iterate convergence results. Beyond static two-player zero-sum games, Wu et al. (2025) consider finite horizon zero-sum linear quadratic dynamical games and establish a linear last-iterate convergence in terms of the objective function-value gap for a *nested* natural policy gradient algorithm. Their proof technique relies on using a gradient domination inequality proved for the primal function. We note that their algorithm is nested and the two players need to take turns whereas we consider a multiplayer setting where our players make decisions simultaneously.

# B Additional Preliminaries

## B.1 Variational Preliminaries

Here, we provide additional preliminaries on the variational perspectives introduced in Section 2.

### B.1.1 Overview of spaces.

Recall that $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_N \subseteq \mathbb{R}^D$, where each $\mathcal{W}_i = \Delta_{d_i}$. Moreover, recall from (9) that we define the linear subspaces $\mathcal{S}$ and $\mathcal{S}^\perp$ by

$$\mathcal{S} := \prod_{i=1}^N \mathrm{Span}(\mathbf{1}_{d_i}) \quad \text{and} \quad \mathcal{S}^\perp := \{v \in \mathbb{R}^D : \langle v, s\rangle = 0\,, \forall s \in \mathcal{S}\}\,.$$

In particular, we note that $\mathcal{S}$ and $\mathcal{S}^\perp$ are the *normal* and *tangent* spaces of the affine hull $\mathrm{aff}(\mathcal{W})$. In particular, recall that $\mathrm{aff}(\mathcal{W}) = w' + \mathcal{S}^\perp$ for any $w' \in \mathcal{W}$. Thus for any $w' \in \mathcal{W}$, we also have that

$$\mathcal{S}^\perp = \mathrm{Span}\{w - w' : w \in \mathcal{W}\}\,. \tag{17}$$

We defer further background to, e.g., (Rockafellar & Wets, 1998, Section 6).

We then recall the following key property of these spaces:

**Lemma B.1.** *Let $w, w' \in \mathcal{W}$. Then $w - w' \in \mathcal{S}^\perp$.*

**Proof.** Fix $w, w' \in \mathcal{W}$. Then for all components $i \in [N]$, it holds by definition of $\mathcal{W}_i$ that for any $s_i \in \mathrm{Span}(\mathbf{1}_{d_i})$, then $\langle s_i, w_i\rangle = \langle s_i, w_i'\rangle = c$ for some $c \in \mathbb{R}$. Taking their difference then yields $\langle s_i, w_i - w_i'\rangle = 0$. The statement of the lemma then follows from the fact that $\mathcal{S}$ is the componentwise product of all $\mathrm{Span}(\mathbf{1}_{d_i})$ for $i \in [N]$, and by definition of $\mathcal{S}^\perp$. ■

### B.1.2 Properties of Global Loss Operator

Recall that $G : \mathcal{W} \to \mathbb{R}^D$ is the operator from (2). By definition, $G$ is continuous and also satisfies the following global Lipschitz and boundedness properties:

**Assumption 1** (Lipschitz and bounded operator)**.** There exist constants $L \geq 1$ and $B \geq 1$ such that (i) $\|G(w') - G(w)\|_* \leq L\|w' - w\|$ and (ii) $\|G(w)\|_* \leq B$ for all $w, w' \in \mathcal{W}$.

Moreover, recall the variational characterization of Nash equilibria from (3): $w^\star \in \mathcal{W}$ is a Nash equilibrium if and only if

$$\langle G(w^\star), w - w^\star\rangle \geq 0 \quad \text{for all } w \in \mathcal{W}\,. \tag{NE}$$

The proof of this equivalence is standard and is based on first-order optimality conditions of the NE condition defined in terms of the individual losses $\ell_i$ (see, e.g., Facchinei & Pang (2003), Proposition 1.4.2). Recall that we then write $\mathrm{NE}(G) \subset \mathcal{W}$ to denote the set of all solutions[2] $x^\star$ of (NE).

**Interior Nash Equilibria and Normal Operator Vectors.**

In the following proposition, we establish an equivalence between interior equilibria $x^\star \in \mathrm{ri}(\mathcal{X})$ and points $x \in$ with normal operator values $G(w) \in \mathcal{S}$:

**Proposition B.2** (NE vs. Normal Operator Value)**.** *Fix $G : \mathcal{W} \to \mathbb{R}^D$ and consider $w' \in \mathcal{W}$. Then:*

*(i)* $G(w') \in \mathcal{S} \implies w' \in \mathrm{NE}(G)$.

*(ii) If* $w' \in \mathrm{ri}(\mathcal{W}) \cap \mathrm{NE}(G)$*, then* $G(w') \in \mathcal{S}$.

**Proof.** We prove the two claims separately:

*Proof of (i).* By definition of $\mathcal{S}$ and $\mathcal{S}^\perp$, we have

$$G(w') \in \mathcal{S} \implies \langle G(w'), v\rangle = 0 \ \text{ for all } v \in \mathcal{S}^\perp.$$

As $w - w' \in \mathcal{S}^\perp$ for all $w \in \mathcal{W}$ (Lemma B.1), this further implies that $\langle G(w'), w - w'\rangle = 0$ for all $w \in \mathcal{W}$, which means $x'$ satisfies (NE).

[2] A reader familiar with variational inequalities will recognize (3) (and equivalently (NE)) as simply the set of Stampacchia VI (SVI) solutions for $G$. For the setting introduced in Section 2, these two sets coincide.

*Proof of (ii).* First, if $w' \in \text{ri}(\mathcal{W})$, then there exists a sufficiently small $\epsilon > 0$ such that

$$w' + \epsilon v \in \mathcal{W} \text{ and } w' - \epsilon v \in \mathcal{W} \text{ for all } v \in \mathcal{S}^{\perp}.$$

Then, since $w'$ satisfies (NE), it holds for all $v \in \mathcal{S}^{\perp}$ that

$$\begin{aligned} \langle G(w'), (w' + \epsilon v) - w' \rangle \geq 0 \implies \epsilon \langle G(w'), v \rangle \geq 0 \\ \text{and also} \quad \langle G(w'), (w' - \epsilon v) - w' \rangle \geq 0 \implies \epsilon \langle G(w'), v \rangle \leq 0 \,. \end{aligned}$$

As $\epsilon > 0$, this together implies $\langle G(w'), v \rangle = 0$ for all $v \in \mathcal{S}^{\perp}$, and thus $G(w') \in \mathcal{S}$. ■

### B.1.3 Polymatrix Games and Linear Operators

For the normal-form game setting introduced in Section 2, we have the linear characterization of $G$ when the underlying game is a *polymatrix* game (e.g., Daskalakis & Papadimitriou (2009)).

**Proposition B.3** (Linear operators and polymatrix games)**.** *Let $G : \mathcal{W} \to \mathbb{R}^D$ be the operator from (2) and suppose the underlying game is polymatrix. Then $G(w) = Pw$ for some $P \in \mathbb{R}^{D \times D}$.*

**Proof.** In a polymatrix game, there exist matrices $A^{i,j} \in \mathbb{R}^{d_i, d_j}$ for every pair $(i, j) \in [N] \times [N]$, where $A^{i,i} = 0$ for all $i \in [N]$. For each player $i \in [N]$, the loss $\ell_i : \mathcal{W} \to \mathbb{R}$ is the sum of bilinear terms

$$\ell_i(w) = \textstyle\sum_{j \neq i}^{N} \langle w_i, A^{i,j} w_j \rangle \text{ for all } w \in \mathcal{W}. \tag{18}$$

Then by definition of $G$ from (2), we have $G_i(w) = \nabla_{w_i} \ell_i(w) = \sum_{j \neq i}^{N} A^{i,j} w_j$ for all $i \in [N]$. Thus, for $P \in \mathbb{R}^{D \times D}$ with blocks $P_{i,j} = A^{i,j}$ for all $(i, j) \in [N] \times [N]$, it follows that $G(w) = Pw$, and hence $G$ is linear. ■

## B.2 Preliminaries on Harmonic Games

In this section, we give several further preliminaries on harmonic games.

### B.2.1 Details on Variational Characterization of Harmonic Games

As introduced in Section 2.2, harmonic games from Definition 2.1 admit the following variational characterization in terms of the operator $G$, a result due to Legacci et al. (2024a):

**Proposition 2.3** (Legacci et al. (2024a), Proposition A.4)**.** *A game with loss operator $G : \mathcal{W} \to \mathbb{R}^D$ is harmonic if and only if there exist (i) a matrix $M = \text{Diag}(m_1 I_{d_1}, \ldots, m_N I_{d_N})$ where $m_i > 0$ for all $i \in [N]$ and (ii) an interior point $w' \in \text{ri}(\mathcal{W})$ such that:*

$$\langle MG(w), w - w' \rangle = 0 \text{ for all } w \in \mathcal{W} \,. \tag{5}$$

Recall from (6) that for weights matrix $M$ satisfying Proposition 2.3, we let $\text{WMEQ}_M(G)$ be the set:

$$\text{WMEQ}_M(G) = \{w' \in \mathcal{W} : \langle MG(w), w - w' \rangle = 0 \text{ for all } w \in \mathcal{W}\} \,. \tag{19}$$

Moreover, we further have by the proposition that $\text{WMEQ}_M(G) \cap \text{ri}(\mathcal{W}) \neq \emptyset$. Note that $\text{WMEQ}_M(G)$ can be viewed as the solution set to an $M$-weighted Minty variational equality. Importantly, this implies that (1) the set $\text{WMEQ}_M(G)$ is convex, and (2) $\text{WMEQ}_M(G) \subseteq \text{NE}(G)$. We prove these properties as follows:

**Lemma B.4** (Convexity of WMEQ set)**.** *Let $G : \mathcal{W} \to \mathbb{R}^D$ satisfy $\text{WMEQ}_M(G) \neq \emptyset$ for a weights matrix $M$. Then $\text{WMEQ}_M(G)$ is convex.*

**Proof.** Let $w', w'' \in \text{WMEQ}_M(G)$. For $\lambda \in [0, 1]$, let $w_\lambda := \lambda w' + (1 - \lambda) w''$. By convexity of $\mathcal{W}$, we have $w_\lambda \in \mathcal{W}$ for all $\lambda \in [0, 1]$. Then for any $w \in \mathcal{W}$, observe that

$$\langle MG(w), w - w_\lambda \rangle = \lambda \langle MG(w), w - w' \rangle + (1 - \lambda) \langle MG(w), w - w'' \rangle = 0$$

Here, the final equality is due to $w', w'' \in \text{WMEQ}_M(G)$. Thus $w_\lambda$ satisfies (19) for all $w \in \mathcal{W}$, meaning also $w_\lambda \in \text{WMEQ}_M(G)$. By definition, this means $\text{WMEQ}_M(G)$ is convex. ■

**Lemma B.5** (WMEQ points are NE). *Let $G : \mathcal{W} \to \mathbb{R}^D$ satisfy $\mathrm{WMEQ}_M(G) \neq \emptyset$ for a weights matrix $M$. Then $\mathrm{WMEQ}_M(G) \subseteq \mathrm{NE}(G)$, and $G(w^\star) \in \mathcal{S}$ for every $w^\star \in \mathrm{WMEQ}_M(G)$.*

**Proof.** Fix $w^\star \in \mathrm{WMEQ}_M(G)$. For each player $i \in [N]$ and any $w_i \in \mathcal{W}_i$, we then apply (19) at the point $w = (w_i, w^\star_{-i})$. By the block diagonal structure of $M = \mathrm{Diag}(m_1 I_{d_1}, \dots, m_N I_{d_N})$, for $m_i > 0$, and since $G_i(w) = G_i(w^\star)$ by definition of $G$ in (2), this yields at the $i$-th component:

$$0 = m_i \langle G_i(w), w_i - w^\star \rangle = m_i \langle G_i(w^\star), w_i - w_i^\star \rangle = 0 \,. \tag{20}$$

Dividing by $m_i > 0$ then gives $\langle G_i(w^\star), w_i - w^\star \rangle = 0$ for all $w_i \in \mathcal{W}_i$. As this latter equality holds for all players simultaneously, summing over $i \in [N]$ then yields

$$\langle G(w^\star), w - w^\star \rangle = 0 \quad \text{for all } w \in \mathcal{W} \,. \tag{21}$$

Thus $w^\star$ satisfies (NE) with equality and so $w^\star \in \mathrm{NE}(G)$.

Moreover, as $\mathcal{S}^\perp = \mathrm{Span}\{w - w^\star : w \in \mathcal{W}\}$ by (17), expression (21) also immediately implies that $G(w^\star) \in \mathcal{S}$, which concludes the proof. ■

### B.2.2 Harmonic games with linear operators

When the operator $G$ satisfying Proposition 2.3 is *linear*, which is implied by all polymatrix settings (Proposition B.3), several further structural properties emerge.

First, we show that for linear harmonic $G$, the sets $\mathrm{WMEQ}_M(G)$ and $\mathrm{NE}(G)$ coincide.

**Proposition B.6** (NE and WMEQ for linear G). *Let $G : \mathcal{W} \to \mathbb{R}^D$ satisfy $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weights matrix $M$. Suppose $G(w) = Pw$ for $P \in \mathbb{R}^{D \times D}$. Then $\mathrm{WMEQ}_M(G) = \mathrm{NE}(G)$.*

**Proof.** As $\mathrm{WMEQ}_M(G) \subseteq \mathrm{NE}(G)$ by Lemma B.5, it suffices to establish $\mathrm{NE}(G) \subseteq \mathrm{WMEQ}_M(G)$ under the assumptions of the proposition. For this, let $w^\star \in \mathrm{NE}(G)$, and let $\widetilde{w} \in \mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W})$. Then since $\widetilde{w} \in \mathrm{WMEQ}_M(G)$, we have that

$$\textstyle\sum_{i=1}^N m_i \langle G_i(w^\star), \widetilde{w}_i - w_i^\star \rangle = 0 \,. \tag{22}$$

Since $w^\star \in \mathrm{NE}(G)$, the componentwise inequalities of (NE) and $m_i > 0$ then imply that $\langle MG(w^\star), w - w^\star \rangle \geq 0$ for all $w \in \mathcal{W}$. Combining this with (22) then yields $\langle MG(w^\star), w - \widetilde{w} \rangle \geq 0$ for all $w \in \mathcal{W}$. Then since $\widetilde{w} \in \mathrm{ri}(\mathcal{W})$, it holds for every $h \in \mathcal{S}^\perp$ that there exists sufficiently small $\epsilon > 0$ such that $\widetilde{w} \pm \epsilon h \in \mathcal{W}$. Substituting these two points into the previous inequality then yields that $\langle MG(w^\star), h \rangle = 0$. This implies that $MG(w^\star) \in \mathcal{S}$, and thus $G(w^\star) \in \mathcal{S}$ since $M^{-1}\mathcal{S} = \mathcal{S}$.

Now for any $w \in \mathcal{W}$, let $v = (w + w^\star)/2$. By linearity, and using the fact that $MG(w^\star) \in \mathcal{S}$ we can compute

$$0 = 4\langle MG(v), v - \widetilde{w} \rangle = \langle MG(w), w + w^\star - 2\widetilde{w} \rangle \,.$$

Then subtracting $\langle MG(w), w - \widetilde{w} \rangle = 0$ yields $\langle MG(w), w^\star - \widetilde{w} \rangle = 0$. Finally, it then follows that

$$\langle MG(w), w - w^\star \rangle = \langle MG(w), w - \widetilde{w} \rangle - \langle MG(w), w^\star - \widetilde{w} \rangle = 0 \,.$$

Thus $w^\star \in \mathrm{WMEQ}_M(G)$. ■

**Zero-sum games with an interior equilibrium.**

When $G$ is the loss operator associated with a two-player zero-sum game with payoff matrix $A \in d_1 \times d_2$, we further show that if $\mathrm{NE}(G) \cap \mathrm{ri}(\mathcal{W})$ is nonempty (meaning there exists an interior equilibrium) then $G$ is harmonic with identity weights matrix $I_D$.

**Proposition B.7** (Zero-sum games with interior NE). *Let $G : \mathcal{W} \to \mathbb{R}^D$ be the loss operator associated to a zero-sum game with payoff matrix $A \in \mathbb{R}^{d_1 \times d_2}$. Suppose that $\mathrm{NE}(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$. Then $G$ satisfies Proposition 2.3 with weights matrix $M = I_D$.*

Note that this result was shown by Legacci et al. (2024b) (Proposition A.10), but we give a separate proof here:

**Proof.** For a zero-sum game with payoff matrix $A$, the operator $G$ is the linear operator $G(w) = Kw$ for the skew-symmetric matrix $K = -K^\top \in \mathbb{R}^{D \times D}$ satisfying

$$K = \begin{pmatrix} 0 & A \\ -A^\top & 0 \end{pmatrix} \,.$$

See, e.g., Wibisono et al. (2022); Lazarsfeld et al. (2026). Moreover, recall by skew-symmetry that $\langle w, Kw\rangle = \langle Kw, w\rangle = 0$ for any $w \in \mathcal{W}$. Then, fix $w^\star \in \mathrm{NE}(G) \cap \mathrm{ri}(\mathcal{W})$. Then by (NE), we have for all $w \in \mathcal{W}$ that

$$\begin{aligned} 0 = \langle G(w^\star), w - w^\star\rangle &= \langle Kw^\star, w - w^\star\rangle \\ &= \langle Kw, w - w^\star\rangle - \langle K(w - w^\star), w - w^\star\rangle \\ &= \langle Kw, w - w^\star\rangle \\ &= \langle G(w), w - w^\star\rangle \ , \end{aligned}$$

where the third equality holds due to skew-symmetry of $K = -K^\top$. Thus, $w^\star \in \mathrm{WMEQ}_M(G)$ with weights matrix $M = I_D$, which by Proposition 2.3 means that the underlying game is harmonic. ■

Note that not every zero-sum game is Harmonic (see, e.g., Candogan et al. (2011); Abdou et al. (2022); Legacci et al. (2024a)), and the characterization of Proposition B.7 is specifically under the assumption that the game has an interior equilibrium.

### B.3 Relationship Between Gap Function and Bregman Divergence

Here, we establish relationships between Bregman divergence and the Gap function, as introduced in Remark 4.2:

**Proposition B.8** (Gap vs. Bregman divergence from WMEQ point)**.** *Let $G : \mathcal{W} \to \mathbb{R}^D$ satisfy $\mathrm{WMEQ}_M(G) \neq \emptyset$ for some weights matrix $M$, and let $G$ satisfy Assumption 1. Let $Q : \mathcal{W} \to \mathbb{R}$ be a regularizer satisfying Assumption 3. Define $\rho := \max_{w', w'' \in \mathcal{W}} \|w' - w''\|$. Then for all $w^\star \in \mathrm{WMEQ}_M(G)$, it holds for all $w \in \mathcal{W}$ that*

$$\tfrac{1}{2(L\rho)^2} \cdot \mathsf{Gap}(w)^2 \ \leq \ \tfrac{1}{2}\|w^\star - w\|^2 \ \leq \ D_Q(w^\star, w) \ .$$

**Proof.** Fix weights matrix $M$ and $w^\star \in \mathrm{WMEQ}_M(G)$. For the first inequality, recall by Lemma B.5 that $w^\star \in \mathrm{WMEQ}_M(G)$ implies $G(w^\star) \in \mathcal{S}$. Thus $\langle G(w^\star), w - w'\rangle = 0$ for all $w, w \in \mathcal{W}$. Thus for all $w, w' \in \mathcal{W}$, we have

$$\begin{aligned} \langle G(w), w - w'\rangle = \langle G(w) - G(w^\star), w - w'\rangle &\leq \|G(w) - G(w^\star)\|_* \cdot \|w - w'\| \\ &\leq L\rho \cdot \|w - w^\star\| \ , \end{aligned} \tag{23}$$

where the final inequality comes from the $L$-Lipschitzness of $G$ and the definition of $\rho$. Then using the definition of $\mathsf{Gap}(w) = \max_{w' \in \mathcal{W}} \langle G(w), w - w'\rangle$, the first inequality of the proposition follows by squaring both sides of (23) and rearranging. The second inequality of the proposition follows directly from the definition of 1-strong-convexity of $Q$ with respect to $\|\cdot\|$. ■

Finally, recall for any $w \in \mathcal{W}$ that if $\mathsf{Gap}(w) \leq \epsilon$, then $w$ is an $\epsilon$-NE (see, e.g., Facchinei & Pang (2003), Section 1.5.3).

### B.4 Examples of Harmonic Games

In this section, we give several additional examples of harmonic games that go beyond the canonical subclass of two-player zero-sum games with an interior NE (e.g., see Proposition B.7). In particular, we give two examples of harmonic games that are (1) multiplayer, (2) non-constant sum, and (3) have multiple equilibria.

#### B.4.1 Example 1: 3-player polymatrix game with WMEQ continuum.

First, we give an three-player polymatrix example. In particular, the plots of Figure 1 use an instantiate of this family. Note that while the underlying operator $G$ is linear for this example (due to Proposition B.3), the underlying game itself is non-constant-sum and non-monotone and has a continuum of WMEQ points. We give these details here:

**Weighted polymatrix setup.**

Consider three players with mixed-strategy space $\mathcal{W} = (\Delta_2)^3$. Write $w_i = (p_i, 1 - p_i) \in \Delta_2$ and $r_i = 2p_i - 1$. For positive weights $m_1, m_2, m_3 > 0$ that are not all equal, define the losses

$\ell_1(w) = r_1(m_2 r_2 - m_3 r_3)$, $\ell_2(w) = r_2(m_3 r_3 - m_1 r_1)$, and $\ell_3(w) = r_3(m_1 r_1 - m_2 r_2)$. The joint loss operator $G = (G_i)_{i=1}^3$ is given by

$$f(r) = \begin{pmatrix} m_2 r_2 - m_3 r_3 \\ m_3 r_3 - m_1 r_1 \\ m_1 r_1 - m_2 r_2 \end{pmatrix}, \quad \text{and} \quad G_i(w) = \begin{pmatrix} f_i(r) \\ -f_i(r) \end{pmatrix}.$$

Since each $r_i = w_{i,+} - w_{i,-}$, the operator $G$ is linear in $w$.

**Continuum of WMEQ points.**

Let $M = \mathrm{Diag}(m_1 I_2, m_2 I_2, m_3 I_2)$ and $m_{\min} = \min_i m_i$. It is straightforward to verify that the set $\mathrm{WMEQ}_M(G)$ is then given by

$$\mathrm{WMEQ}_M(G) = \{w \in \mathcal{W} : r_i = (t/m_i) \text{ for all } i, \ \ |t| \le m_{\min}\}. \tag{24}$$

Equivalently, in terms of the coordinates $p_i \in (0,1)$, each $w = (w_1, w_2, w_3) \in \mathrm{WMEQ}_M(G)$ is parameterized by $p_i = (1 + t/m_i)/2$ for $|t| \le m_{\min}$. It is straightforward to verify the WMEQ identity for every point of this segment by substituting into $\langle MG(w), w - w^\star \rangle$. As $G$ is linear and clearly $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, it follows by Proposition B.6 that $\mathrm{NE}(G) = \mathrm{WMEQ}_M(G)$.

As an example, the left-subplot in Figure 2 plots the continuum of $\mathrm{WMEQ}_M(G)$ for the case when $m_1 = 3, m_2 = 6, m_3 = 4$.

**Nonmonotonicity and nonconstant sum.**

While the operator $G$ is linear, we show that $G$ is non-monotone, and the underlying game is additionally non-constant-sum. For this, let $w^0$ be the uniform strategy profile. Then by definition of $G$ and $\ell_i$, we can compute for any other $w \in \mathcal{W}$

$$\begin{aligned} &\langle G(w) - G(w^0), w - w^0 \rangle \\ &\quad = \textstyle\sum_{i=1}^3 \ell_i(w) = (m_2 - m_1) r_1 r_2 + (m_1 - m_3) r_1 r_3 + (m_3 - m_2) r_2 r_3. \end{aligned} \tag{25}$$

As the weights $\{m_i\}$ are not all equal, at least one term in (25) has non-zero coefficient $m_i - m_j$. Supposing without loss of generality that $m_2 - m_1 > 0$, then set $w_3 = 0$, and choose $w_1, w_2 \in \Delta_2$ such that $r_1 r_2 < 0$. Then it follows from (25) that $\langle G(w) - G(w^0), w - w^0 \rangle < 0$, and thus $G$ is non-monotone. At this profile, the total loss is also non-zero. However, observe that when $w = w^0$, then (25) vanishes. Thus the total loss $\sum_{i=1}^3 \ell_i(w)$ is also non-constant-sum.

### B.4.2 Example 2: non-linear operator with non-convex NE set.

Here, we give an example family of harmonic games that do not correspond to linear polymatrix games. In particular, the following example induces non-linear loss operators $G$, have a continuum of WMEQ points, have a nonconvex set of NE points, are nonmonotone and correspond to a non-constant-sum game, and which additionally satisfy the restricted regularity condition of Assumption 5. We give more details here:

**Game setup.**

Consider three players with mixed-strategy space $\mathcal{W} = (\Delta_2)^3$. Write $w_i = (p_i, 1 - p_i)$ and $r_i = 2p_i - 1$, and let $w(r)$ denote the strategy profile with centered coordinates $r \in [-1, 1]^3$. Define the losses $\ell_1(w) = r_1 r_2 r_3$, $\ell_2(w) = -r_1 r_2 r_3$, and $\ell_3(w) = r_1 r_2$. These are the mixed extensions of the pure-action losses obtained by substituting $r_i = a_i \in \{-1, +1\}$.

**Nonlinear loss operator.**

Ordering each player's actions as $(+1, -1)$, the joint loss operator is given by

$$G(w) = \big(r_2 r_3,\ -r_2 r_3,\ -r_1 r_3,\ r_1 r_3,\ r_1 r_2,\ r_1 r_2\big).$$

The cross-interactions between $r_i r_j$ mean that $G$ is nonlinear on $\mathcal{W}$. For example, $G_{1,+}(w) = (2p_2 - 1)(2p_3 - 1)$ has mixed derivative $\partial^2 G_{1,+}/(\partial p_2\, \partial p_3) = 4$, so it is thus not affine in the probability coordinates. Therefore, there is no matrix $P$ such that $G(w) = Pw$ for $w \in \mathcal{W}$.

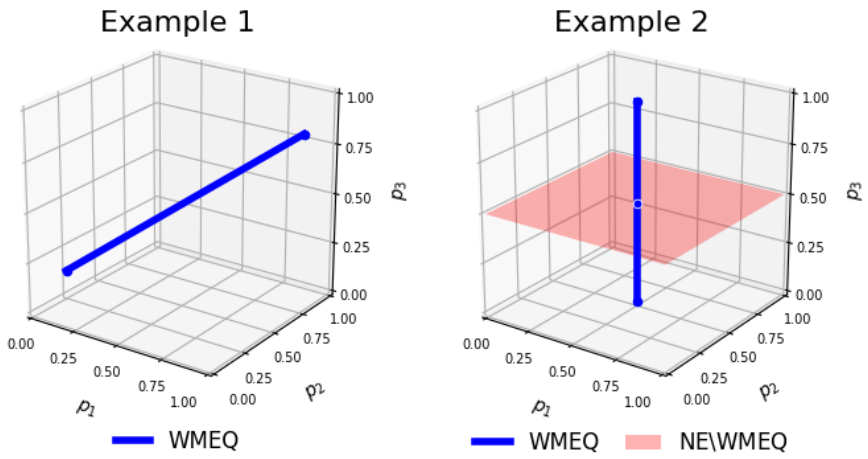


Figure 2: The sets $\mathrm{WMEQ}_M(G)$ and $\mathrm{NE}(G) \setminus \mathrm{WMEQ}_M(G)$ for the instantiation of Example 1 from Section B.4.1 with $m_1 = 3$, $m_2 = 6$, $m_3 = 4$ (left), and Example 2 from Section B.4.2.

**Continuum of MEQ points.**

Let $M = I_D$, where $D = 6$. For any centered profiles $r, s$, we can compute

$$\langle G(w(r)), w(r) - w(s) \rangle = -s_1 r_2 r_3 + s_2 r_1 r_3. \tag{26}$$

For every $r \in [-1, 1]^3$, observe that (26) vanishes if and only if $s_1 = s_2 = 0$. Hence

$$\mathrm{WMEQ}_{I_D}(G) = \mathrm{MEQ}(G) = \{w(0, 0, t) : t \in [-1, 1]\}. \tag{27}$$

Equivalently, the MEQ set is the segment $p_1 = p_2 = 1/2$, $p_3 \in [0, 1]$. Since it contains the uniform profile in $\mathrm{ri}(\mathcal{W})$, the game is harmonic due to Proposition 2.3.

**Nonconvex Nash equilibrium set.**

It is straightforward to verify that a profile correpsonds to an NE if and only if $r_2 r_3 = r_1 r_3 = 0$. Thus, including boundary profiles, we have that

$$\mathrm{NE}(G) = \mathrm{MEQ}(G) \cup \{w(r_1, r_2, 0) : (r_1, r_2) \in [-1, 1]^2\},$$
$$\text{and thus} \quad \mathrm{NE}(G) \setminus \mathrm{MEQ}(G) = \{w(r_1, r_2, 0) : (r_1, r_2) \in [-1, 1]^2 \setminus \{(0, 0)\}\}.$$

In particular, there is a continuum of interior but non-MEQ Nash equilibria. Thus, the full set $\mathrm{NE}(G)$ is nonconvex: $w(0, 0, 1/2)$ and $w(1/2, 0, 0)$ are both in $\mathrm{NE}(G)$, while their midpoint $w(1/4, 0, 1/4) \notin \mathrm{NE}(G)$. In particular, the right-subplot in Figure 2 plots both the continuum of $\mathrm{WMEQ}_M(G) = \mathrm{MEQ}(G)$ in blue and the continuum of $\mathrm{NE}(G) \setminus \mathrm{WMEQ}_M(G)$ in red. The full set $\mathrm{NE}(G)$ is the union of the two sets, which illustrates that $\mathrm{NE}(G)$ is non-convex.

**Nonmonotonicity and nonconstant sum.**

For $r = (1/2, 0, 1/2)$ and $s = (1/2, -1/2, 0)$, we have using a direct calculation that

$$\langle G(w(r)) - G(w(s)), w(r) - w(s) \rangle = -\tfrac{1}{8} < 0.$$

Thus $G$ is not monotone. As $\sum_{i=1}^{3} \ell_i(w) = r_1 r_2$ varies with $w$, the game is also not constant sum.

**Restricted regularity.**

We further show that $G$ satisfies the regularity condition from Assumption 5, and when the regularizer $Q$ and its energy function $H = Q^*$ satisfy Assumptions 2 and 4.

For this, first consider the standard (componentwise) negative Shannon entropy regularizer $Q(w) = \sum_{i=1}^{3} \sum_{a \in \{+,-\}} w_{i,a} \log w_{i,a}$, and let $H = Q^*$, which satisfy Assumptions 2 and 4, respectively. In reduced dual coordinates $b_i = (z_{i,+} - z_{i,-})/2$, its choice map is $r_i = \tanh b_i$, where $\tanh$ is the

hyperbolic tangent. Since player 3's loss block is normal to the simplex, the effective dual update space and reachable set are

$$\mathcal{U} = \{(u_1, -u_1, u_2, -u_2, 0, 0) : u \in \mathbb{R}^2\}, \qquad \mathcal{R}(w_0) = \{w(r_1, r_2, c) : (r_1, r_2) \in (-1, 1)^2\},$$

where $c = r_{3,0} = \tanh b_{3,0}$. Identifying $\mathcal{U}$ with $\mathbb{R}^2$, the effective stationarity map is then given by

$$\pi_0(u) = \begin{pmatrix} c\tanh(b_{2,0} + u_2) \\ -c\tanh(b_{1,0} + u_1) \end{pmatrix}.$$

If $c \neq 0$, the unique zero is $u^\star = (-b_{1,0}, -b_{2,0})$. Then we have $J\pi_0(u^\star) = \begin{pmatrix} 0 & c \\ -c & 0 \end{pmatrix}$ and thus $\det J\pi_0(u^\star) = c^2 > 0$. If $c = 0$, the entire reachable set $\mathcal{R}(w_0) \subset \mathrm{NE}(G)$ and is exempt from the regularity requirement. Thus Assumption (5) holds.

Note also that the same argument applies to any other separable regularizer $Q$ and energy function $H = Q^*$ satisfying Assumptions 2 and 4, respectively: for this, we can replace $\tanh$ by the scalar choice maps $h_i$, whose derivatives are strictly positive, which follows from the fact that $\mathrm{Null}(\nabla^2 H(z)) = \mathcal{S}$ (see Proposition C.5 (iii)). At a zero, the determinant of $J_{\pi_0}$ then becomes $c^2 h_1'(b_1^\star) h_2'(b_2^\star) > 0$ when $c \neq 0$, where $h_i(b_i^\star) = 0$, and the case $c = 0$ remains exempt.

# C Details on FTRL Dynamics

In this section, we give more details on the global dynamics of FTRL variants introduced in Section 3. First, we give in Section C.1 a more complete derivation of the global dynamics from their playerwise definitions. Then, we present in Section C.2 several key primal-dual relationships of the global regularizer and corresponding energy function that are used throughout. This also includes the proof of Proposition 3.2.

## C.1 Playerwise and Global FTRL Dynamics

Recall from Section 3 that each player $i \in [N]$ uses a strictly convex regularizer $Q_i : \mathcal{W}_i \to \mathbb{R}$ and a fixed stepsize $\eta_i > 0$. Recall also that we define the global stepsize $\eta := \max_{i \in [N]} \eta_i > 0$.

**Playerwise primal update.** The standard playerwise FTRL update rule (Shalev-Shwartz, 2012, Section 2.3), along with its Optimistic, Extragradient, and Implicit variants, all follow a similar primal update structure. First, we assume each player $i \in [N]$ has initial $w_{0,i} \in \mathrm{ri}(\mathcal{W}_i)$. Then for $t \geq 0$, they choose $w_{t+1,i} \in \mathcal{W}_i$ according to

$$w_{t+1,i} = \mathrm{argmin}_{w_i \in \mathcal{W}_i} \{-\langle w_i, v_{t+1,i}\rangle + Q_i(w_i)\} \tag{28}$$

where $v_{t+1,i} \in \mathbb{R}^{d_i}$ is a variant-specific *dual iterate* (specified below) that accumulates $\eta_i$-scaled operator vectors (i.e., of the form $\eta_i G_i(w)$). Note here that the argmin in (28) is unique due to strict convexity of $Q_i$.

**Playerwise, variant-specific dual iterates.** Recall that $w_t = (w_{t,1}, \ldots, w_{t,N}) \in \mathbb{R}^D$ is the globally concatenated primal iterate at times $t \geq 0$. Moreover, for each $i \in [N]$, each player has dual initialization $v_{0,i} \in \mathbb{R}^{d_i}$. Then variant-specific dual iterate $v_{t+1,i} \in \mathbb{R}^{d_i}$ is given by:

$$\begin{aligned} \text{for FTRL}: &\quad v_{t+1,i} = v_{t,i} - \eta_i G_i(w_t) \\ \text{for IFTRL}: &\quad v_{t+1,i} = v_{t,i} - \eta_i G_i(w_{t+1}) \\ \text{for OFTRL}: &\quad v_{t+1,i} = v_{t,i} - 2\eta_i G_i(w_t) + \eta_i G_i(w_{t-1}) \\ \text{for EGFTRL}: &\quad v_{t+1/2,i} = v_{t,i} - \eta_i G_i(w_t), \quad v_{t+1,i} = v_{t,i} - \eta_i G_i(w_{t+1/2}). \end{aligned} \tag{29}$$

Here, we use the convention for OFTRL that $w_{-1} = w_0$, and for EGFTRL, we let $w_{t+1/2}$ denote the concatenation of the minimizers in (28) with respect to $v_{t+1/2} = (v_{t+1/2,i})_{i \in [N]}$. When instantiating (28) with one of the dual iterate definitions in (29), we exactly obtain the standard form of the corresponding FTRL-type update rule (see, e.g., Syrgkanis et al. (2015) and Legacci et al. (2024a)).

Then, for a fixed playerwise dual sequence $\{v_{t,i}\}$, we also use the global stepsize $\eta > 0$ to define the $\eta$-normalized dual iterates $\{z_{t,i}\}$. Specifically, for all $i \in [N]$, let $z_{t,i} = \frac{\eta}{\eta_i} \cdot v_{t,i}$ for all $t \geq 0$. It then follows that, for any of the dual sequences in (29), the playerwise update in (28) for $t \geq 0$ can be equivalently written in terms of $z_{t,i}$ as

$$w_{t+1,i} = \mathrm{argmin}_{w_i \in \mathcal{W}_i} \{-\langle w_i, \tfrac{\eta_i}{\eta} z_{t+1,i}\rangle + Q_i(w_i)\} = \mathrm{argmin}_{w_i \in \mathcal{W}_i} \{-\langle w_i, z_{t+1,i}\rangle + \tfrac{\eta}{\eta_i} Q_i(w_i)\}. \tag{30}$$

Moreover, the variant-specific *scaled* dual sequences are explicitly given by:

$$\begin{aligned} \text{for FTRL}: &\quad z_{t+1,i} = z_{t,i} - \eta G_i(w_t) \\ \text{for IFTRL}: &\quad z_{t+1,i} = z_{t,i} - \eta G_i(w_{t+1}) \\ \text{for OFTRL}: &\quad z_{t+1,i} = z_{t,i} - 2\eta G_i(w_t) + \eta G_i(w_{t-1}) \\ \text{for EGFTRL}: &\quad z_{t+1/2,i} = z_{t,i} - \eta G_i(w_t)\,, \quad z_{t+1,i} = z_{t,i} - \eta G_i(w_{t+1/2})\,. \end{aligned} \tag{31}$$

**Global primal update.** Now consider the collection of any of the scaled playerwise dual sequences of (31), and let $z_t = (z_{t,1}, \dots, z_{t,N}) \in \mathbb{R}^D$ for all $t \geq 0$. Moreover, recall from (7) that we define the global regularizer $Q : \mathcal{W} \to \mathbb{R}$ by

$$Q(w) = \textstyle\sum_{i=1}^N \frac{\eta}{\eta_i} Q_i(w_i) \quad \text{for all } w = (w_1, \dots, w_N) \in \mathcal{W}\,.$$

Using the fact that $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_N$ is a product space and that $Q$ is separable, it then follows that, for $w_t = (w_{t,1}, \dots, w_{t,N}) \in \mathbb{R}^D$ where each $w_{t,i}$ is given by (30), it holds that

$$w_t = \operatorname{argmin}_{w \in \mathcal{W}} \{-\langle w, z_t\rangle + Q(w)\}\,, \tag{32}$$

which is exactly the global primal update rule from (8).

**Global dual updates.** Fix any of the playerwise dual sequences in (31), and let $z_t = (z_{t,i})_{i=1}^N \in \mathbb{R}^D$ for all $t \geq 0$. For the sequence $\{w_t\}$ using (32), it is then straightforward to see that we then exactly recover the global dual sequences of (FTRL), (IFTRL), (OFTRL), and (EGFTRL) from Section 3.

### C.2 Primal-Dual Relationships Between Regularizer and Energy

We now present several key properties of the global regularizer $Q : \mathcal{W} \to \mathbb{R}^D$ and its corresponding energy function (convex conjugate) $H = Q^* : \mathbb{R}^D \to \mathbb{R}$. Recall that $H$ is given by

$$H(z) := Q^*(z) = \max_{w \in \mathcal{W}} \{\langle w, z\rangle - Q(w)\} \quad \text{for all } z \in \mathbb{R}^D\,. \tag{33}$$

For further background and preliminaries on conjugate functions, see, e.g., (Boyd & Vandenberghe, 2004, Section 3.3). Moreover, recall from Section 3.1 that we consider the following assumption:

**Assumption 2** (Relatively Legendre)**.** $Q$ is (i) strictly convex and twice continuously differentiable on $\operatorname{ri}(\mathcal{W})$, and (ii) $\|\nabla Q(w)\| \to \infty$ as $w$ approaches the relative boundary of $\mathcal{W}$.

We remark that Assumption 2 is the standard notion of Legendre-type functions (see, e.g. Rockafellar (1997), Chapter 26), but defined with respect to the relative interior of $\mathcal{W}$ (which is lower-dimensional). Nevertheless, Assumption 2 implies several standard primal-dual relationships between $Q$ and $H$, which we state in the following sequence of propositions. We first make the following technical remark regarding differentiability and our notation:

**Remark on differentiability.** Throughout, derivatives of the regularizer $Q$ are taken only on $\operatorname{ri}(\mathcal{W})$ and relative to its affine hull, whose space of feasible directions is $\mathcal{S}^\perp$. We also assume $Q$ as a proper, lower semicontinuous, convex function on $\mathcal{W}$, allowing the value $+\infty$ at boundary points, and extended to $+\infty$ outside $\mathcal{W}$. The relative steepness condition of Assumption 2 (ii) should thus be interpreted relative to the affine hull.

Moreover, we will also assume throughout that Assumptions 3 and 4 hold for $Q$ and $H$.

#### C.2.1 Relative Duality and Proof of Proposition 3.2.

We start with the following property which gives a useful characterization of $\nabla H$. This should be thought of as the basic property of Legendre-type functions (e.g., as in Rockafellar (1997), Theorem 26.5), but adapted to hold in a relative sense:

**Proposition C.1.** *For every $z \in \mathbb{R}^D$, the maximizer of $H(z)$ in (33) is unique and lies in $\operatorname{ri}(\mathcal{W})$. It holds specifically that $\nabla H(z) = \operatorname{argmax}_{w \in \mathcal{W}}\{\langle w, z\rangle - Q(w)\} \in \operatorname{ri}(\mathcal{W})$.*

**Proof.** Fix $z \in \mathbb{R}^D$. By compactness of $\mathcal{W}$ and lower semicontinuity of $Q$, the maximum in the definition of $H(z)$ from (33) is finite and attained. By the relative steepness property of Assumption 2 (ii), every maximizer lies in $\operatorname{ri}(\mathcal{W})$. By strict convexity of $Q$, the maximizer is also unique. This proves the first claim of the proposition.

For the second claim, let $w \in \text{ri}(\mathcal{W})$ denote the maximizer defining $H(z)$ from (33). Then for every $h \in \mathbb{R}^D$ and $t \in \mathbb{R}$, evaluating the objective in (33) for $H(z+th)$ at the point $w$ gives $H(z+th) \geq H(z) + t\langle w, h\rangle$. Then taking the derivative at $t=0$ from both directions and using the differentiability of $H$ (which holds by Assumption 4) then yields $\langle \nabla H(z), h\rangle = \langle w, h\rangle$ for all $h \in \mathbb{R}^D$. Thus $\nabla H(z) = w$. ■

Note that as a consequence of Proposition C.1, we obtain a short proof of the primal-dual invariants of the global FTRL-type dynamics from Proposition 3.2, restated here:

**Proposition 3.2** (Primal-dual invariant)**.** *Let $\{w_t\}$ be the iterates from (8) using the dual sequence $\{z_t\}$ of (FTRL), (IFTRL), (OFTRL), or (EGFTRL). Then $w_t = \nabla H(z_t) \in \text{ri}(\mathcal{W})$ for all $t \geq 1$.*

**Proof.** Observe by the global primal update rule of (8) that each

$$w_t = \text{argmin}_{w\in\mathcal{W}}\{-\langle w, z_t\rangle + Q(w)\} = \text{argmax}_{w\in\mathcal{W}}\{\langle w, z_t\rangle - Q(w)\} \,.$$

Then by Proposition C.1, we have $w_t = \nabla H(z_t) \in \text{ri}(\mathcal{W})$. ■

### C.2.2 First-Order and Second-Order Primal-Dual Relationships.

We now state several further primal-dual relationships between $Q$ and $H$ that hold under Assumption 2. Note that these relationships can be thought of as the usual consequences of Legendre duality, but working over the quotient space $\mathbb{R}^D/\mathcal{S}$.

**Proposition C.2.** *For $z \in \mathbb{R}^D$, $w \in \text{ri}(\mathcal{W})$: $w = \nabla H(z) \iff z = \nabla Q(w) + s$ for some $s \in \mathcal{S}$.*

**Proof.** By Proposition C.1, $w = \nabla H(z)$ if and only if $w = \text{argmin}_{w'\in\mathcal{W}}\{-\langle w', z\rangle + Q(w)\}$. As $w \in \text{ri}(\mathcal{W})$, first-order optimality conditions yield

$$\langle \nabla Q(w) - z, h\rangle = 0 \quad \text{for all } h \in \mathcal{S}^\perp \,,$$

which is necessary due to $w \in \text{ri}(\mathcal{W})$, and sufficient due to convexity. Moreover, the condition is equivalent to $z - \nabla Q(w) \in \mathcal{S}$, which yields the statement of the proposition. ■

We also have that $\nabla H$ is invariant to shifts under directions in $\mathcal{S}$:

**Proposition C.3.** *For all $z \in \mathbb{R}^D$ and $s \in \mathcal{S}$: $\nabla H(z+s) = \nabla H(z)$.*

**Proof.** Fix any $w' \in \mathcal{W}$ and $s \in \mathcal{S}$. For any $w \in \mathcal{W}$, recall that $w - w' \in \mathcal{S}^\perp$. Thus we have that $\langle w - w', s\rangle = 0$, which implies that $\langle w, s\rangle = \langle w', s\rangle$ for all $w \in \mathcal{W}$. Then using the definition of $H$ from (33), we have that

$$\begin{aligned} H(z+s) &= \max_{w\in\mathcal{W}}\{\langle w, z+s\rangle - Q(w)\} \\ &= \max_{w\in\mathcal{W}}\{\langle w, z\rangle - Q(w) + \langle w', s\rangle\} = H(z) + \langle w', s\rangle \,. \end{aligned}$$

Thus the maximizing $w \in \mathcal{W}$ for $H(z+s)$ and $H(z)$ is identical. By Proposition C.1, this gives $\nabla H(z+s) = \nabla H(z)$. ■

As a direct corollary of Proposition C.3, we also have the following property:

**Proposition C.4.** *For all $z \in \mathbb{R}^D$ and any $u \in \mathbb{R}^D$: $\nabla H(z+v) = \nabla H(z + \Pi_{\mathcal{S}^\perp} v)$.*

Next, we establish the following second-order primal-dual relationships:

**Proposition C.5.** *Then for all $z \in \mathbb{R}^D$, and $w = \nabla H(z) \in \text{ri}(\mathcal{W})$:*

*(i)* $\nabla^2 H(z+s) = \nabla^2 H(z)$ *for all* $s \in \mathcal{S}$.

*(ii)* $\nabla^2 H(z)\nabla^2 Q(w)u = u$ *for all* $u \in \mathcal{S}^\perp$.

*(iii)* $\text{Null}(\nabla^2 H(z)) = \mathcal{S}$ *and* $\text{Range}(\nabla^2 H(z)) = \mathcal{S}^\perp$.

**Proof.** For (i), fix $s \in \mathcal{S}$. Then differentiating the identity $\nabla H(z+s) = \nabla H(z)$ with respect to $z$ gives $\nabla^2 H(z+s) = \nabla^2 H(z)$.

For (ii), note by the first-order property of Proposition C.2 that $\nabla H(\nabla Q(v)) = v$ for any $v \in \text{ri}(\mathcal{W})$. Then differentiating this identity at $v = w$ along any direction $u \in \mathcal{S}^\perp$ then yields by the chain rule

$$\nabla^2 H(\nabla Q(w))\nabla^2 Q(w)u = u \,. \tag{34}$$

As $z - \nabla Q(w) \in \mathcal{S}$ by Proposition C.2, we then have by part (i) that $\nabla^2 H(\nabla Q(w)) = \nabla^2 H(z)$, which by substituting into (34) yields claim (ii).

For claim (iii), let $s \in \mathcal{S}$. Then we differentiate $\nabla H(z + ts) - \nabla H(z)$ at $t = 0$, which yields $\nabla^2 H(z)s = 0$. By symmetry of $\nabla^2 H(z)$, this implies that $\nabla^2 H(z)v \in \mathcal{S}^\perp$ for all $v \in \mathbb{R}^D$. Also, part (ii) establishes $\mathcal{S}^\perp \subseteq \text{Range}(\nabla^2 H(z))$. Thus together we have $\text{Range}(\nabla^2 H(z)) = \mathcal{S}^\perp$. Again by symmetry of $\nabla^2 H(z)$, it follows that

$$\text{Null}(\nabla^2 H(z)) = \text{Range}(\nabla^2 H(z))^\perp = \mathcal{S} \, ,$$

which concludes the proof. ■

### C.2.3 Properties of Bregman Divergences

Recall that the Bregman divergence $\mathsf{D}_Q : \mathcal{W} \times \text{ri}(\mathcal{W}) \to \mathbb{R}$ is given by

$$\mathsf{D}_Q(\widetilde{w}, w) = Q(\widetilde{w}) - Q(w) - \langle \nabla Q(w), \widetilde{w} - w \rangle \, . \tag{35}$$

Here, note that $\mathsf{D}_Q(\widetilde{w}, w) = +\infty$ whenever $Q(\widetilde{w}) = +\infty$.

Then we further have the following useful primal-dual characterization of Bregman divergence for $Q$ satisfying Assumption 2:

**Proposition C.6.** *Fix $z, z' \in \mathbb{R}^D$, and let $w = \nabla H(z), w' = \nabla H(z') \in \text{ri}(\mathcal{W})$. Then for $\widetilde{w} \in \mathcal{W}$ with $Q(\widetilde{w}) < \infty$, the following properties hold:*

*(i)* $\mathsf{D}_Q(\widetilde{w}, w) = Q(\widetilde{w}) + H(z) - \langle z, \widetilde{w} \rangle$

*(ii)* $\mathsf{D}_Q(\widetilde{w}, w') - \mathsf{D}_Q(\widetilde{w}, w) = H(z') - H(z) - \langle \widetilde{w}, z' - z \rangle$.

**Proof.** For the first claim (i), recall by the Fenchel-Young equality that $Q(w) + H(z) = \langle z, w \rangle$, which holds due to $w = \nabla H(z)$ (see, e.g., (Shalev-Shwartz, 2012, Section 2.7)). Then substituting this into (35) and using the fact that $\nabla Q(w) = z + s$ for some $s \in \mathcal{S}$ from Proposition C.2:

$$\begin{aligned} \mathsf{D}_Q(\widetilde{w}, w) &= Q(\widetilde{w}) - Q(w) - \langle \nabla Q(w), \widetilde{w} - w \rangle \\ &= Q(\widetilde{w}) + H(z) - \langle z, w \rangle - \langle z + s, \widetilde{w} - w \rangle \\ &= Q(\widetilde{w}) + H(z) - \langle z, \widetilde{w} \rangle \end{aligned}$$

Here, we used the fact that $\widetilde{w} - w \in \mathcal{S}^\perp$ and thus $\langle s, \widetilde{w} - w \rangle = 0$ for $s \in \mathcal{S}$. This yields claim (i).

For claim (ii), apply claim (i) with the pairs $(w, z)$ and $(w', z')$, and take their difference. ■

Finally, we establish the following boundary divergence property for the Bregman divergence $\mathsf{D}_Q$:

**Proposition C.7** (Boundary blowup of Bregman divergence)**.** *Fix $\widetilde{w} \in \text{ri}(\mathcal{W})$. Then $\mathsf{D}_Q(\widetilde{w}, w) \to \infty$ as $w \in \text{ri}(\mathcal{W}) \to \text{rbd}(\mathcal{W})$. Then for finite $\alpha \geq 0$, $\mathcal{C} = \{w \in \text{ri}(\mathcal{W}) : \mathsf{D}_Q(\widetilde{w}, w) \leq \alpha\} \subset \text{ri}(\mathcal{W})$ is a compact subset of $\text{ri}(\mathcal{W})$.*

**Proof.** We choose some $r > 0$ such that the relative closed Euclidean ball $\mathcal{B} = \{\widetilde{w} + h : h \in \mathcal{S}^\perp, \ \|h\|_2 \leq r\}$ lies in $\text{ri}(\mathcal{W})$. By continuity, $Q$ is bounded above on $\mathcal{B}$, and we let $c > 0$ denote such an upper bound. Now for $w \in \text{ri}(\mathcal{W})$, let $z = \nabla Q(w) \in \mathcal{S}^\perp$. By Proposition C.2, we then have $w = \nabla H(z)$. If $z \neq 0$, then evaluating the maximum defining $H(z)$ from (33) at the point $\widetilde{w} + rz/\|z\|_2$ yields $H(z) \geq \langle \widetilde{w}, z \rangle + r\|z\|_2 - c$. Evaluating the maximum defining $H(0)$, evaluated at $\widetilde{w}$, this same inequality also holds for $z = 0$. Then by Proposition C.6, part (i):

$$\mathsf{D}_Q(\widetilde{w}, w) \geq Q(\widetilde{w}) + r\|\nabla Q(w)\|_2 - c. \tag{36}$$

By the relative steepness condition of Assumption 2, part (ii), it follows that the right hand side of (36) tends to $\infty$ as $w \to \text{rbd}(\mathcal{W})$. Thus, as $\mathsf{D}_Q(\widetilde{w}, \cdot)$ is continuous on $\text{ri}(\mathcal{W})$, this boundary blowup prevents the finite sublevelset $\mathcal{C}$ from accumulating at the relative boundary. Thus, $\mathcal{C}$ is closed in $\mathcal{W}$, and compactness of $\mathcal{W}$ then implies compactness of $\mathcal{C}$. ■

## C.3 Examples of Regularizers

Here, we give details on instantiations of regularizers $Q : \mathcal{W} \to \mathbb{R}^D$ that satisfy Assumptions 2, 3, and 4. In particular, we very the properties for the canonical examples of negative Shannon

entropy (C.3.1), negative Tsallis entropy (C.3.2) of order $q \in (0,1)$, and for the log-barrier regularizer (C.3.3). Throughout, we verify the properties of $Q$ and $H$ at a single component, which suffices due to separability. For readability, we will write (by slight abuse of notation) $Q$ to denote a single component $Q_i$, and $H$ for a single component $H_i$, and we will assume that $\mathcal{W} = \Delta_D$.

### C.3.1 Negative Shannon Entropy

For the negative Shannon entropy regularizer, we have that

$$Q(w) = \textstyle\sum_{i=1}^D w_i \log w_i \quad \text{for all } w \in \mathcal{W}$$

It is well known that the corresponding $H = Q^* : \mathbb{R}^D \to \mathbb{R}$ and $\nabla H : \mathbb{R}^D \to \mathrm{ri}(\mathcal{W})$ are given by

$$H(z) = \log\left(\textstyle\sum_{i=1}^D \exp(z_i)\right), \quad \nabla H(z) = \mathrm{softmax}(z) = \frac{\exp(z_i)}{\sum_{j=1}^D \exp(z_j)} \quad \text{for all } z \in \mathbb{R}^D.$$

**Relative Legendre and Strong Convexity.** It is well-known for negative Shannon entropy that $Q$ satisfies Assumption 2 (see, e.g., (Xiao, 2022, Section 4.1)). Moreover, by Pinsker's inequality, we have that $Q$ is 1-strongly-convex with respect to the $\ell_1$ norm $\|\cdot\|_1$, and thus satisfies Assumption 3.

**Local hessian stability.** The energy $H$ additionally satisfies the (LHS) property of Assumption 4 with $\omega = 2$, as proven in (Lazarsfeld et al., 2026, Proposition 4.3).

### C.3.2 Negative Tsallis Entropy

Fix $q \in (0,1)$. Then the (normalized) negative $q$-Tsallis entropy is given by

$$Q_q(w) = \frac{1 - \sum_{j=1}^d w_j^q}{q(1-q)} \quad \text{for all } w \in \mathcal{W},$$

with $0^q = 0$. The corresponding energy $H_q = Q_q^*$ and gradient map are given by

$$H_q(z) = \lambda_q(z) + \tfrac{1}{q}\textstyle\sum_{j=1}^d w_j(z)^q - \frac{1}{q(1-q)}, \quad (\nabla H_q(z))_j = w_j(z) = \big((1-q)(\lambda_q(z) - z_j)\big)^{-1/(1-q)},$$

where $\lambda_q(z) > \max_j z_j$ is uniquely determined by $\sum_j w_j(z) = 1$. These expressions follow directly from the optimality conditions defining $H_q$. The implicit function theorem implies that $\lambda_q$ and $\nabla H_q$ are smooth on $\mathbb{R}^d$.

**Relatively Legendre.** On $\mathrm{ri}(\mathcal{W})$, note that $Q$ smooth, with

$$\partial_j Q_q(w) = -\frac{w_j^{q-1}}{1-q}, \qquad \nabla^2 Q_q(w) = \mathrm{Diag}(w_j^{q-2}).$$

The Hessian is positive definite, so $Q_q$ is strictly convex. As $w$ approaches the relative boundary, a smallest coordinate tends to zero, and its gradient component tends to $-\infty$. Meanwhile, some coordinate is at least $1/d$, so its gradient component remains bounded. Thus the projection of the gradient onto the simplex tangent space diverges, establishing Assumption 2.

**Strong Convexity.** Since $q < 1$ and $w_j \le 1$, we have by Cauchy-Schwarz

$$\langle v, \nabla^2 Q_q(w) v\rangle = \textstyle\sum_{j=1}^d w_j^{q-2} v_j^2 \ge \sum_{j=1}^d \frac{v_j^2}{w_j} \ge \|v\|_1^2 \qquad \text{for all } w \in \mathrm{ri}(\mathcal{W}) \text{ and } v \in \mathbb{R}^d.$$

Integrating along line segments and using continuity of $Q_q$ on $\mathcal{W}$ yields $D_{Q_q}(w', w) \ge \frac{1}{2}\|w' - w\|_1^2$ for every $w' \in \mathcal{W}$ and $w \in \mathrm{ri}(\mathcal{W})$. Thus, $Q_q$ is 1-strongly convex with respect to $\|\cdot\|_1$ and satisfies Assumption 3.

**Local Hessian Stability.** The energy $H_q$ satisfies the (LHS) property of Assumption 4 with parameter $\omega_q = 2(2-q)$. To see this, set $a_j(z) = w_j(z)^{2-q}$ and $A(z) = \sum_j a_j(z)$. Differentiating the optimality conditions gives

$$\nabla^2 H_q(z) = \mathrm{Diag}(a(z)) - \frac{a(z)a(z)^\top}{A(z)}.$$

Fix $h = z' - z$ and $v \in \mathbb{R}^d$. Along $z_s = z + sh$, write $w_j = w_j(z_s)$, $a_j = a_j(z_s)$, and let $\bar{h}_s = \sum_j a_j h_j / A$ and $\bar{v}_s = \sum_j a_j v_j / A$. Then $\dot{w}_j = a_j(h_j - \bar{h}_s)$, and $f_v(s) = \langle v, \nabla^2 H_q(z_s) v\rangle$ satisfies

$$f_v(s) = \textstyle\sum_j a_j (v_j - \bar{v}_s)^2, \qquad f_v'(s) = (2-q)\sum_j a_j w_j^{1-q}(h_j - \bar{h}_s)(v_j - \bar{v}_s)^2.$$

Since $w_j^{1-q} \le 1$ and $|h_j - \bar{h}_s| \le 2\|h\|_\infty$, we obtain $|f_v'(s)| \le \omega_q \|h\|_\infty f_v(s)$. Applying Grönwall's inequality and using the fact that $v$ was arbitrary then yields

$$\exp(-\omega_q\|z-z'\|_\infty)\cdot\nabla^2 H_q(z') \preceq \nabla^2 H_q(z) \preceq \exp(\omega_q\|z-z'\|_\infty)\cdot\nabla^2 H_q(z')\ ,$$

where $\omega_q = 2(2-q) \ge 1$.

### C.3.3 Log-Barrier Regularizer

Let $\mathcal{W} = \Delta_d$. The log-barrier regularizer is given by

$$Q_{\mathrm{lb}}(w) = -\textstyle\sum_{j=1}^d \log w_j \quad \text{for all } w \in \mathrm{ri}(\mathcal{W}),$$

and $Q_{\mathrm{lb}}(w) = +\infty$ on the relative boundary. Here, the regularizer assumptions are understood with this extended-valued convention. The corresponding energy $H_{\mathrm{lb}} = Q_{\mathrm{lb}}^*$ and gradient map are:

$$H_{\mathrm{lb}}(z) = \lambda(z) - d - \textstyle\sum_{j=1}^d \log(\lambda(z) - z_j), \quad (\nabla H_{\mathrm{lb}}(z))_j = w_j(z) = \frac{1}{\lambda(z)-z_j},$$

where $\lambda(z) > \max_j z_j$ is uniquely determined by $\sum_j w_j(z) = 1$. These expressions follow directly from the optimality conditions defining $H_{\mathrm{lb}}$. The implicit function theorem implies that $\lambda$ and $\nabla H_{\mathrm{lb}}$ are smooth on $\mathbb{R}^d$.

**Relatively Legendre.** On $\mathrm{ri}(\mathcal{W})$, the regularizer is smooth, with

$$\partial_j Q_{\mathrm{lb}}(w) = -\tfrac{1}{w_j}, \qquad \nabla^2 Q_{\mathrm{lb}}(w) = \mathrm{Diag}(w_j^{-2}).$$

The Hessian is positive definite, so $Q_{\mathrm{lb}}$ is strictly convex. As $w$ approaches the relative boundary, a smallest coordinate tends to zero, and its gradient component tends to $-\infty$. Meanwhile, some coordinate is at least $1/d$, so its gradient component remains bounded. Thus the projection of the gradient onto the simplex tangent space diverges, establishing Assumption 2 under the extended-valued convention.

**Strong Convexity.** Since $w_j \le 1$, we have by Cauchy–Schwarz

$$\langle v, \nabla^2 Q_{\mathrm{lb}}(w) v\rangle = \textstyle\sum_{j=1}^d \frac{v_j^2}{w_j^2} \ge \sum_{j=1}^d \frac{v_j^2}{w_j} \ge \|v\|_1^2 \qquad \text{for all } w \in \mathrm{ri}(\mathcal{W}) \text{ and } v \in \mathbb{R}^d.$$

Integrating along line segments yields $D_{Q_{\mathrm{lb}}}(w', w) \ge \frac{1}{2}\|w'-w\|_1^2$ for every $w, w' \in \mathrm{ri}(\mathcal{W})$. For $w'$ on the relative boundary, the inequality holds in the extended-valued sense because $D_{Q_{\mathrm{lb}}}(w', w) = +\infty$. Thus, $Q_{\mathrm{lb}}$ is 1-strongly convex with respect to $\|\cdot\|_1$ and satisfies Assumption 3.

**Local Hessian Stability.** The energy $H_{\mathrm{lb}}$ satisfies the (LHS) property of Assumption 4 with parameter $\omega_{\mathrm{lb}} = 4$. To see this, set $a_j(z) = w_j(z)^2$ and $A(z) = \sum_j a_j(z)$. Differentiating the optimality conditions gives

$$\nabla^2 H_{\mathrm{lb}}(z) = \mathrm{Diag}(a(z)) - \tfrac{a(z)a(z)^\top}{A(z)}.$$

Fix $h = z' - z$ and $v \in \mathbb{R}^d$. Along $z_s = z + sh$, $s \in [0,1]$, write $w_j = w_j(z_s)$, $a_j = a_j(z_s)$, $A = A(z_s)$, and let $\bar{h}_s = \sum_j a_j h_j / A$ and $\bar{v}_s = \sum_j a_j v_j / A$. Then $\dot{w}_j = a_j(h_j - \bar{h}_s)$, and $f_v(s) = \langle v, \nabla^2 H_{\mathrm{lb}}(z_s) v\rangle$ satisfies

$$f_v(s) = \textstyle\sum_j a_j (v_j - \bar{v}_s)^2, \qquad f_v'(s) = 2\sum_j a_j w_j (h_j - \bar{h}_s)(v_j - \bar{v}_s)^2.$$

Since $w_j \le 1$ and $|h_j - \bar{h}_s| \le 2\|h\|_\infty$, we obtain $|f_v'(s)| \le 4\|h\|_\infty f_v(s)$. Applying Grönwall's inequality and using the fact that $v$ was arbitrary then yields

$$\exp(-\omega_{\mathrm{lb}}\|z-z'\|_\infty)\cdot\nabla^2 H_{\mathrm{lb}}(z') \preceq \nabla^2 H_{\mathrm{lb}}(z) \preceq \exp(\omega_{\mathrm{lb}}\|z-z'\|_\infty)\cdot\nabla^2 H_{\mathrm{lb}}(z')\ ,$$

where $\omega_{\mathrm{lb}} = 4 \ge 1$.

## D Details on Reduction to Symmetrized System

This section provides more details on the $M$-symmetrized system that was introduced in Section 5, and that is studied to prove of our main results in Theorem 4.1 Theorem 4.5. In particular, the proofs of Proposition 5.1 and 5.2 are given below in Sections D.1 and D.2, respectively. We further summarize all properties of the symmetrized system in Setting D.1, which will referenced in the proofs of our later results.

**Overview of underlying setting.** Throughout this section, we assume an operator $G : \mathcal{W} \to \mathbb{R}^D$ such that $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weights matrix $M = \mathrm{Diag}(m_1 I_{d_1}, \ldots, m_N I_{d_N}) \in \mathbb{R}^{D\times D}$, where each $m_i > 0$. We assume that $G$ satisfies Assumption 1 with Lipschitz constant $L \geq 1$ and boundedness constant $B \geq 1$. We also assume that $Q : \mathcal{W} \to \mathbb{R}$ is a regularizer satisfying Assumption 2, and its corresponding energy funciton $H = Q^* : \mathbb{R}^D \to \mathbb{R}$ satisfies Assumption 4. We further assume that Assumption 5 holds.

### D.1 Structural Components and Properties

**Scaled Primal and Dual Spaces.** Recall we define the scaled primal domain $\mathcal{X} := M^{1/2}\mathcal{W} \subseteq \mathbb{R}^D$. For each $w \in \mathcal{W}$ and $z \in \mathbb{R}^D$, we associate $x = M^{1/2}w$ and $y = M^{1/2}z$, from which it follows that $w = M^{-1/2}x$ and $z = M^{-1/2}y$. Recalling that $\mathcal{W} = \prod_{i=1}^N \mathcal{W}_i$, where each $\mathcal{W}_i = \Delta_{d_i}$, observe then that $\mathcal{X} = \prod_{i=1}^N \mathcal{X}_i$ where each $\mathcal{X}_i = \sqrt{m_i}\Delta_{d_i}$. Thus, for $\mathcal{S} = \prod_{i=1}^N \mathrm{Span}(\mathbf{1}_{d_i}) \subseteq \mathbb{R}^D$ and its orthogonal complement $\mathcal{S}^\perp \subseteq \mathbb{R}^D$ from (9), we have the following invariant, the proof of which is identical to that of Lemma B.1:

**Lemma D.1.** *Let $x, x' \in \mathcal{X}$. Then $x - x' \in \mathcal{S}^\perp$.*

In other words, as each component of $\mathcal{X}_i$ is a positive constant scaling of $\mathcal{W}_i$, $\mathcal{S}$ and $\mathcal{S}^\perp$ remain the normal and tangent spaces of $\mathcal{X}$, respectively.

**Symmetrized operator.** Recall that the $M$-symmetrized operator $V : \mathcal{X} \to \mathbb{R}^D$ is given by

$$V(x) = M^{1/2}G(M^{-1/2}x) \quad \text{for all } x \in \mathcal{X} \,. \tag{37}$$

By slight abuse of notation, we let $\mathrm{NE}(V) \subset \mathcal{X}$ denote the set of solutions $x' \in \mathcal{X}$ satisfying

$$\langle V(x'), x - x' \rangle \geq 0 \quad \text{for all } x \in \mathcal{X} \,. \tag{NE}$$

As introduced in Section 5, the key structural property of $V$ is that, for all solutions $w^\star \in \mathrm{WMEQ}_M(G)$, the scaled variable $x^\star = M^{1/2}w^\star \in \mathcal{X}$ satisfies the *unweighted* Minty variational equality with respect to $V$:

**Proposition 5.1** (Unweighted Minty equality)**.** *Let $V : \mathcal{X} \to \mathbb{R}^D$ be the operator from (12). For any solution $w^\star \in \mathrm{WMEQ}_M(G)$, let $x^\star := M^{1/2}w^\star \in \mathcal{X}$. Then it holds that:*

$$\langle V(x), x - x^\star \rangle = 0 \ \textit{for all } x \in \mathcal{X} \,. \tag{MEQ}$$

**Proof.** Fix $w^\star \in \mathrm{WMEQ}_M(G)$ and $w \in \mathcal{W}$. Let $x^\star = M^{1/2}w^\star \in \mathcal{X}$ and $x = M^{1/2}w \in \mathcal{X}$, which means $w^\star = M^{-1/2}x^\star$ and $w = M^{-1/2}x$. As $w^\star$ satisfies (6), we then have

$$\begin{aligned} 0 = \langle MG(w), w - w^\star \rangle &= \langle MG(M^{-1/2}x), M^{-1/2}(x - x^\star) \rangle \\ &= \langle M^{1/2}G(M^{-1/2}x), (x - x^\star) \rangle = \langle V(x), x - x^\star \rangle \,. \end{aligned}$$

Here, the third equality follows by symmetry of $M = M^\top$, and the final equality by definition of $V$. Since the equality holds for all $w \in \mathcal{W}$, it also holds for all $x = M^{1/2}w \in \mathcal{X}$. ■

For convenience recall that we write

$$\mathrm{MEQ}(V) = \{x^\star \in \mathcal{X} : \langle V(x), x - x^\star \rangle = 0 \text{ for all } x \in \mathcal{X}\} \subseteq \mathcal{X} \tag{38}$$

to denote the set of all solutions $x^\star$ of (MEQ). Then by reversing the calculation in Proposition 5.1 and also using the product structure of $\mathcal{X}$ and Lemma B.5, we have as an immediate corollary the following relationships:

**Corollary D.2.** *The following hold: (i) $w^\star \in \mathrm{WMEQ}_M(G)$ if and only if $M^{1/2}w^\star \in \mathrm{MEQ}(V)$ and (ii) $w^\star \in \mathrm{NE}(G)$ if and only if $M^{1/2}w^\star \in \mathrm{NE}(V)$.*

**Lipschitzness and boundedness of V.** Under Assumption 1 fo $G$, it follows that $V$ is Lipschitz and bounded, with parameters depending on $L, B$ and $m_{\max}, m_{\min}$. For this, recall that we let $\kappa_M := m_{\max}/m_{\min}$. Then we have the following properties of $V$:

**Proposition D.3.** *The operator $V$ is $\widehat{L}$-Lipschitz and $\widehat{B}$ bounded, for $\widehat{L} = \sqrt{\kappa_M}L$ and $\widehat{B} = \sqrt{m_{\max}}B$.*

**Proof.** First, using the block-diagonal structure of $M = \mathrm{Diag}(m_1 I_{d_1}, \dots, m_N I_{d_N})$, observe that $\|M^{1/2}v\|_* \leq \sqrt{m_{\max}}\|v\|_*$ and $\|M^{-1/2}v\| \leq \frac{1}{\sqrt{m_{\min}}}\|v\|$ for all $v \in \mathbb{R}^D$. Then by definition of $V(x) = M^{1/2}G(M^{-1/2}x)$ and using the $L$-Lipschitzness of $G$ from Assumption 1, we have for all $x, x' \in \mathcal{X}$:

$$\|V(x') - V(x)\|_* = \|M^{1/2}(G(M^{-1/2}x') - G(M^{-1/2}x))\| \leq \frac{\sqrt{m_{\max}}}{\sqrt{m_{\min}}} L\|x - x'\| \,.$$

Similarly, by $B$-boundedness of $G$, we have for all $x \in \mathcal{X}$

$$\|V(x)\|_* = \|M^{1/2}G(M^{-1/2}x)\|_* \leq \sqrt{m_{\max}} B \,,$$

which concludes the proof. ■

### D.2 Algorithmic Components and Properties

We now expand on the algorithmic components of the symmetrized system. In particular, we develop the proof of Proposition 5.2, restated here:

**Proposition 5.2** (Symmetrized dynamics)**.** *Let $V : \mathcal{X} \to \mathbb{R}^D$ be the operator from (12). Then there exist a global regularizer $R : \mathcal{X} \to \mathbb{R}$ and energy function $F = R^*$ such that, if $\{(w_t, z_t)\}$ is an (OFTRL) or (EGFTRL) trajectory for $(G, H, \eta)$, then letting $x_t = M^{1/2}w_t$ and $y_t = M^{1/2}z_t$ for all $t \geq 0$, it follows that $\{(x_t, y_t)\}$ is an (OFTRL) or (EGFTRL) trajectory for $(V, F, \eta)$.*

We will prove the proposition for the regularizer $R : \mathcal{X} \to \mathbb{R}$ defined as follows: first, recall from Section 4 that we defined the $M$-symmetrized regularizer $Q_M : \mathcal{W} \to \mathbb{R}$ by

$$Q_M(w) = \textstyle\sum_{i=1}^N m_i \cdot \left(\frac{\eta}{\eta_i} Q_i(w_i)\right) \quad \text{for all } w \in \mathcal{W} \,. \tag{39}$$

Then $R$ is the separable regularizer given by

$$R(x) := Q_M(M^{-1/2}x) = \textstyle\sum_{i=1}^N m_i \cdot \left(\frac{\eta}{\eta_i} Q_i(\frac{x_i}{\sqrt{m_I}})\right) \text{ for all } x \in \mathcal{X} \,. \tag{40}$$

Recall that $F = R^* : \mathbb{R}^D \to \mathbb{R}$ is then the convex conjugate of $R$. Importantly, recall that $Q$ satisfies assumption 2 with respect to $\mathcal{W}$. Then due to the positive weights in $M$ and the invertibility of the change of coordinates defining $\mathcal{X} = M^{1/2}\mathcal{W}$ it follows that $R$ is twice continuously differentiable on $\mathrm{ri}(\mathcal{X})$, and that the relative steepness of $\nabla R$ also holds with respect to $\mathrm{rbd}(\mathcal{X})$. Thus $R$ satisfies the properties of Assumption 2 with respect to $\mathcal{X}$, and the primal-dual relationships of Section C.2 then identically apply to the pair $R$ and $F$.

We then proceed with the proof of proposition 5.2.

**Proof.** The proof proceeds in two steps: first, we establish a relationship between the original and symmetrized regularizer and energy gradients. Then, we apply the relationship to the iterates of OFTRL and EGFTRL under the original system, and we show that the update structure is exactly preserved for the symmetrized system.

*1. Relationship between gradient maps.* First, note that by separability of $Q_M$ (and using the definition of $Q$ from 7), we have that $\nabla Q_M(w) = M\nabla Q(w)$ for all $w \in \mathrm{ri}(\mathcal{W})$. Then for $x \in \mathrm{ri}(\mathcal{X})$, let $w = M^{-1/2}x \in \mathcal{W}$. Using the fact that $R(x) = \mathrm{Q}_M(M^{-1/2}x)$, it then follows by the chain rule and the symmetry of $M$ that

$$\nabla R(x) = (M^{-1/2})^\top \nabla Q_M(M^{-1/2}x) = M^{-1/2}M\nabla Q(w) = M^{1/2}\nabla Q(w) \,. \tag{41}$$

Now, for fixed $y \in \mathbb{R}^D$, set $z = M^{-1/2}y \in \mathbb{R}^D$, Then let $w = \nabla H(z) \in \mathrm{ri}(\mathcal{W})$, and $x = M^{1/2}w \in \mathrm{ri}(\mathcal{X})$. We will show that $x = \nabla F(y)$. For this, first recall that due to Proposition C.2 applied to the pair $(w, z)$, we have that $z = \nabla Q(w) + s$ for some $s \in \mathcal{S}$. It follows using (41) that

$$M^{1/2}z = M^{1/2}\nabla Q(w) + M^{1/2}s \implies y = \nabla R(x) + s' \tag{42}$$

where $s' = M^{1/2}s \in \mathcal{S}$. However, by Proposition C.2 applied with $R$ and $F$, this implies the identity $x = \nabla F(y) \in \mathrm{ri}(\mathcal{X})$. Thus for every $z \in \mathbb{R}^D$ and $w = \nabla H(z) \in \mathrm{ri}(\mathcal{W})$, it holds for $y = M^{1/2}z \in \mathbb{R}^D$ and $x = M^{1/2}w \in \mathrm{ri}(\mathcal{X})$ that

$$\nabla F(y) = x = M^{1/2}w = M^{1/2}\nabla H(M^{-1/2}y) \,. \tag{43}$$

*2. Preservation of FTRL-type update structure.* Using the relationship of (43), we then have the following equivalence in the primal and dual update structure of FTRL-type dynamics for the symmetrized system. For generality, consider $z', z \in \mathbb{R}^D$ and $w' = \nabla H(z') \in \mathrm{ri}(\mathcal{W})$ and $w = \nabla H(z) \in \mathrm{ri}(\mathcal{W})$. Let $y' = M^{1/2}z'$ and $y = M^{1/2}z$, and let $x' = M^{1/2}w' \in \mathrm{ri}(\mathcal{X})$ and $x = M^{1/2}w \in \mathrm{ri}(\mathcal{X})$. Suppose further that for some $\eta > 0$, we have $z' = z - \eta G(w)$. Then multiplying by $M^{1/2}$, it follows that:

$$y' = M^{1/2}z' = M^{1/2}z - \eta M^{1/2}G(M^{-1/2}x) = y - \eta V(x)\,, \tag{44}$$

where we use the definition of $V$ from (12). As $z'$ and $w'$ satisfy $w' = \nabla H(z')$, then applying (43) to $y'$, we further obtain $x' = \nabla F(y') \in \mathrm{ri}(\mathcal{X})$.

For the iterates $\{(w_t, z_t)\}$ of (OFTRL) and (EGFTRL) under the original system, this immediately implies that, for $x_t = M^{1/2}w_t$ and $y_t = M^{1/2}z_t$, we have the following update structure in the symmetrized coordinates for all $t \geq 0$:

$$\begin{aligned} &\text{for OFTRL}: \quad y_{t+1} = y_t - 2\eta V(x_t) + \eta V(x_{t-1}) \quad x_{t+1} = \nabla F(y_{t+1}) \\ &\text{for EGFTRL}: \quad \begin{cases} y_{t+1/2} = y_t - \eta V(x_t) & , x_{t+1/2} = \nabla F(y_{t+1/2}) \\ y_{t+1} = y_t - \eta V(x_{t+1/2}) & , x_{t+1} = \nabla F(y_{t+1})\,. \end{cases} \end{aligned} \tag{45}$$

This completes the proof of the proposition. ■

**Local Hessian stability.** As the original energy function $H = Q^*$ satisfies the Assumption 4 with constant $\omega > 0$, we further have for the symmetrized energy function a similar property:

**Proposition D.4** (LHS property for $F$)**.** *For every $y, y' \in \mathbb{R}^D$ the energy hessians satisfy*

$$\exp(-\widehat{\omega}\|y - y'\|_*) \cdot \nabla^2 F(y') \preceq \nabla^2 F(y) \preceq \exp(\widehat{\omega}\|y - y'\|_*) \cdot \nabla^2 F(y')\,. \tag{LHS}$$

*where $\widehat{\omega} = \omega/\sqrt{m_{\min}}$.*

**Proof.** First, recall from (43) that for $y \in \mathbb{R}^D$ and $z = M^{-1/2}y$, we have the relationship $\nabla F(y) = M^{1/2}\nabla H(M^{-1/2}y)$. Differentiating then yields

$$\nabla^2 F(y) = M^{1/2}\nabla^2 H(M^{-1/2}y)M^{-1/2} = \nabla^2 H(M^{-1/2y})\,.$$

Here, the final equality holds due to the separability of $H = Q^*$, which means that the the block diagonal Hessian $\nabla^2 H$ comutes with the block-diagonal matrix $M$. Now fix also $y' \in \mathbb{R}^D$ and let $z' = M^{-1/2}y'$. By applying Assumption 4 for $H$ and using the expression above, we then have for a constant $\widehat{\omega} \geq 1$ that

$$\exp(-\omega\|z - z'\|_*) \cdot \nabla^2 H(z') \preceq \nabla^2 H(z) \preceq \exp(\omega\|z - z'\|_*) \cdot \nabla^2 H(z')\,. \tag{46}$$

Moreover, observe that $\omega\|z - z'\|_* = \omega\|M^{-1/2}(y - y')\|_*$. Substituting this inequality into (46) then yields the statement of the proposition. ■

**Proof of Proposition 5.3** For the energy function $F = R^*$, recall that we define for any $x^\star \in \mathrm{ri}(\mathcal{X})$ the *shifted energy function* $F_{x^\star} : \mathbb{R}^D \to \mathbb{R}$ by

$$F_{x^\star}(y) = F(y) - \langle x^\star, y\rangle \quad \text{for all } y \in \mathbb{R}^D\,. \tag{47}$$

We then prove the following relationship between the Bregman divergence $\mathsf{D}_R$ and $F_{x^\star}$, as introduced in Section 5.1:

**Proposition 5.3** (Change in shifted energy and Bregman divergence)**.** *Let $y, y' \in \mathbb{R}^D$, and let $x = \nabla F(y), x' = \nabla F(y') \in \mathrm{ri}(\mathcal{X})$. Then for any $x^\star \in \mathrm{ri}(\mathcal{X})$, it holds that:*

$$\mathsf{D}_R(x^\star, x') - \mathsf{D}_R(x^\star, x) \;=\; F_{x^\star}(y') - F_{x^\star}(y)\,.$$

**Proof.** As $F$ and $R$ satisfy Assumption 2, apply Proposition C.6 (ii) for the pairs $(x', y')$ and $(x, y)$, and use the definition of $F_{x^\star}(\cdot)$ from (47). ■

**Regularity condition under symmetrized system.** Finally, when Assumption 5 holds for the original system using $G$, $Q$, and $H$, then the analogous condition also holds for the symmetrized system:

**Proposition D.5** (Regularity under symmetrized system.)**.** *Fix $z_0 \in \mathbb{R}^D$ and let $w_0 = \nabla H(z_0)$, and let $y_0 = M^{1/2} z_0$ and $x_0 = M^{1/2} w_0$ If condition of Assumption 5 holds for the original system with $(G, Q, H)$ at $w_0$, then it holds for the symmetrized system at $x_0$ with $(V, R, F)$*

**Proof.** First, we extend the definitions of the effective dual space and the map $\pi_0$ introduced in 3.2 to the symmetrized system. In particular, to distinguish between the original and symmetrized systems, we define $\mathcal{U}_0^G := \Pi_{\mathcal{S}^\perp} \mathrm{Span}(G(\mathcal{W}))$ and $\mathcal{U}_0^V := \Pi_{\mathcal{S}^\perp} \mathrm{Span}(V(\mathcal{X}))$. Since we have $V(M^{1/2} w) = M^{1/2} G(w)$, and because $M^{1/2}$ commutes with the projection $\Pi_{\mathcal{S}^\perp}$, we have that $\mathcal{U}_V = M^{1/2} \mathcal{U}_G$. Thus $M^{1/2}$ restricts to a linear isomorphism from $\mathcal{U}_G$ onto $\mathcal{U}_V$.

Now, we further distinguish between the two respective restricted stationarity maps, where we let $\pi_0^G(v) := \Pi_{\mathcal{S}^\perp} G(\nabla H(z_0 + v))$ for $v \in \mathcal{U}_G$ and $\pi_0^V(u) := \Pi_{\mathcal{S}^\perp} V(\nabla F(y_0 + u))$ for $u \in \mathcal{U}_V$. Using the relationship $y_0 = M^{1/2} z_0$ and the identity of (43), we then have

$$\nabla F(y_0 + u) = M^{1/2} \nabla H(z_0 + M^{-1/2} u) \; . \tag{48}$$

It then follows that

$$\pi_0^V(u) = \Pi_{\mathcal{S}^\perp} M^{1/2} G(\nabla H(z_0 + M^{-1/2} u)) = M^{1/2} \pi_G(M^{-1/2} u), \tag{49}$$

which implies that $u$ is a zero of $\pi_0^V$ if and only if $v = M^{-1/2} u$ is a zero of $\pi_0^G$. Thus, differentiating on the effective dual spaces gives the relationship $J_{\pi_0^V}(u) = M^{1/2} J_{\pi_0^G}(v) M^{-1/2}$. Now at every such zero, recall that Assumption 5 implies that the map $J_{\pi_0^G}(v) : \mathcal{U}_G \to \mathcal{U}_G$ is nonsingular. Thus, since $M^{1/2} : \mathcal{U}_G \to \mathcal{U}_V$ is an isomorphism, it follows that $J_{\pi_0^V}(u) : \mathcal{U}_V \to \mathcal{U}_V$ is also nonsingular.

Finally, write $\mathcal{R}_G(w_0)$ and $\mathcal{R}_V(x_0)$ to denote the respective reachable primal sets of the two systems, Then by Corollary D.2, it follows immediately that $\mathcal{R}_G(w_0) \not\subseteq \mathrm{NE}(G)$ if and only if $\mathcal{R}_V(x_0) \not\subseteq \mathrm{NE}(V)$. Thus, when Assumption 5 holds for the original system, then the two conditions similarly hold for the symmetrized system. ■

### D.3 Summary of Symmetrized Setting

Here, we give a summary of the key properties of the symmetrized system that will be used in the remainder of the analysis:

**Setting D.1** (Summary of symmetrized system)**.** *Assume the operator $G : \mathcal{W} \to \mathbb{R}^D$ satisfies $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weights matrix $M$ satisfying Assumption 1. Let $Q : \mathcal{W} \to \mathbb{R}$ and $H = Q^* : \mathbb{R}^D \to \mathbb{R}$ satisfy Assumptions 2, 3, and 4. Further let Assumption 5 hold.*

*Then the following properties hold for the symmetrized system:*

- *$V : \mathcal{X} \to \mathbb{R}^D$ is $\widehat{L}$-Lipschitz and $\widehat{B}$-bounded with $\widehat{L} = \sqrt{\kappa_M} L$ and $\widehat{B} = \sqrt{m_{\max}} B$.*
- *$w^\star \in \mathrm{WMEQ}_M(G)$ if and only if $x^\star \in \mathrm{MEQ}(V)$.*
- *$R : \mathcal{X} \to \mathbb{R}$ and $F = R^* : \mathbb{R}^D \to \mathbb{R}$ satisfy Assumption 2 with respect to $\mathrm{ri}(\mathcal{X})$.*
- *$F$ satisfies (LHS) with parameter $\widehat{\omega} = \omega / \sqrt{m_{\min}}$.*
- *For the iterates $\{(w_t, z_t)\}$ of (FTRL), (IFTRL), (OFTRL), or (EGFTRL) with stepsize $\eta > 0$, letting $x_t = M^{1/2} w_t$ and $y_t = M^{1/2} z_t$ for all $t \geq 0$, it holds correspondingly that:*

$$y_{t+1} = y_t - \eta\, V(x_t) \tag{FTRL}$$
$$y_{t+1} = y_t - \eta\, V(x_{t+1}) \tag{IFTRL}$$
$$y_{t+1} = y_t - 2\eta\, V(x_t) + \eta\, V(x_{t-1}) \tag{OFTRL}$$
$$y_{t+1/2} = y_t - \eta\, V(x_t)\,, \quad y_{t+1} = y_t - \eta\, V(x_{t+1/2})\,. \tag{EGFTRL}$$

  *Moreover, for each variant, it holds for all $t \geq 0$ that $\|y_{t+1} - y_t\|_* \leq 3\eta \hat{B}$.*

- *Under Assumption 5 for the original system with $(G, Q, H)$ the corresponding properties from the assumption also hold for the symmetrized system with $(V, R, F)$.*

*Given the equivalence between the original and symmetrized systems, throughout, we will freely apply preliminaries proved for the original system (e.g., from Sections C.2).*

# E Details on Effective Dual Space and Reachable Primal Sets

In this section, we give more details on properties on the reachable primal set, effective dual space, and the stationarity map introduce in Section 3.2. Specifically, given the reduction to the symmetrized system introduced in Section D (as summarized in Setting D.1), we consider here the analogous objects for the symmetrized system. We begin by recalling these definitions:

**Recap of Definitions.** Let $\mathcal{Y} = \text{Span}(V(\mathcal{X}))$ and $\mathcal{U} = \Pi_{\mathcal{S}^\perp}\mathcal{Y} \subseteq \mathcal{S}^\perp$ for $\mathcal{S}^\perp$ is defined in (9). For a fixed primal and dual initialization $x_0 = \nabla F(y_0) \in \text{ri}(\mathcal{X})$, recall that we define

$$\mathcal{R}(x_0) = \{\nabla F(y_0 + y) \ : \ y \in \mathcal{Y}\} \ .$$

The effective stationarity map $\pi_0 : \mathcal{U} \to \mathcal{U}$ is given by

$$\pi_0(u) = \Pi_{\mathcal{S}^\perp} V(\nabla F(y_0 + u)) \ \text{ for all } u \in \mathcal{U} \ .$$

**Key property of effective stationarity map.**

To begin we note the following important relationship between zeros of $\pi_0$ with the set $\text{NE}(V)$:

**Proposition E.1.** *Fix $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0) \in \text{ri}(\mathcal{X})$ and consider the map $\pi_0 : \mathcal{U} \to \mathcal{U}$. Then for all $u \in \mathcal{U}$, it holds that $\pi_0(u) = 0$ if and only if $\nabla F(y_0 + u) \in \text{NE}(V)$.*

**Proof.** Fix $u \in \mathcal{U}$, and let $x_0(u) := \nabla F(y_0 + u) \in \text{ri}(\mathcal{X})$. For the forward direction suppose $\pi_0(u) = 0$. Then by definition, we must have $V(x_0(u)) \in \mathcal{S}$. By Proposition B.2 (i) (applied to the symmetrized system), this implies that $x_0(u) \in \text{NE}(V)$. For the other direction, suppose $x_0(u) \in \text{NE}(V)$. Then since also $x_0(u) \in \text{ri}(\mathcal{X})$, it follows by Proposition B.2 (ii) that $V(x_0(u)) \in \mathcal{S}$. Thus by definition, $\pi_0(u) = \Pi_{\mathcal{S}^\perp} V(x_0(u)) = 0$. ■

Next, we give in Section E.1 several properties of the reachable primal sets $\mathcal{R}(x_0)$. Then, in Section E.2, we derive several consequences of the regularity property of Assumption 5. These properties and consequences are used to later establish our equilibrium selection result and operator domination inequality.

## E.1 Properties of Reachable Primal Set

**Unique reachable MEQ point.**

First, we establish that for any initialization, the corresponding reachable primal set contains at most one MEQ point (note this holds without Assumption 5).

**Proposition E.2** (Unique reachable MEQ point)**.** *Fix $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0) \in \text{ri}(\mathcal{X})$. Then*

$$\left|\text{MEQ}(V) \cap \mathcal{R}(x_0)\right| \leq 1 \ .$$

**Proof.** Let $x_1^\star, x_2^\star \in \text{MEQ}(V) \cap \mathcal{R}(x_0)$. We will prove that this implies $x_1^\star = x_2^\star$.

*Step 1. Consequence of membership in MEQ set.*

First, observe by (MEQ) that $x_1^\star \in \text{MEQ}(V)$ and $x_2^\star \in \text{MEQ}(V)$ implies that $0 = \langle V(x), x - x_1^\star\rangle = \langle V(x), x - x_2^\star\rangle$ for all $x \in \mathcal{X}$. Subtracting then implies that

$$\langle V(x), x_1^\star - x_2^\star\rangle = 0 \ \ \forall x \in \mathcal{X} \ .$$

Thus $x_1^\star - x_2^\star$ is orthogonal to $V(x)$ for all $x \in \mathcal{X}$, meaning $x_1^\star - x_2^\star \in \mathcal{Y}^\perp = \text{Span}(V(\mathcal{X}))^\perp$.

*Step 2. Consequence of membership in reachable primal set.*

Now as $x_1^\star, x_2^\star \in \mathcal{R}(x_0)$, this implies there exists $y_1, y_2 \in \mathcal{Y}$ such that $x_1^\star = \nabla F(y_0 + y_1)$ and $x_2^\star = \nabla F(y_0 + y_2)$. Define $u_1 := \Pi_{\mathcal{S}^\perp} y_1 \in \mathcal{U}$ and $u_2 := \Pi_{\mathcal{S}^\perp} y_2 \in \mathcal{U}$. Now by the shift-invariance property of $\nabla F$ from Proposition C.4, it also follows that $x_1^\star = \nabla F(y_0 + u_1)$ and $x_2^\star = \nabla F(y_0 + u_2)$.

*Step 3. Combining the pieces.*

As $x_1^\star - x_2^\star \in \mathcal{Y}^\perp$ and $y_1 - y_2 \in \mathcal{Y}$, it follows that $\langle x_1^\star - x_2^\star, y_1 - y_2\rangle = 0$. Moreover, we can express $y_1 - y_2$ via the orthogonal decomposition

$$y_1 - y_2 = \Pi_{\mathcal{S}^\perp}(y_1 - y_2) + \Pi_{\mathcal{S}}(y_1 - y_2) = u_1 - u_2 + \Pi_{\mathcal{S}}(y_1 - y_2) \ ,$$

where the final equality uses the definitions of $u_1, u_2 \in \mathcal{U} \subseteq \mathcal{S}^\perp$. Combining these two facts yields

$$\begin{aligned} 0 &= \langle x_1^\star - x_2^\star, y_1 - y_2 \rangle \\ &= \langle x_1^\star - x_2^\star, u_1 - u_2 \rangle + \langle x_1^\star - x_2^\star, \Pi_{\mathcal{S}}(y_1 - y_2) \rangle \\ &= \langle \nabla F(y_0 + u_1) - \nabla F(y_0 + u_2), u_1 - u_2 \rangle \ . \end{aligned} \tag{50}$$

Here, the final equality comes from the fact that $x_1^\star, x_2^\star \in \mathcal{X}$ and thus $x_1^\star - x_2^\star \in \mathcal{S}^\perp$, and also using the characterization of $x_1^\star, x_2^\star$ from Step 2. Now let $d = u_1 - u_2$, and observe by the fundamental theorem of calculus that

$$\langle \nabla F(y_0 + u_1) - \nabla F(y_0 + u_2), d \rangle = \int_0^1 \langle d, \nabla^2 F(y_0 + u_2 + sd) d \rangle \, ds.$$

By convexity of $F$, and using Proposition C.5 (iii), the Hessian $\nabla^2 F$ in the integral is positive definite on $\mathcal{S}^\perp$. As $u_1, u_2 \in \mathcal{U} \subseteq \mathcal{S}^\perp$, then also $d = u_1 - u_2 \in \mathcal{S}^\perp$. Thus by positive definiteness, the integral must be strictly positive for $d \neq 0$. Thus (50) subsequently implies that we must have $u_1 = u_2$, and therefore $x_1^\star = \nabla F(y_0 + u_1) = \nabla F(y_0 + u_2) = x_2^\star$, which concludes the proof. ■

**Useful characterization of reachable primal set.**

We further give the following useful characterization of the reachable primal set $\mathcal{R}(x_0)$:

**Proposition E.3.** *Fix initial $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0) \in \mathrm{ri}(\mathcal{X})$. Then for all $x \in \mathrm{ri}(\mathcal{X})$:*

$$x \in \mathcal{R}(x_0) \iff \Pi_{\mathcal{S}^\perp}(\nabla R(x) - y_0) \in \mathcal{U} \ .$$

**Proof.** Recall that $\mathcal{Y} = \mathrm{Span}(V(\mathcal{X}))$, that $\mathcal{U} = \Pi_{\mathcal{S}^\perp}\mathcal{Y}$, and $\mathcal{R}(x_0) = \{\nabla F(y_0 + v) : v \in \mathcal{Y}\}$. Now by the primal-dual relationship of Proposition C.2, recall that for $x \in \mathrm{ri}(\mathcal{X})$ and $y \in \mathbb{R}^D$ we have $x = \nabla F(y)$ if and only if $y - \nabla R(x) \in \mathcal{S}$. Then fixing such an $x \in \mathrm{ri}(\mathcal{X})$, we then have the following string of equivalences:

$$\begin{aligned} x \in \mathcal{R}(x_0) &\iff \exists v \in \mathcal{Y} : y_0 + v - \nabla R(x) \in \mathcal{S} \\ &\iff \exists v \in \mathcal{Y} : \Pi_{\mathcal{S}^\perp}(y_0 + v - \nabla R(x)) = 0 \\ &\iff \exists v \in \mathcal{Y} : \Pi_{\mathcal{S}^\perp}(\nabla R(x) - y_0) = \Pi_{\mathcal{S}^\perp} v \\ &\iff \Pi_{\mathcal{S}^\perp}(\nabla R(x) - y_0) \in \mathcal{U} \ , \end{aligned}$$

where the final equivalence uses the definition of $\mathcal{U} = \Pi_{\mathcal{S}^\perp}\mathcal{Y}$. ■

### E.2 Regularity of Effective Stationary Map

In this section, we derive an important consequence of the regularity condition of Assumption 5. In particular, we show in the following proposition that *regular* zeros of the stationarity map $\pi_0$ must also be MEQ points. As a result, under the conditions of Assumption 5 (when applied to the symmetrized system), every reachable zero of the map $\pi_0$ corresponds to an MEQ point (see Corollary E.5). We begin with the following proposition:

**Proposition E.4** (Regular reachable zeros are MEQ points)**.** *Let $\mathrm{MEQ}(V) \neq \emptyset$. Fix initial $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0) \in \mathrm{ri}(\mathcal{X})$. For any $y \in \mathcal{Y}$, let $u = \Pi_{\mathcal{S}^\perp} y \in \mathcal{U}$. If $\pi_0(u) = 0$ and $J_{\pi_0}(u) : \mathcal{U} \to \mathcal{U}$ is non-singular, then $\nabla F(y_0 + u) \in \mathrm{MEQ}(V)$.*

**Proof.** First, fixing $y \in \mathcal{Y}$ and $u = \Pi_{\mathcal{S}^\perp} y \in \mathcal{U} \subseteq \mathcal{S}^\perp$, define $x_0(u) = \nabla F(y_0 + u) \in \mathrm{ri}(\mathcal{X})$. Conditioned on $\pi_0(u) = 0$, we will prove the contrapositive statement that $x_0(u) \notin \mathrm{MEQ}(V)$ implies $J_{\pi_0}(u) : \mathcal{U} \to \mathcal{U}$ is singular. For this, we proceed in four steps:

*Step 1. Consequence of non-empty MEQ set.*

As $\mathrm{MEQ}(V) \neq \emptyset$, consider $x^\star \in \mathrm{MEQ}(V)$. Let $\phi : \mathcal{X} \to \mathbb{R}$ be the scalar function given by

$$\phi(x) = \langle V(x), x - x^\star \rangle \ \text{ for all } x \in \mathcal{X}. \tag{51}$$

As $x^\star \in \mathrm{MEQ}(V)$, we have that $\phi(x) = 0$ for all $x \in \mathcal{X}$. Thus, differentiating along any direction $v \in \mathcal{S}^\perp$ yields $J_\phi(x)v = 0$ for all $x \in \mathrm{ri}(\mathcal{X})$. By the product rule, this implies

$$J_\phi(x)v = \langle J_V(x)v, x - x^\star \rangle + \langle V(x), v \rangle = 0 \ \text{ for all } x \in \mathrm{ri}(\mathcal{X}) \text{ and } v \in \mathcal{S}^\perp. \tag{52}$$

*Step 2. Consequence of stationarity.*

Recall by assumption that $\pi_0(u) = \Pi_{\mathcal{S}^\perp} V(x_0(u)) = 0$. This implies $V(x_0(u)) \in \mathcal{S}$, and thus $\langle V(x_0(u)), v\rangle = 0$ for all $v \in \mathcal{S}^\perp$. Applying expression (52) at $x_0(u)$ then yields

$$\langle J_V(x_0(u))v, x_0(u) - x^\star\rangle = 0 \;\text{ for all } v \in \mathcal{S}^\perp\,. \tag{53}$$

As $x_0(u), x^\star \in \mathcal{X}$, we have $x_0(u) - x^\star \in \mathcal{S}^\perp$, and thus applying the orthogonal decomposition

$$J_V(x_0(u))v = \Pi_{\mathcal{S}} J_V(x_0(u))v + \Pi_{\mathcal{S}^\perp} J_V(x_0(u))v \;\text{ for any } v \in \mathcal{S}^\perp.$$

We then further obtain from (53) that

$$\left\langle \Pi_{\mathcal{S}^\perp} J_V(x_0(u))v, x_0(u) - x^\star\right\rangle = 0 \;\text{ for all } v \in \mathcal{S}^\perp\,. \tag{54}$$

*Step 3. Construction of null vector.*

Now by the chain rule, observe for $\pi_0(u) = \Pi_{\mathcal{S}^\perp} V(\nabla F(y_0 + u))$ that the action of the Jacobian $J_{\pi_0}(u) : \mathcal{U} \to \mathcal{U}$ on a vector $h \in \mathcal{U}$ is given by

$$J_{\pi_0}(u)h = \Pi_{\mathcal{S}^\perp} J_V(x_0(u))\nabla^2 F(y_0 + u)h \;\text{ for all } h \in \mathcal{U}\,. \tag{55}$$

Here, we also use notational substitution $x_0(u) = \nabla F(y_0 + u)$. Moreover, for every $h \in \mathcal{U}$, we also have by Proposition C.5 (iii) that $\nabla^2 F(y_0 + u)h \in \mathcal{S}^\perp$. Thus applying expression (54) at $v = \nabla^2 F(y_0 + u)h$ for all $h \in \mathcal{U}$ and substituting (55) yields

$$\begin{aligned} 0 &= \left\langle \Pi_{\mathcal{S}^\perp} J_V(x_0(u))\nabla^2 F(y_0 + u)h, x_0(u) - x^\star\right\rangle \\ &= \left\langle J_{\pi_0}(u)h, x_0(u) - x^\star\right\rangle \\ &= \left\langle J_{\pi_0}(u)h, \Pi_{\mathcal{U}}(x_0(u) - x^\star)\right\rangle \\ &= \left\langle h, J_{\pi_0}(u)^\top \Pi_{\mathcal{U}}(x_0(u) - x^\star)\right\rangle \;\text{ for all } h \in \mathcal{U}\,. \end{aligned} \tag{56}$$

Here, the penultimate equality uses the fact that $J_{\pi_0}(u)h \in \mathcal{U}$ for all $h \in \mathcal{U}$, and thus we can project $x_0(u) - x^\star$ onto $\mathcal{U}$ and preserve equality of the inner product. As (56) holds for all $h \in \mathcal{U}$:

$$J_{\pi_0}(u)^\top \Pi_{\mathcal{U}}(x_0(u) - x^\star) = 0\,, \tag{57}$$

meaning $\Pi_{\mathcal{U}}(x_0(u) - x^\star)$ is in the nullspace of $J_{\pi_0}(u)^\top$.

*Step 4. Non-MEQ stationary point implies non-trivial nullspace.*

To complete the proof, it now suffices to establish that $x_0(u) \notin \mathrm{MEQ}(V)$ implies that the vector $\Pi_{\mathcal{U}}(x_0(u) - x^\star) \neq 0$. By (57), this will imply that $J_{\pi_0}(u)^\top$ has a non-trivial nullspace, and thus $J_{\pi_0}(u) : \mathcal{U} \to \mathcal{U}$ is singular.

For this, recall that we assume $x^\star \in \mathrm{MEQ}(V)$, and thus for every $x \in \mathcal{X}$, we have

$$\langle V(x), x - x_0(u)\rangle = -\langle V(x), x_0(u) - x^\star\rangle\,. \tag{58}$$

Thus $x_0(u) \in \mathrm{MEQ}(V)$ if and only if the right hand side of (58) vanishes for all $x \in \mathcal{X}$, and thus

$$x_0(u) \in \mathrm{MEQ}(V) \iff x_0(u) - x^\star \in \mathcal{Y}^\perp\,. \tag{59}$$

Now since $x_0(u), x^\star \in \mathcal{X}$, then also $x_0(u) - x^\star \in \mathcal{S}^\perp$ by Lemma D.1. Thus we also have

$$\langle y, x_0(u) - x^\star\rangle = \langle \Pi_{\mathcal{S}^\perp} y, x_0(u) - x^\star\rangle \;\text{ for all } y \in \mathcal{Y}\,. \tag{60}$$

Therefore, since $\mathcal{U} = \Pi_{\mathcal{S}^\perp}\mathcal{Y}$, then

$$x_0(u) - x^\star \in \mathcal{Y}^\perp \iff x_0(u) - x^\star \in \mathcal{U}^\perp\,. \tag{61}$$

Combining the equivalences of (59) and (61) then implies that

$$x_0(u) \in \mathrm{MEQ}(V) \iff x_0(u) - x^\star \in \mathcal{U}^\perp\,. \tag{62}$$

Thus, if $x_0(u) \notin \mathrm{MEQ}(V)$, then $x_0(u) - x^\star \notin \mathcal{U}^\perp$. This implies that $\Pi_{\mathcal{U}}(x_0(u) - x^\star) \neq 0$, which yields the desired claim and concludes the proof. ■

As an immediate corollary of Proposition E.1 and Proposition E.4, we then have under Assumption 5 (and by Proposition D.5 which extends these properties to the present, symmetrized system) the following characterization:

**Corollary E.5.** *Fix $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0)$, and suppose Assumption 5 holds for this initialization. Then for every $x \in \mathcal{R}(x_0) \subseteq \mathrm{ri}(\mathcal{X})$, if $x \in \mathrm{NE}(V)$ then $x \in \mathrm{MEQ}(V)$.*

### E.2.1 Regularity Under Linear Harmonic Operators

When $V : \mathcal{X} \to \mathbb{R}^D$ is the linear map $V(x) = Kx$ for some $K \in \mathbb{R}^{D\times D}$ satisfying the harmonic property $\mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, then $V$ always satisfies the regularity condition of Assumption 5 (for the symmetrized system). We formally state and prove this here:

**Proposition E.6** (Regularity under linearity)**.** *Let $V : \mathcal{X} \to \mathbb{R}^D$ be given by the linear map $x \mapsto Kx$ for some $K \in \mathbb{R}^{D\times D}$, and suppose that $\mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$. Let $R$ satisfy Assumption 2 with respect to $\mathcal{X}$ and let $F = R^*$ satisfy Assumption 4. Fix an initialization pair $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0) \in \mathrm{ri}(\mathcal{X})$ such that $\mathcal{R}(x_0) \not\subseteq \mathrm{NE}(V)$. Then the Jacobian $J_{\pi_0}(u) : \mathcal{U} \to \mathcal{U}$ is nonsingular at every $u \in \mathcal{U}$, and thus Assumption 5 holds whenever $\mathcal{R}(x_0) \not\subseteq \mathrm{NE}(V)$.*

**Proof.** First, define the linear map $A : \mathcal{S}^\perp \to \mathcal{S}^\perp$ by $Av := \Pi_{\mathcal{S}^\perp} Kv$. Then fix some $p \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$, and observe by Proposition B.6 and Proposition B.2 that $V(p) = Kp \in \mathcal{S}$. Now as $p \in \mathrm{ri}(\mathcal{X})$, then for any $v \in \mathcal{S}^\perp$, there exists a sufficiently small $t > 0$ such that $p + tv \in \mathcal{X}$. Then applying the (MEQ) at $p + tv$ gives $0 = \langle K(p + tv), tv\rangle = t^2\langle Av, v\rangle$. Thus $\langle Av, v\rangle = 0$ for every $v \in \mathcal{S}^\perp$, which implies that $A^\top = -A$ as an operator on $\mathcal{S}^\perp$. Moreover, since $\Pi_{\mathcal{S}^\perp} Kp = 0$ and $p \in \mathrm{ri}(\mathcal{X})$, we have $\mathrm{Span}(\mathcal{X} - p) = \mathcal{S}^\perp$ and thus $\mathcal{U} = \mathrm{Span}\{\Pi_{\mathcal{S}^\perp} Kx : x \in \mathcal{X}\} = A(\mathcal{S}^\perp)$.

Thus, when viewing $A$ as an operator on $\mathcal{S}^\perp$, it then follows that

$$\mathrm{Null}(A) = \mathrm{Range}(A)^\perp \cap \mathcal{S}^\perp = \mathcal{U}^\perp \cap \mathcal{S}^\perp.$$

Now fix $u \in \mathcal{U}$, and let $P := \nabla^2 F(y_0+u)$. Recall by Proposition C.5 that $P \succeq 0$ with $\mathrm{Null}(P) = \mathcal{S}$. In particular, $P$ maps into $\mathcal{S}^\perp$ and is positive definite on $\mathcal{S}^\perp$. Differentiating $\pi_0$ then yields

$$J_{\pi_0}(u)v = \Pi_{\mathcal{S}^\perp} KPv = APv \quad \text{for every } v \in \mathcal{U}\ .$$

If $J_{\pi_0}(u)v = 0$ for some $v \in \mathcal{U}$, then $Pv \in \mathrm{Null}(A) \subseteq \mathcal{U}^\perp$. Since $v \in \mathcal{U}$, it follows that $\langle v, Pv\rangle = 0$. Then as $P$ is positive definite on $\mathcal{S}^\perp$, and $\mathcal{U} \subseteq \mathcal{S}^\perp$, then we must have $v = 0$. Thus $J_{\pi_0}(u) : \mathcal{U} \to \mathcal{U}$ has only the trivial nullspace and hence is injective and nonsingular. This holds at every $u \in \mathcal{U}$, and in particular at every zero of $\pi_0$. Together with the assumption $\mathcal{R}(x_0) \not\subseteq \mathrm{NE}(V)$, this establishes the conditions of Assumption 5. ■

## F Intuition from Continuous Time

In this section, we give more details on the continuous-time intuition introduced in Section 5.1.

**Continuous-time FTRL flow.**

First, recall that we can consider the FTRL flow:

$$\dot{y}_t = -V(x_t) \quad \text{and} \quad x_t = \nabla F(y_t)\ . \tag{FTRL flow}$$

We showed in (15) that under (FTRL flow) that

$$\tfrac{d}{dt}\mathsf{D}_R(x^\star, x_t) = \tfrac{d}{dt}F_{x^\star}(y_t) = -\langle V(x_t), x_t - x^\star\rangle = 0 \tag{63}$$

for all points $x^\star \in \mathrm{MEQ}(V)$.

**Forward and backward discretizations of FTRL flow.**

Under the forward and backward discretizations of (FTRL flow), we then obtain the following bounds on change in shifted energy, as stated in (16):

**Proposition F.1** (Change in shifted energy for FTRL and IFTRL)**.** *Consider the symmetrized system of Setting D.1. Then along* (FTRL) *and* (IFTRL)*, and for any $x^\star \in \mathrm{MEQ}(V)$:*

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) = \begin{cases} \mathsf{D}_F(y_{t+1}, y_t) \ \geq\ 0 & \text{for FTRL} \\ -\mathsf{D}_F(y_t, y_{t+1}) \ \leq\ 0 & \text{for IFTRL}\ . \end{cases}$$

**Proof.** Let $x^\star \in \mathrm{MEQ}(V)$. Then, by definition of $\mathrm{MEQ}(V)$, we have:

$$\langle V(x), x - x^\star\rangle = 0\,, \quad \forall x \in \mathcal{X}\,. \tag{64}$$

By definition of the shifted energy function $F_{x^\star}$ (see (14)), we have for any $t \geq 0$,

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) = F(y_{t+1}) - F(y_t) - \langle x^\star, y_{t+1} - y_t \rangle \,. \tag{65}$$

We start by proving the identity for vanilla FTRL. By definition of the Bregman divergence, we have for any $t \geq 0$,

$$\begin{aligned} \mathsf{D}_F(y_{t+1}, y_t) &= F(y_{t+1}) - F(y_t) - \langle \nabla F(y_t), y_{t+1} - y_t \rangle \\ &= F(y_{t+1}) - F(y_t) - \langle x_t, y_{t+1} - y_t \rangle \,, \end{aligned} \tag{66}$$

where the last identity follows from the primal-dual relationship $x_t = \nabla F(y_t)$.

Rearranging (66) and combining it with (65) yields:

$$\begin{aligned} F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) &= \mathsf{D}_F(y_{t+1}, y_t) + \langle x_t - x^\star, y_{t+1} - y_t \rangle \\ &= \mathsf{D}_F(y_{t+1}, y_t) - \eta \langle x_t - x^\star, V(x_t) \rangle \\ &= \mathsf{D}_F(y_{t+1}, y_t) \,, \end{aligned} \tag{67}$$

where the penultimate identity follows from using FTRL's dual update rule ($y_{t+1} = y_t - \eta V(x_t)$), and the last equality follows from (64) since $x^\star \in \mathrm{MEQ}(V)$ and $x_t \in \mathcal{X}$ for all $t \geq 0$. This concludes the proof of the first identity for FTRL.

As for IFTRL, the proof follows similar lines. Using again the definition of the Bregman divergence, we have for any $t \geq 0$,

$$\begin{aligned} \mathsf{D}_F(y_t, y_{t+1}) &= F(y_t) - F(y_{t+1}) - \langle \nabla F(y_{t+1}), y_t - y_{t+1} \rangle \\ &= F(y_t) - F(y_{t+1}) - \langle x_{t+1}, y_t - y_{t+1} \rangle \,, \end{aligned} \tag{68}$$

where the last identity follows from the primal-dual relationship $x_{t+1} = \nabla F(y_{t+1})$.

Rearranging and combining the above identity (68) with (65) yields for any $t \geq 1$,

$$\begin{aligned} F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) &= -\mathsf{D}_F(y_t, y_{t+1}) + \langle x^\star - x_{t+1}, y_t - y_{t+1} \rangle \\ &= -\mathsf{D}_F(y_{t+1}, y_t) + \eta \langle x^\star - x_{t+1}, V(x_{t+1}) \rangle \\ &= -\mathsf{D}_F(y_{t+1}, y_t) \,, \end{aligned} \tag{69}$$

where the penultimate identity follows from using IFTRL's dual update rule ($y_{t+1} = y_t - \eta V(x_{t+1})$), and the last equality follows from (64) since $x^\star \in \mathrm{MEQ}(V)$ and $x_{t+1} \in \mathcal{X}$ for all $t \geq 0$. This concludes the proof of the second identity for IFTRL. ■

## G Properties of Energy Function Hessian

In this section, we develop the machinery that will be used in the proof of Lemma 5.4, which establishes one-step bounds on the dissipation of the shifted energy function along the OFTRL and EGFTRL iterates. We establish this machinery for the symmetrized system with operator $V : \mathcal{X} \to \mathbb{R}^D$, regularizer $R$, and energy function $F = R^*$ (as summarized in Proposition 5.2). In particular, most of the bounds derived in this section stem from the local Hessian stability property (LHS) for the energy Hessian, i.e. $\nabla^2 F$.

We start with a brief recap of the assumptions satisfied under the symmetrized system:

**Summary of global assumptions for symmetrized system.**

Recall from D.1 that the operator $V : \mathcal{X} \to \mathbb{R}^D$, regularizer $R : \mathcal{X} \to \mathbb{R}$, and energy $F : \mathbb{R}^D \to \mathbb{R}$ satisfy the following properties:

- $V$ is $\widehat{L}$-Lipschitz and $\widehat{B}$-Bounded.
- $R$ is relatively Legendre and 1-strongly-convex with respect to $\|\cdot\|$.
- $F = R^*$ is 1-smooth with respect to $\|\cdot\|_*$ and satisfies (LHS) with constant $\widehat{\omega} > 0$.

### G.1 Properties of Local Semi-norms.

We start by recalling the notation for the *local semi-norms* introduced in Section 5. Specifically, for all $y \in \mathbb{R}^D$, recall that $\nabla^2 F(y) \succeq 0$ is positive semi-definite, but with the non-trivial nullspace

$\text{Null}(\nabla^2 F(y)) = \mathcal{S}$ (Proposition C.5). Thus, $\nabla^2 F(y)$ can be used to define a local degenerate inner-product, which in turn induces a local semi-norm as below:

$$\langle u, v\rangle_y := \langle u, \nabla^2 F(y) v\rangle \;\text{ and }\; \|v\|_y := \sqrt{\langle v, \nabla^2 F(y) v\rangle} \;\text{ for all } u, v \in \mathbb{R}^D \,. \tag{70}$$

Importantly, note that $\|v\|_y = 0$ if and only if $v \in \mathcal{S} = \text{Null}(\nabla^2 F(y))$ (which holds even if $v \neq 0$).

When restricted to the subspace $\mathcal{S}^\perp$, we also define the dual semi-norm $\|\cdot\|_{y,*}$. In particular, for $x = \nabla F(y)$, recall that we have the primal-dual relationship between the operators $\nabla^2 F$ and $\nabla^2 R$ given by $(\nabla^2 F(y))^{-1} = \nabla^2 R(x)$, when the operators are restricted to the subspace $\mathcal{S}^\perp$, e.g., when applied to vectors $v \in \mathcal{S}^\perp$ (Proposition C.5). Then for $y \in \mathbb{R}^D$, we define

$$\|u\|_{y,*} := \sqrt{\langle u, (\nabla^2 F(y))^{-1} u\rangle} \;\text{ for } u \in \mathcal{S}^\perp. \tag{71}$$

For $y \in \mathbb{R}^D$, the semi-norm $\|\cdot\|_y$ satisfies the Cauchy-Schwarz inequality:

$$|\langle v, v'\rangle_y| \leq \|v\|_y \cdot \|v'\|_y \;\text{ for } v, v' \in \mathbf{r}^D \,. \tag{72}$$

Moreover, the pair $(\|\cdot\|_y, \|\cdot\|_{y,*})$ satisfies the generalized Cauchy-Schwarz inequality:

$$|\langle u, v\rangle| \leq \|u\|_{y,*} \cdot \|v\|_y \;\text{ for } u \in \mathcal{S}^\perp \text{ and } v \in \mathbb{R}^D \,. \tag{73}$$

Finally, fixing $y \in \mathbb{R}^D$ and letting $P_y = \nabla^2 F(y)$ and $P_y^{-1} = (\nabla^2 F(y))^{-1}$ (restricted to $\mathcal{S}^\perp$), we further have by definition the following equivalent and useful characterizations:

$$\|v\|_y = \|P_y^{1/2} v\|_2 \;\text{ and }\; \|u\|_{y,*} = \|P_y^{-1/2} u\|_2 \;\text{ for all } v \in \mathbb{R}^D \text{ and } u \in \mathcal{S}^\perp \,. \tag{74}$$

See Rudin (1991) for further background and preliminaries on the degenerate inner products and semi-norms induced by positive semi-definite matrices.

### G.2 Relationship between Local and Global (Semi)-Norms

The following lemma relates the above local norms to particular global norms assuming strong convexity of the regularizer $R$ with respect to the global norm.

**Lemma G.1.** *Assume that $R$ is 1-strongly-convex with respect to a particular norm $\|\cdot\|$ (equivalently, that $F$ is 1-smooth with respect to its dual norm $\|\cdot\|_*$). Then, for all $y \in \mathbb{R}^D$ the following global relationships hold.*

*(i)* $\|v\|_y \leq \|v\|_*$ *for all* $v \in \mathbb{R}^D$.

*(ii)* $\|u\| \leq \|u\|_{y,*}$ *for all* $u \in \mathcal{S}^\perp$.

**Proof.** Fix $y \in \mathbb{R}^D$ and $x = \nabla F(y) \in \text{ri}(\mathcal{X})$. First, by 1-strong-convexity of $R$ with respect to the norm $\|\cdot\|$, we have the second-order characterization: $\langle u, \nabla^2 R(x) u\rangle \geq \|u\|^2$ for all $u \in \mathcal{S}^\perp$ (see, e.g., Boyd & Vandenberghe (2004), Section 9.1). As $\nabla^2 R(x) = (\nabla^2 F(y))^{-1}$ when restricted to the subspace $\mathcal{S}^\perp$ it then follows for all $u \in \mathcal{S}^\perp$ that

$$\|u\|^2 \leq \langle u, \nabla^2 R(x) u\rangle = \langle u, (\nabla^2 F(y))^{-1} u\rangle = \|u\|_{y,*}^2 \,.$$

Taking square roots yields claim (ii) of the lemma. For claim (i), we similarly have by 1-smoothness of $F = R^*$ with respect to the dual norm $\|\cdot\|_*$ that $\|v\|_y^2 = \langle v, \nabla^2 F(y) v\rangle \leq \|v\|_*^2$ for all $v \in \mathbb{R}^D$. Taking square roots then yields claim (i) and completes the proof. ■

### G.3 Local Semi-Norm Transfer Inequalities

Under the local Hessian stability property of Assumption 4, the following inequalities are immediate:

**Lemma G.2.** *Let $y, y' \in \mathbb{R}^D$ satisfy $\|y' - y\|_* \leq \alpha$ for $\alpha > 0$. Then for every $v \in \mathbb{R}^D$ it holds that*

$$\exp(-\widehat{\omega}\alpha) \cdot \|v\|_{y'}^2 \leq \|v\|_y^2 \leq \exp(\widehat{\omega}\alpha) \|v\|_{y'}^2 \,.$$

### G.4 Two-Sided Bounds on Bregman Divergences

In this section, we derive bounds on the Bregman divergence $\mathsf{D}_F(\cdot, \cdot)$ induced by the energy Hessian that hold under the local semi-norm.

**Bregman divergence in integral form.** First, recall that the Bregman divergence $\mathsf{D}_F(y', y)$ admits the following integral formulation as a result of Taylor's theorem:

**Proposition G.3.** *Let $y, y' \in \mathbb{R}^D$, and let $y_s := y + s(y' - y)$ for all $s \in [0, 1]$. Then the Bregman divergence $\mathsf{D}_F(y', y)$ is given by*

$$\mathsf{D}_F(y', y) = \int_0^1 (1 - s)\langle y' - y, \nabla^2 F(y_s)(y' - y)\rangle \, ds \, .$$

This result is standard, and thus we omit a proof (see, e.g., Nocedal & Wright (2006), Theorem 2.1).

Then, under the (LHS) property, the energy-function-induced Bregman divergence $\mathsf{D}_F(y', y)$ for sufficiently close vectors under the global dual norm $\|\cdot\|_*$ is proportional to their squared distance under the local semi-norm $\|\cdot\|_y$. Formally, we have the following bounds:

**Lemma G.4.** *Let $y, y' \in \mathbb{R}^D$ be such that $\|y' - y\|_* \leq \alpha$ for $\alpha > 0$. Then it holds that:*

$$\left(\tfrac{\exp(-\widehat{\omega}\alpha)+\widehat{\omega}\alpha-1}{(\widehat{\omega}\alpha)^2}\right) \cdot \|y' - y\|_y^2 \leq \mathsf{D}_F(y', y) \leq \left(\tfrac{\exp(\widehat{\omega}\alpha)-\widehat{\omega}\alpha-1}{(\widehat{\omega}\alpha)^2}\right) \cdot \|y' - y\|_y^2 \, .$$

Note that this is the generalization of (Lazarsfeld et al., 2026, Lemma E.9), and we include a full proof here for completeness:

**Proof.** We first prove the upper bound. For this, let $y_s := y + s(y' - y)$ for $s \in [0, 1]$, and observe for all such $s$ that $\|y_s - y\|_* = s\|y' - y\|_* \leq s\alpha < 1$, where the final (strict) inequality follows from the assumption that $\|y' - y\|_* \leq \alpha$ for $\alpha \in (0, 1)$. Then using the integral representation of Bregman divergence (Proposition G.3) and the upper bound part of Lemma G.2, we obtain

$$\begin{aligned}
\mathsf{D}_F(y', y) &= \int_0^1 (1 - s)\langle y' - y, \nabla^2 F(y_s)(y' - y)\rangle \, ds \\
&= \int_0^1 (1 - s)\|y' - y\|_{y_s}^2 \, ds \\
&\leq \int_0^1 (1 - s)\Big(\exp(\widehat{\omega}\alpha s)\|y' - y\|_y^2\Big) \, ds \\
&= \Big(\int_0^1 (1 - s)\exp(\widehat{\omega}\alpha s) \, ds\Big) \cdot \|y' - y\|_y^2 \, .
\end{aligned}$$

Integration by parts yields $\int_0^1 (1 - s)\exp(\widehat{\omega}\alpha s)ds = \frac{\exp(\widehat{\omega}\alpha)-\widehat{\omega}\alpha-1}{(\widehat{\omega}\alpha)^2}$, which yields the upper bound.

For the lower bound, we use the lower bound part of Lemma G.2 to similarly find

$$\mathsf{D}_F(y', y) \geq \Big(\int_0^1 (1 - s)\exp(-\widehat{\omega}\alpha s) \, ds\Big)\|y' - y\|_y^2 \, .$$

Then integration by parts again yields $\int_0^1 (1 - s)\exp(-\widehat{\omega}\alpha s)ds = \frac{\exp(-\widehat{\omega}\alpha)+\widehat{\omega}\alpha-1}{(\widehat{\omega}\alpha)^2}$, which yields the lower bound of the lemma and concludes the proof. ■

### G.5 Dual-to-Primal Local Semi-Norm Inequalities

We next establish a relationship between differences of primal iterates $x', x \in \mathcal{X}$ under the local dual seminorm $\|\cdot\|_{y,*}$, and differences of the corresponding dual iterates $y', y \in \mathbb{R}^D$ under the seminorm $\|\cdot\|_y$. Formally, we prove the following:

**Lemma G.5.** *Let $y, y' \in \mathbb{R}^D$ such that $\|y' - y\|_* \leq \alpha$ for $\alpha > 0$. Let $x = \nabla F(y)$ and $x' = \nabla F(y')$. Then it holds that:*

$$\|x' - x\|_{y,*} \leq \left(\tfrac{\exp(\widehat{\omega}\alpha)-1}{\widehat{\omega}\alpha}\right) \cdot \|y' - y\|_y \, .$$

**Proof.** The proof proceeds in several steps:

*1. Shorthand notation.* For readability, we begin by introducing several pieces of shorthand notation. First, we drop the subscript $y$ to denote $P := \nabla^2 F(y)$. Recall that when restricted to $\mathcal{S}^\perp$ the Hessian $P$ is positive-definite, and thus over this subspace we will denote the energy Hessian inverse by

$P^{-1} := (\nabla^2 F(y))^{-1}$. For $s \in [0,1]$, let $y_s := y + s(y' - y)$, and further define $P_s := \nabla^2 F(y_s)$, and the normalized $B_s := P^{-1/2} P_s P^{-1/2}$. As with $P^{-1}$, we only use $B_s$ when explicitly working over $\mathcal{S}^\perp$.

*2. Bound on* $\|B_s\|_2$. Recall that for all $s \in [0,1]$ that $\|y_s - y\|_* = s\|y' - y\|_* \leq \alpha s$. Thus, the (LHS) property with respect to the pair $(y_s, y)$ can be expressed in our shorthand notation as

$$\exp(-\widehat{\omega}\alpha s)P \preceq P_s \preceq \exp(\widehat{\omega}\alpha s)P \; .$$

Then by congruence (using the fact that $P^{-1/2}$ is symmetric), it follows for all $s \in [0,1]$ that

$$\exp(-\widehat{\omega}\alpha s)I \preceq B_s \preceq \exp(\widehat{\omega}\alpha s)I \; .$$

By definition of the spectral norm, this implies for all $s \in [0,1]$ that $\|B_s\|_2 \leq \exp(\widehat{\omega}\alpha s)$.

*3. Expand primal difference in integral form.* Recall by assumption that $x = \nabla F(y)$ and $x' = \nabla F(y')$. It follows by the fundamental theorem of calculus (see, e.g., Nocedal & Wright (2006), Theorem 2.1) that we can write

$$x' - x = \nabla F(y') - \nabla F(y) = \int_0^1 \nabla^2 F(y_s)(y' - y)\, ds = \int_0^1 P_s(y' - y)\, ds \; . \tag{75}$$

As $x' - x \in \mathcal{S}^\perp$, we can take its dual local seminorm. Using the triangle inequality with respect to the integral then yields

$$\|x' - x\|_{y,*} = \Big\| \int_0^1 P_s(y' - y)\, ds \Big\|_{y,*} \leq \int_0^1 \|P_s(y' - y)\|_{y,*}\, ds \; . \tag{76}$$

Now by definition of $\|\cdot\|_{y,*}$ and $B_s = P^{-1/2} P_s P^{-1/2}$, we have for all $s \in [0,1]$ that

$$\begin{aligned} \|P_s(y' - y)\|_{y,*} = \|P^{-1/2} P_s (y' - y)\|_2 &= \|B_s P^{1/2}(y' - y)\|_2 \\ &\leq \|B_s\|_2 \cdot \|P^{1/2}(y' - y)\|_2 \\ &\leq \exp(\widehat{\omega}\alpha s) \cdot \|P^{1/2}(y' - y)\|_2 \\ &= \exp(\widehat{\omega}\alpha s) \cdot \|y' - y\|_y \; . \end{aligned} \tag{77}$$

Here, the first inequality follows by definition of the spectral norm $\|B_s\|_2$, the second inequality follows from the bound in Step 2, and the final equality comes from the definition of $\|\cdot\|_y$.

Substituting (77) into (76) and integrating then yields

$$\|x' - x\|_{y,*} \leq \Big( \int_0^1 \exp(\widehat{\omega}\alpha s)\, ds \Big) \cdot \|y' - y\|_y = \Big( \frac{\exp(\widehat{\omega}\alpha) - 1}{\widehat{\omega}\alpha} \Big) \cdot \|y' - y\|_y \; ,$$

which completes the proof. ■

### G.6 Upper Bound on Hessian Remainder Term

In this section, we derive bounds on the magnitude of the following *Hessian remainder term*:

**Hessian remainder term.** For $y', y \in \mathbb{R}^D$, let $\mathsf{G}_F(y', y) \in \mathbb{R}^D$ be the vector

$$\mathsf{G}_F(y', y) = \nabla F(y') - \nabla F(y) - \nabla^2 F(y)(y' - y) \; . \tag{78}$$

In particular, the $\mathsf{G}_F(y', y)$ can be thought of as a higher-order analogue of the Bregman divergence $\mathsf{D}_F(y', y)$. Specifically, $\mathsf{G}_F(y', y)$ is the first-order error vector associated with approximating the difference of gradients $\nabla F(y') - \nabla F(y)$ by the matrix-vector product $\nabla^2 F(y)(y' - y)$.

Under the (LHS) property, we can then bound the norm of $\mathsf{G}_F(y', y)$ under the semi-norm $\|\cdot\|_{y,*}$. Formally, we prove the following:

**Lemma G.6.** *Let $y, y' \in \mathbb{R}^D$ such that $\|y' - y\|_* \leq \alpha$ for $\alpha \geq 0$. Let $\mathsf{G}_F(y', y)$ be the Hessian remainder term as defined in (78). Then it holds that $\mathsf{G}_F(y', y) \in \mathcal{S}^\perp$ and that*

$$\|\mathsf{G}_F(y', y)\|_{y,*} \leq \big( \tfrac{\exp(\widehat{\omega}\alpha) - 1 - \widehat{\omega}\alpha}{\widehat{\omega}\alpha} \big) \cdot \|y' - y\|_y \; .$$

**Proof.** The proof proceeds in several steps and reuses several calculations from the proof of Lemma G.5. We start by re-introducing several pieces of notation from that proof:

*1. Shorthand notation.* We let $P := \nabla^2 F(y)$, and let $P^{-1} = (\nabla^2 F(y))^{-1}$ when restricted to $\mathcal{S}^\perp$. For all $s \in [0,1]$, let $y_s := y + s(y' - y)$, let $P_s := \nabla^2 F(y_s)$, and let $B_s := P^{-1/2} P_s P^{-1/2}$.

*2. Properties of $B_s$ from Lemma G.5.* From Step 2 of the proof of Lemma G.5, recall that we established for all $s \in [0,1]$ that $\exp(-\widehat{\omega}\alpha s) I \preceq B_s \preceq \exp(\widehat{\omega}\alpha s) I$. This further implies for $B_s - I$ that:

$$(\exp(-\widehat{\omega}\alpha s) - 1) I \preceq B_s - I \preceq (\exp(\widehat{\omega}\alpha s) - 1) I \; .$$

As $\exp(-x) - 1 \geq 1 - \exp(x)$ for all $x \in \mathbb{R}$ (by the AM-GM inequality), we further have that $B_s - I \succeq 1 - \exp(\widehat{\omega}\alpha s)$. Thus, it follows by definition of spectral norm that $\|B_s - I\|_2 \leq \exp(\widehat{\omega}\alpha s) - 1$ for all $s \in [0,1]$.

*3. Expand $\mathsf{G}_F$ in integral form.* Then, using the definition of $\mathsf{G}_F(y', y)$ from (78), together with the expression for $\nabla F(y') - \nabla F(y)$ from (75) in the proof of Lemma G.5, it follows that

$$\mathsf{G}_F(y', y) = \int_0^1 (P_s - P)(y' - y) ds \; . \tag{79}$$

Now recall from Proposition C.5 (applied to $\nabla^2 F$) that $\mathrm{Range}(\nabla^2 F(y'')) = \mathcal{S}^\perp$ for all $y'' \in \mathbb{R}^D$. As $\mathcal{S}^\perp$ is a linear subspace, then also $(P_s - P)(y' - y) \in \mathcal{S}^\perp$ for all $s \in [0,1]$, and thus $\mathsf{G}_F(y', y) \in \mathcal{S}^\perp$. This establishes the first claim of the lemma.

*4. Bound on $\mathsf{G}_F$ in dual seminorm.* Using expression (79) and noting that we have shown that $\mathsf{G}_F(y', y) \in \mathcal{S}^\perp$, we can take the dual semi-norm $\|\cdot\|_{y,*}$ and apply the triangle inequality to write:

$$\|\mathsf{G}_F(y', y)\|_{y,*} \leq \int_0^1 \|(P_s - P)(y' - y)\|_{y,*} ds \; . \tag{80}$$

Then for each $s \in [0,1]$, we can use the definition of $B_s = P^{-1/2} P_s P^{-1/2}$ to further write

$$\begin{aligned}
\|(P_s - P)(y' - y)\|_{y,*} &= \|P^{-1/2}(P_s - P)(y' - y)\|_2 \\
&= \|(B_s P^{1/2} - P^{1/2})(y' - y)\|_2 \\
&= \|(B_s - I) P^{1/2}(y' - y)\|_2 \\
&\leq \|B_s - I\|_2 \cdot \|P^{1/2}(y' - y)\|_2 \leq (\exp(\widehat{\omega}\alpha s) - 1) \cdot \|y' - y\|_y \; .
\end{aligned} \tag{81}$$

Here, the final inequality uses the bound on $\|B_s - I\|_2$ from Step 2 and the definition of $\|\cdot\|_y$.

Substituting (81) into (80) and integrating then yields

$$\|\mathsf{G}_F(y', y)\|_{y,*} \leq \Big( \int_0^1 (\exp(\widehat{\omega}\alpha s) - 1) ds \Big) \cdot \|y' - y\|_y = \Big( \frac{\exp(\widehat{\omega}\alpha) - 1 - \widehat{\omega}\alpha}{\widehat{\omega}\alpha} \Big) \cdot \|y' - y\|_y \; ,$$

which concludes the proof. ■

### G.7 Upper Bound on Operator Differences

Under $\widehat{L}$-Lipschitzness of $V$, 1-strong-convexity of $R$, and using the dual-to-primal local semi-norm bound of Lemma G.5, we can derive the following key lemma that bounds the magnitude of operator differences $\|V(x') - V(x)\|_y$ in terms of the magnitude of dual increments $\|y' - y\|_y$. Formally:

**Lemma G.7.** *Fix $y, y' \in \mathbb{R}^D$ such that $\|y' - y\|_* \leq \alpha$ for $\alpha > 0$, and let $x = \nabla F(y) \in \mathrm{ri}(\mathcal{X})$ and $x' = \nabla F(y') \in \mathrm{ri}(\mathcal{X})$. Then it holds that*

$$\|V(x') - V(x)\|_y \leq \widehat{L} \cdot \big( \tfrac{\exp(\widehat{\omega}\alpha) - 1}{\widehat{\omega}\alpha} \big) \cdot \|y' - y\|_y \; .$$

**Proof.** Observe the following sequence of inequalities:

$$\begin{aligned}
\|V(x') - V(x)\|_y &\leq \|V(x') - V(x)\|_* && \text{(Lemma G.1, part (i))} \\
&\leq \widehat{L} \cdot \|x' - x\| && \text{(Lipschitzness of } V\text{)} \\
&\leq \widehat{L} \cdot \|x' - x\|_{y,*} && \text{(Lemma G.1, part (ii))} \\
&\leq \widehat{L} \cdot \big( \tfrac{\exp(\widehat{\omega}\alpha) - 1}{\widehat{\omega}\alpha} \big) \cdot \|y' - y\|_y && \text{(Lemma G.5) .}
\end{aligned}$$

Note that third inequality holds due to the fact that $x, x' \in \mathcal{X}$, which means $x' - x \in \mathcal{S}^\perp$ and allows for applying Lemma G.1, part (ii). This completes the proof. ■

## H Proof of Lemma 5.4: Energy Dissipation

In this section, we develop the proof of Lemma 5.4. This lemma establishes tight bounds on the one-step change in (shifted) energy dissipation for OFTRL and EGFTRL under the symmetrized system of Proposition 5.2 (as summarized in Setting D.1). We restate the proposition here:

**Lemma 5.4** (Dissipation of shifted energy)**.** *Let Assumptions 1, 2, 3, 4 hold. Let* $\{(x_t, y_t)\}$ *be the* OFTRL *or* EGFTRL *iterates from Proposition 5.2 with* $\eta \in (0, 1/(880LB\omega\kappa_M)]$*. Then for* $t \geq 1$*:*

$$-\tfrac{7}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \;\leq\; F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \;\leq\; -\tfrac{1}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 , \quad \forall\, x^\star \in \mathrm{MEQ}(V).$$

Before giving the proof of the lemma, we mention several further remarks:

**Remark H.1** (Assumptions and generality)**.** Note that the energy dissipation bounds of Lemma 5.4 *do not* require the game regularity condition of Assumption 5. Moreover, the bounds hold simultaneously for all points $x^\star \in \mathrm{MEQ}(V)$.

**Remark H.2** (Constant stepsize)**.** The tight dissipation bounds of Lemma 5.4 hold for a sufficiently small *constant* stepsize. In general, our proof does not attempt to optimize the final stepsize constraint (and the resulting absolute constant dependencies in the lemma statement), and thus the result likely also holds under larger constant stepsize settings. Note that within the stepsize constraint, the inverse dependencies on the Lipschitz and boundedness parameters of the operator are standard. On the other hand, the constraint also depends inversly on the ratio $\kappa_M = \frac{m_{\max}}{m_{\min}}$ for the weights matrix $M$. A similar constraint is required for the asymptotic result of Legacci et al. (2024a), and we leave the design and analysis of adaptive stepsize methods as an interesting direction for future work.

### Preliminaries and Recap of Setting

For convenience, we recall here some properties of the symmetrized system and dynamics. Specifically, for a weights matrix $M = \mathrm{Diag}(m_1 I_{d_1}, \dots, m_N I_{d_N}) \succ 0$, we consider the $M$-symmetrized dynamics from Proposition 5.2 using operator $V$, regularizer $R$, and energy $F = R^*$. Recall that we define $\kappa_M = m_{\max}/m_{\min}$.

**Definition of constants.** Recall that the Lipschitz and boundedness constants for the operator $G$ (from Assumption 1) are given by $L \geq 1$ and $B \geq 1$. Moreover, the (LHS) constant for the energy $H$ in Assumption 4 is given by $\omega \geq 1$. Then for the symmetrized system, we define the constants:

$$\widehat{L} := \sqrt{\kappa_M} L\,, \quad \widehat{B} := \sqrt{m_{\max}} B\,, \quad \widehat{\omega} := \tfrac{1}{\sqrt{m_{\min}}}\omega\,, \quad \widehat{\tau} := 3\widehat{B}\,. \tag{82}$$

**Properties of symmetrized system.** For the symmetrized system with operator $V$ regularizer $R$, recall that $V$ is $\widehat{L}$-Lipschitz and $\widehat{B}$-bounded (Setting D.1):

$$\|V(x') - V(x)\|_* \leq L\sqrt{\kappa_M} \cdot \|x' - x\|\,, \quad \|V(x)\|_* \leq \sqrt{m_{\max}} B \;\text{ for all } x, x' \in \mathcal{X}\,. \tag{83}$$

Moreover, the energy function $F = R^*$ satisfies the (LHS) with parameter $\widehat{\omega}$.

**Dual update rules.** For convenience, we also restate here the dual update rules of OFTRL and EGFTRL of the symmetrized system. For all $t \geq 0$, we have:

$$y_{t+1} = y_t - \eta 2V(x_t) + \eta V(x_{t-1})\,, \quad x_{t+1} = \nabla F(y_{t+1}) \tag{OFTRL}$$

$$\begin{cases} y_{t+1/2} = y_t - \eta V(x_t)\,, & x_{t+1/2} = \nabla F(y_{t+1/2}) \\ y_{t+1} = y_t - \eta V(x_{t+1/2})\,, & x_{t+1} = \nabla F(y_{t+1}) \end{cases} \tag{EGFTRL}$$

Recall from Setting D.1 that for both (OFTRL) and (EGFTRL), we have $\|y_{t+1} - y_t\|_* \leq \eta\widehat{\tau}$ for all $t \geq 0$ (and also $\|y_{t+1/2} - y_t\|_* \leq \eta\widehat{\tau}$ for (EGFTRL) for all $t \geq 0$).

**Stepsize constraint.** We assume the following bound on the stepsize $\eta$:

$$0 < \eta \leq (880\widehat{L}\widehat{B}\widehat{\omega})^{-1} = (880LB\omega\kappa_M)^{-1}\,. \tag{84}$$

Observe that (84) corresponds to the condition from the statement of Lemma 5.4.

### H.1 Proof of Lemma 5.4

We now give the proof of Lemma 5.4, which builds off the high-level overview from Section 5.1.

**Scalar functions.** For convenience, we first define the scalar functions $\phi(\cdot)$, $\mu(\cdot)$, $\nu(\cdot)$, and $\gamma(\cdot)$:

$$\phi(\alpha) = \exp(\widehat{\omega}\alpha)\,,\ \mu(\alpha) = \tfrac{\exp(\widehat{\omega}\alpha)-(\widehat{\omega}\alpha)-1}{(\widehat{\omega}\alpha)^2}\,,\ \nu(\alpha) = \tfrac{\exp(\widehat{\omega}\alpha)-1-\widehat{\omega}\alpha}{\widehat{\omega}\alpha}\,,\ \gamma(\alpha) = \tfrac{\exp(\widehat{\omega}\alpha)-1}{\widehat{\omega}\alpha}\,. \quad (85)$$

Each of these scalar functions corresponds to a multiplicative factor arising from the bounds developed in Section G, which hold for the energy $F$ satisfying (LHS) with parameter $\widehat{\omega}$. In particular, each function in (85) is non-negative and increasing in $\alpha$. Throughout the proof of the lemma, we will usually apply the functions of (85) with $\alpha = \eta\widehat{\tau}$, which stems from the upper bound $\|y_{t+1} - y_t\|_* \le \eta\widehat{\tau}$ that holds for the (OFTRL) and (EGFTRL) iterates.

**1. Coupling with implicit step.**

The first step of the proof involves setting up a general coupling between a sequence of FTRL-type iterates and an *implicit* update step:

**Proposition 5.5** (Implicit step coupling). *Let $\{(x_t, y_t)\}$ be a sequence satisfying $x_t = \nabla F(y_t)$ for all $t \ge 0$. Further, define the implicit update $\widehat{y}_{t+1} := y_t - \eta V(x_{t+1})$ for $t \ge 0$. Then for $t \ge 1$:*

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \;=\; \underbrace{-\mathsf{D}_F(y_t, y_{t+1})}_{\text{Dissipation term}} + \underbrace{\langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle}_{\text{Coupling error term } C_t}, \quad \forall\, x^\star \in \mathrm{MEQ}(V)\,.$$

To streamline the presentation, we give the proof of the proposition further below in Section H.1.1. It thus remains to instantiate Proposition 5.5 with the (OFTRL) and (EGFTRL) iterates and to control at each step the Bregman divergence $\mathsf{D}_F(y_t, y_{t+1})$ and the coupling error term $C_t := \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle$. We describe these steps below.

**2. Bounds on Bregman divergence terms.**

Suppose that $\|y_{t+1} - y_t\|_* \le \alpha$ for some $\alpha > 0$. First, we apply the two-sided bounds on the Bregman divergence $\mathsf{D}_F$ from Lemma G.4. Together, with an application of Lemma G.2, which transfers the local seminorm to the iterate $y_t$, we obtain the inequalities:

$$\mu(-\alpha)\phi(-\alpha)\cdot\|y_{t+1} - y_t\|_{y_t}^2 \;\le\; \mathsf{D}_F(y_t, y_{t+1}) \;\le\; \mu(\alpha)\phi(\alpha)\cdot\|y_{t+1} - y_t\|_{y_t}^2\,, \quad (86)$$

where $\mu(\cdot)$ and $\phi(\cdot)$ are as defined in (85). To instantiate (86) with the (OFTRL) and (EGFTRL) iterates, we thus require two-sided bounds on the magnitude of increments $\|y_{t+1} - y_t\|_{y_t}$ under the local semi-norm. For this, we prove, under the constraint on $\eta$ in (84), the following inequalities that relate the increments $y_{t+1} - y_t$ to the operator vector $V(x_t)$ in local semi-norm:

**Proposition H.3.** *For (OFTRL): $\frac{6}{7}\eta\|V(x_t)\|_{y_t} \le \|y_{t+1} - y_t\|_{y_t} \le \frac{6}{5}\eta\|V(x_t)\|_{y_t}$ for all $t \ge 1$.*

**Proposition H.8.** *For (EGFTRL): $\frac{5}{6}\eta\|V(x_t)\|_{y_t} \le \|y_{t+1} - y_t\|_{y_t} \le \frac{7}{6}\eta\|V(x_t)\|_{y_t}$ for all $t \ge 1$.*

We give the proofs of these propositions in Sections H.2.1 and H.3.1, respectively. Then, recall that $\|y_{t+1} - y_t\|_* \le \eta\widehat{\tau}$ for all $t \ge 1$ under both (OFTRL) and (EGFTRL) (where $\widehat{\tau} = 3\widehat{B}$, as in (82)), meaning that we can set $\alpha = \eta\widehat{\tau}$. Substituting these statements into (86) yields the following bounds:

- For (OFTRL), we obtain for all $t \ge 1$:

$$\tfrac{36}{49}\mu(-\eta\widehat{\tau})\phi(-\eta\widehat{\tau})\cdot\eta^2\|V(x_t)\|_{y_t}^2 \le \mathsf{D}_F(y_t, y_{t+1}) \le \tfrac{36}{25}\mu(\eta\widehat{\tau})\phi(\eta\widehat{\tau})\cdot\eta^2\|V(x_t)\|_{y_t}^2\,. \quad (87)$$

- For (EGFTRL), we obtain for all $t \ge 1$:

$$\tfrac{25}{36}\mu(-\eta\widehat{\tau})\phi(-\eta\widehat{\tau})\cdot\eta^2\|V(x_t)\|_{y_t}^2 \le \mathsf{D}_F(y_t, y_{t+1}) \le \tfrac{49}{36}\mu(\eta\widehat{\tau})\phi(\eta\widehat{\tau})\cdot\eta^2\|V(x_t)\|_{y_t}^2\,. \quad (88)$$

Now, the product of scalar functions $\mu(-\eta\widehat{\tau})\cdot\phi(-\eta\widehat{\tau})$ is decreasing with $\eta$ and can be made a sufficiently large positive constant bounded above by $1/2$ when $\eta$ is sufficiently small. Similarly, $\mu(\eta\widehat{\tau})\cdot\phi(\eta\widehat{\tau})$ is increasing with $\eta$ and be made an arbitrarily small positive constant bounded below by $1/2$ for sufficiently small $\eta$. In particular, for $\eta$ as in (84), it is straightforward to check that

$$\mu(-\eta\widehat{\tau})\cdot\phi(-\eta\tau) \ge \tfrac{49}{50} \;\text{ and }\; \mu(\eta\widehat{\tau})\cdot\phi(\eta\tau) \le \tfrac{51}{100}\,. \quad (89)$$

**3. Bounds on coupling error terms.**

Recall that $C_t = \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1} \rangle$ denotes the coupling error term at each step $t \geq 1$. For the iterates of (OFTRL) and (EGFTRL), we obtain the following bounds for $\eta$ as in (84):

**Proposition H.4** *For* (OFTRL)*, for all* $x^\star \in \mathrm{MEQ}(V)$ *and* $t \geq 1$*:*

$$|C_t| \leq \tfrac{24}{5} \cdot \nu(\eta\widehat{\tau}) \cdot \eta^2 \|V(x_t)\|_{y_t}^2 + 110\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 \,. \tag{90}$$

**Proposition H.9** *For* (EGFTRL)*, for all* $x^\star \in \mathrm{MEQ}(V)$ *and* $t \geq 0$*:*

$$|C_t| \leq 9\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 \,. \tag{91}$$

We give the proofs of Propositions H.4 and H.9 in Sections H.2.2 and H.3.2, respectively. Similar to the prior step, we note that $\nu(\eta\widehat{\tau})$ is non-negative and strictly increasing in $\eta$ and can be made an arbitrarily small positive constant for $\eta$ sufficiently small. Moreover, the $O(\eta^3)$ terms in Propositions H.4 and H.9 can also be made small positive constants for $\eta$ sufficiently small. In particular, for $\eta$ as in (84), we have the bounds:

$$\nu(\eta\widehat{\tau}) \leq \tfrac{1}{500} \quad \text{and} \quad \eta \cdot 110\widehat{L} \leq \tfrac{1}{8} \,. \tag{92}$$

**Combining the pieces: upper bound.** To derive the upper bound of Lemma 5.4, fix $x^\star \in \mathrm{MEQ}(V)$ and $t \geq 1$. Then we combine the lower bounds on $\mathsf{D}_F(y_t, y_{t+1})$ from (87) and (88) and the upper bounds on $|C_t|$ from Proposition H.4 and H.9. Taking the worse bound of the two instances, this yields for both (OFTRL) and (EGFTRL):

$$\begin{aligned} F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) &= -\mathsf{D}_F(y_t, y_{t+1}) + C_t \\ &\leq -\big(\tfrac{25}{36}\mu(-\eta\widehat{\tau})\phi(-\eta\widehat{\tau}) - \tfrac{24}{5}\nu(\eta\widehat{\tau})\big) \cdot \eta^2 \|V(x_t)\|_{y_t}^2 + 110\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 \\ &\leq -\tfrac{1}{4} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 + \tfrac{1}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \\ &= -\tfrac{1}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \,. \end{aligned}$$

Here, the penultimate inequality comes from substituting the lower bound on $\mu(-\eta\widehat{\tau}) \cdot \phi(-\eta\tau)$ from (89) and the upper bounds of (92), both of which hold for $\eta$ as in (84). Thus, for such sufficiently small constant $\eta$, the negative second order $\mathsf{D}_F(y_t, y_{t+1})$ term dominates the coupling error and establishes a dissipative one-step upper bound on $F_{x^\star}$.

**Combining the pieces: lower bound.** For the lower bound of Lemma 5.4, we now use the upper bounds on the $\mathsf{D}_F(y_t, y_{t+1})$ terms from (87) and (88). Then for any $x^\star \in \mathrm{MEQ}(V)$ and $t \geq 1$, we similarly find for both (OFTRL) and (EGFTRL) that:

$$\begin{aligned} F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) &= -\mathsf{D}_F(y_t, y_{t+1}) + C_t \\ &\geq -\mathsf{D}_F(y_t, y_{t+1}) - |C_t| \\ &\geq -\big(\tfrac{36}{25}\mu(\eta\widehat{\tau})\phi(\eta\widehat{\tau}) + \tfrac{24}{5}\nu(\eta\widehat{\tau})\big) \cdot \eta^2 \|V(x_t)\|_{y_t}^2 - 110\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 \\ &\geq -\tfrac{3}{4} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 - \tfrac{1}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \\ &= -\tfrac{7}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \,. \end{aligned}$$

Here, the penultimate inequality now follows from the upper bound on $\mu(\eta\widehat{\tau}) \cdot \phi(\eta\widehat{\tau})$ from (89) (and again using (92)), which holds for $\eta$ as in (84). This completes the proof of the lemma. ■

### H.1.1 Proof of Proposition 5.5

Here we give the proof of the implicit-step coupling of Proposition 5.5:

**Proof.** The proof relies on two steps: (1) expressing the difference in shifted energy in terms of the (unshifted) Bregman divergence $\mathsf{D}_F(y_t, y_{t+1})$ and (2) coupling the true iterate $y_{t+1}$ with the implicit-step iterate $\widehat{y}_{t+1}$. We proceed with these two steps:

*Step (1).* Fix $x^\star \in \mathrm{MEQ}(V)$ and $t \geq 1$, and observe for any $t \geq 1$ that

$$\begin{aligned} F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) &= F(y_{t+1}) - F(y_t) - \langle x^\star, y_{t+1} - y_t \rangle \\ &= \langle \nabla F(y_{t+1}), y_{t+1} - y_t \rangle - \mathsf{D}_F(y_t, y_{t+1}) - \langle x^\star, y_{t+1} - y_t \rangle \\ &= \langle x_{t+1}, y_{t+1} - y_t \rangle - \mathsf{D}_F(y_t, y_{t+1}) - \langle x^\star, y_{t+1} - y_t \rangle \\ &= -\mathsf{D}_F(y_t, y_{t+1}) + \langle x_{t+1} - x^\star, y_{t+1} - y_t \rangle \ . \end{aligned} \tag{93}$$

Here, the first equality comes from the definition of $\mathsf{D}_F(y_t, y_{t+1})$ and the second equality comes from the assumption that $x_{t+1} = \nabla F(y_{t+1})$ for all $t \geq 0$.

*Step (2).* To further simplify (93), we use the implicit update $\widehat{y}_{t+1} = y_t - \eta V(x_{t+1})$ to write:

$$y_{t+1} - y_t = \widehat{y}_{t+1} - y_t + y_{t+1} - \widehat{y}_{t+1} = -\eta V(x_{t+1}) + y_{t+1} - \widehat{y}_{t+1} \ . \tag{94}$$

Then it follows that

$$\begin{aligned} \langle x_{t+1} - x^\star, y_{t+1} - y_t \rangle &= -\eta \langle x_{t+1} - x^\star, V(x_{t+1}) \rangle + \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1} \rangle \\ &= \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1} \rangle \ . \end{aligned} \tag{95}$$

Here the last equality follows due to the fact that $x^\star \in \mathrm{MEQ}(V)$, and thus $\langle x_{t+1} - x^\star, V(x_{t+1}) \rangle = 0$. Substituting (95) into (93) then yields the desired statement. ■

### H.2 Instantiating Implicit Step Coupling for OFTRL

In this section, we provide details on the instantiation of Lemma 5.4 along (OFTRL), as outlined in the proof in Section H.1. This requires proving two main propositions, which we state here.

First, the following proposition relates the dual increments $y_{t+1} - y_t$ to the operator value $V(x_t)$:

**Proposition H.3.** *For* (OFTRL) *with* $\eta > 0$ *as in* (84)*, it holds for all* $t \geq 1$*:*

$$\tfrac{6}{7}\eta \|V(x_t)\|_{y_t} \leq \|y_{t+1} - y_t\|_{y_t} \leq \tfrac{6}{5}\eta \|V(x_t)\|_{y_t} \ .$$

Second, we establish bounds on the coupling error terms $C_t := \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1} \rangle$ that hold for all $x^\star \in \mathrm{MEQ}(V)$. Formally we prove the following:

**Proposition H.4.** *For* (OFTRL) *with* $\eta > 0$ *as in* (84)*, it holds for all* $t \geq 1$ *and* $x^\star \in \mathrm{MEQ}(V)$*:*

$$|C_t| \leq \tfrac{24}{5}\nu(\eta\widehat{\tau}) \cdot \eta^2 \|V(x_t)\|_{y_t}^2 + 110\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 \ .$$

The proofs of Proposition H.3 and Proposition H.4 are given below in Section H.2.1 and Section H.2.2, respectively. Both proofs additionally use the following helper lemma, which establishes the *stability* of the OFTRL dual iterates in local seminorm. Formally:

**Lemma H.5** (OFTRL dual stability)**.** *Under* (OFTRL) *with* $\eta$ *satisfying* (84)*, for all* $t \geq 2$*:*

$$\tfrac{1}{2} \cdot \|y_{t-1} - y_{t-2}\|_{y_t} \ \leq \ \|y_t - y_{t-1}\|_{y_t} \ \leq \ 3 \cdot \|y_{t+1} - y_t\|_{y_t} \ .$$

We will grant this result as true for now and defer its proof to Section H.2.3.

#### H.2.1 Proof of Proposition H.3

This section gives the proof of Proposition H.3, which relates the magnitude of increments $y_{t+1} - y_t$ to the magnitude of the operator vector $V(x_t)$ under the local semi-norm $\|\cdot\|_{y_t}$.

**Proof.** First, recall the scalar function $\gamma(\cdot)$ from (85) and the constant $\widehat{\tau}$ from (82). In particular, recall for (OFTRL) that $\|y_t - y_{t-1}\|_* \leq \eta\widehat{\tau}$ for all $t \geq 2$. Moreover, it is straightforward to check that for $\eta$ as in (84), we have $3\eta\widehat{L}\gamma(\eta\widehat{\tau}) \leq \frac{1}{6}$. We then prove the lower and upper inequalities of the proposition separately:

**Upper inequality.** By definition of (OFTRL) and using the triangle inequality with respect to the local norm $\|\cdot\|_{y_t}$, we obtain the bound

$$\begin{aligned} \|y_{t+1} - y_t\|_{y_t} &\leq \eta \|V(x_t)\|_{y_t} + \eta \|V(x_t) - V(x_{t-1})\|_{y_t} \\ &\leq \eta \|V(x_t)\|_{y_t} + \eta\widehat{L}\gamma(\eta\widehat{\tau}) \cdot \|y_t - y_{t-1}\|_{y_t} \qquad \text{(Lemma G.7)} \\ &\leq \eta \|V(x_t)\|_{y_t} + 3\eta\widehat{L}\gamma(\eta\widehat{\tau}) \cdot \|y_{t+1} - y_t\|_{y_t} \qquad \text{(Lemma H.5)} \ . \end{aligned}$$

Rearranging and using the fact that $3\eta\widehat{L}\gamma(\eta\widehat{\tau}) < 1/6$, we conclude that

$$\|y_{t+1} - y_t\|_{y_t} \leq \tfrac{1}{1 - 3\eta\widehat{L}\gamma(\eta\widehat{\tau})} \cdot \eta \|V(x_t)\|_{y_t} \leq \tfrac{6}{5} \cdot \eta \|V(x_t)\|_{y_t} \ .$$

**Lower inequality.** For the lower inequality, we again use the definition of (OFTRL) and the triangle inequality to write

$$\begin{aligned}\eta\|V(x_t)\|_{y_t} &\le \|y_t - y_{t+1}\|_{y_t} + \eta\|V(x_t) - V(x_{t-1})\|_{y_t} \\ &\le \|y_t - y_{t+1}\|_{y_t} + \widehat{L}\gamma(\eta\widehat{\tau})\eta \cdot \|y_t - y_{t-1}\|_{y_t} && \text{(Lemma G.7)} \\ &\le \|y_t - y_{t+1}\|_{y_t} + 3\widehat{L}\gamma(\eta\widehat{\tau})\eta \cdot \|y_{t+1} - y_t\|_{y_t} && \text{(Lemma H.5)} \\ &\le \tfrac{7}{6} \cdot \|y_{t+1} - y_t\|_{y_t} && (3\eta\widehat{L}\gamma(\eta\widehat{\tau}) \le 1/6)\ .\end{aligned}$$

Rearranging then yields the desired lower inequality and concludes the proof. ■

### H.2.2 Proof of Proposition H.4

This section gives the proof of Proposition H.4, which establishes a bound on the coupling error terms $C_t$ from Proposition 5.5 along the (OFTRL) iterates:

**Proof.** The proof proceeds in two steps: first, deriving an exact expansion of the terms $C_t$ into four constituent *error terms*, and second, deriving bounds on the absolute magnitudes of the error terms.

**1. Expanding the inner product term.**

First, recall that $\widehat{y}_{t+1} = y_t - \eta V(x_{t+1})$. Then under (OFTRL), it follows that we can express the difference $y_{t+1} - \widehat{y}_{t+1}$ as

$$\begin{aligned}y_{t+1} - \widehat{y}_{t+1} &= y_t - 2\eta V(x_t) + \eta V(x_{t-1}) - (y_t - \eta V(x_{t+1})) \\ &= \eta(V(x_{t+1}) - 2V(x_t) + V(x_{t-1}))\ .\end{aligned}$$

Now fix $x^\star$ satisfying (MEQ). Then we can expand $C_t$ as follows:

$$\begin{aligned}C_t := \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle &= \eta\big\langle x_{t+1} - x^\star, V(x_{t+1}) - 2V(x_t) + V(x_{t-1})\big\rangle \\ &= -2\eta\big\langle x_{t+1} - x^\star, V(x_t)\big\rangle + \eta\big\langle x_{t+1} - x^\star, V(x_{t-1})\big\rangle\,, \qquad (96)\end{aligned}$$

where the final equality follows from observing that $\langle x_{t+1} - x^\star, V(x_{t+1})\rangle = 0$ since $x^\star$ satisfies (MEQ). Both terms in (96) can then further be expanded as:

$$\begin{aligned}-2\eta\big\langle x_{t+1} - x^\star, V(x_t)\big\rangle &= -2\eta\big\langle x_t - x^\star + (x_{t+1} - x_t), V(x_t)\big\rangle \\ &= -2\eta\big\langle x_{t+1} - x_t, V(x_t)\big\rangle\,, \qquad (97)\end{aligned}$$

$$\begin{aligned}\eta\big\langle x_{t+1} - x^\star, V(x_{t-1})\big\rangle &= \eta\big\langle x_{t-1} - x^\star + (x_{t+1} - x_{t-1}), V(x_{t-1})\big\rangle \\ &= \eta\big\langle x_{t+1} - x_{t-1}, V(x_{t-1})\big\rangle \\ &= \eta\big\langle x_{t+1} - x_{t-1}, V(x_{t-1}) - V(x_t)\big\rangle + \eta\big\langle x_{t+1} - x_{t-1}, V(x_t)\big\rangle. \qquad (98)\end{aligned}$$

In the expansion of both terms, the second equalities follow from the fact that $x^\star$ satisfies (MEQ). Substituting (97) and (98) into (96) and simplifying then yields

$$C_t = \underbrace{\eta\big\langle V(x_t), x_t - x_{t-1} + x_t - x_{t+1}\big\rangle}_{\text{Term I}} + \eta\big\langle x_{t+1} - x_{t-1}, V(x_{t-1}) - V(x_t)\big\rangle\ . \qquad (99)$$

We then further simplify the first term of (99) as follows:

*Expansion of Term I.* First recall the primal-dual relationships $x_{t-1} = \nabla F(y_{t-1})$, $x_t = \nabla F(y_t)$, and $x_{t+1} = \nabla F(y_{t+1})$. Using the definition of the Hessian remainder terms $\mathsf{G}_F(y_{t+1}, y_t)$ and $\mathsf{G}_F(y_{t-1}, y_t)$, it follows that

$$\begin{aligned}&x_t - x_{t-1} + x_t - x_{t+1} \\ &\quad= \nabla F(y_t) - \nabla F(y_{t-1}) + \nabla F(y_t) - \nabla F(y_{t+1}) \\ &\quad= \nabla^2 F(y_t)(y_t - y_{t-1}) - \mathsf{G}_F(y_{t-1}, y_t) + \nabla^2 F(y_t)(y_t - y_{t+1}) - \mathsf{G}_F(y_{t+1}, y_t) \\ &\quad= \nabla^2 F(y_t)(y_t - y_{t-1} + y_t - y_{t+1}) - \mathsf{G}_F(y_{t-1}, y_t) - \mathsf{G}_F(y_{t+1}, y_t)\ . \qquad (100)\end{aligned}$$

By definition of (OFTRL), observe that

$$\begin{aligned}y_t - y_{t-1} + y_t - y_{t+1} &= -2\eta V(x_{t-1}) + \eta V(x_{t-2}) - \eta V(x_{t-1}) + 2\eta V(x_t) \\ &= 2\eta\big(V(x_t) - V(x_{t-1})\big) + \eta\big(V(x_{t-2}) - V(x_{t-1})\big)\ . \qquad (101)\end{aligned}$$

Substituting (101) into (100) then allows for expanding (Term I) in (99) as

$$\text{(Term I)} = 2\eta^2\big\langle V(x_t), V(x_t) - V(x_{t-1})\big\rangle_{y_t} + \eta^2\big\langle V(x_t), V(x_{t-2}) - V(x_{t-1})\big\rangle_{y_t} - \eta\big\langle V(x_t), \mathsf{G}_F(y_{t+1}, y_t) + \mathsf{G}_F(y_{t-1}, y_t)\big\rangle \,. \quad (102)$$

*Final expansion of $C_t$.* Finally, substituting (102) into (99) yields the full expansion

$$\begin{aligned} C_t = &\underbrace{-\eta\big\langle V(x_t), \mathsf{G}_F(y_{t+1}, y_t) + \mathsf{G}_F(y_{t-1}, y_t)\big\rangle}_{\mathcal{E}_{t,1}} \\ &+ \underbrace{2\eta^2\big\langle V(x_t), V(x_t) - V(x_{t-1})\big\rangle_{y_t}}_{\mathcal{E}_{t,2}} + \underbrace{\eta^2\big\langle V(x_t), V(x_{t-2}) - V(x_{t-1})\big\rangle_{y_t}}_{\mathcal{E}_{t,3}} \\ &+ \underbrace{\eta\big\langle x_{t+1} - x_{t-1}, V(x_{t-1}) - V(x_t)\big\rangle}_{\mathcal{E}_{t,4}} \,. \end{aligned} \quad (103)$$

**2. Bounds on magnitude of coupling error terms.**

We proceed to derive upper bounds on the magnitude of the error terms $\mathcal{E}_{t,1}$, $\mathcal{E}_{t,2}$, $\mathcal{E}_{t,3}$, and $\mathcal{E}_{t,4}$ that comprise the exact expansion of the inner product $C_t = \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle$ from (103). We assume $\eta$ is constrained as in (84) and we also use the definitions of the scalar functions $\nu(\cdot)$ and $\gamma(\cdot)$ from (85).

**Bound on second-order error terms.** The first term $\mathcal{E}_{t,1}$ scales like $O(\eta^2\|V(x_t)\|_{y_t}^2)$, and with a leading constant that can be made arbitrarily small for $\eta$ sufficiently bounded. In particular, by an application of the generalized Cauchy-Schwarz inequality with respect to the local norm $\|\cdot\|_{y_t}$, we can write and further simplify

$$\begin{aligned} |\mathcal{E}_{t,1}| &= \big|\eta\big\langle V(x_t), \mathsf{G}_F(y_{t+1}, y_t) + G_F(y_{t-1}, y_t)\big\rangle\big| \\ &\le \eta\|V(x_t)\|_{y_t} \cdot \|\mathsf{G}_F(y_{t+1}, y_t) + \mathsf{G}_F(y_{t-1}, y_t)\|_{y_t,*} \\ &\le \eta\|V(x_t)\|_{y_t} \cdot \big(\|\mathsf{G}_F(y_{t+1}, y_t)\|_{y_t,*} + \|\mathsf{G}_F(y_t, y_{t-1})\|_{y_t,*}\big) \\ &\le \eta\|V(x_t)\|_{y_t} \cdot \nu(\eta\widehat{\tau}) \cdot \big(\|y_{t+1} - y_t\|_{y_t} + \|y_t - y_{t-1}\|_{y_t}\big) && \text{(Lemma G.6)} \\ &\le \eta\|V(x_t)\|_{y_t} \cdot 4\nu(\eta\widehat{\tau}) \cdot \|y_{t+1} - y_t\|_{y_t} && \text{(Lemma H.5)} \\ &\le \frac{24}{5}\nu(\eta\widehat{\tau}) \cdot \eta^2\|V(x_t)\|_{y_t}^2 \,. && \text{(Proposition H.3)} \end{aligned}$$

**Bounds on third-order error terms.** We now derive bounds on the magnitudes of $\mathcal{E}_{t,2}$, $\mathcal{E}_{t,3}$, and $\mathcal{E}_{t,4}$ from (103). In particular, we show each of these three terms scales like $O(\eta^3\|V(x_t)\|_{y_t}^2)$.

For the term $\mathcal{E}_{t,2}$, we can apply the Cauchy-Schwarz inequality with respect to the local induced norm $\|\cdot\|_{y_t}$ and Lemma G.7 and further simplify to find

$$\begin{aligned} |\mathcal{E}_{t,2}| &= \big|2\eta^2\big\langle V(x_t), V(x_t) - V(x_{t-1})\big\rangle_{y_t}\big| \\ &\le 2\eta^2\|V(x_t)\|_{y_t} \cdot \|V(x_t) - V(x_{t-1})\|_{y_t} \\ &\le 2\eta^2\|V(x_t)\|_{y_t} \cdot \widehat{L}\gamma(\eta\widehat{\tau}) \cdot \|y_t - y_{t-1}\|_{y_t} && \text{(Lemma G.7)} \\ &\le 2\eta^2\|V(x_t)\|_{y_t} \cdot 3\widehat{L}\gamma(\eta\widehat{\tau}) \cdot \|y_{t+1} - y_t\|_{y_t} && \text{(Lemma H.5)} \\ &\le \frac{36}{5}\widehat{L}\gamma(\eta\widehat{\tau}) \cdot \eta^3\|V(x_t)\|_{y_t}^2 \,. && \text{(Proposition H.3)} \end{aligned}$$

For the term $\mathcal{E}_{t,3}$, we use identical calculations as for $\mathcal{E}_{t,2}$, but now the application of Lemma H.5 gives $\|y_{t-1} - y_{t-2}\|_{y_t} \le 6\|y_{t+1} - y_t\|_{y_t}$. Together with the fact that $\mathcal{E}_{t,3}$ does not contain leading multiplicative factor 2, we obtain the same bound

$$|\mathcal{E}_{t,3}| = \big|\eta^2\big\langle V(x_t), V(x_{t-2}) - V(x_{t-1})\big\rangle_{y_t}\big| \le \tfrac{36}{5}\widehat{L}\gamma(\eta\widehat{\tau}) \cdot \eta^3\|V(x_t)\|_{y_t}^2 \,.$$

For the term $\mathcal{E}_{t,4}$, we use generalized Cauchy-Schwarz with respect to the local norm $\|\cdot\|_{y_t}$, together with applications of Lemma G.5 and Lemma G.7, to obtain

$$\begin{aligned}
|\mathcal{E}_{t,4}| &= \big|\eta\langle x_{t+1} - x_{t-1}, V(x_t) - V(x_{t-1})\rangle\big| \\
&= \eta\big|\langle x_{t+1} - x_t + x_t - x_{t-1}, V(x_t) - V(x_{t-1})\rangle\big| \\
&\le \eta \cdot \big(\|x_{t+1} - x_t\|_{y_t,*} + \|x_t - x_{t-1}\|_{y_t,*}\big) \cdot \|V(x_t) - V(x_{t-1})\|_{y_t} && \text{(Generalized C-S)} \\
&\le \eta\gamma(\eta\widehat{\tau}) \cdot \big(\|y_{t+1} - y_t\|_{y_t} + \|y_t - y_{t-1}\|_{y_t}\big) \cdot \widehat{L}\gamma(\eta\widehat{\tau})\|y_t - y_{t-1}\|_{y_t} && \text{(Lemmas G.5 and G.7)} \\
&\le 12\widehat{L}\gamma(\eta\widehat{\tau})^2\eta \cdot \|y_{t+1} - y_t\|_{y_t}^2 && \text{(Lemma H.5)} \\
&\le \frac{432}{5}\widehat{L}\gamma(\eta\widehat{\tau})^2\eta^3\|V(x_t)\|_{y_t}^2 \,. && \text{(Proposition H.3)}
\end{aligned}$$

Combining the bounds on the third-order terms $|\mathcal{E}_{t,2}|$, $|\mathcal{E}_{t,3}|$ and $|\mathcal{E}_{t,4}|$, and also using the fact that $\gamma(\eta\widehat{\tau}) \ge 1$ for all $\eta > 0$, we then have that

$$|\mathcal{E}_{t,2}| + |\mathcal{E}_{t,3}| + |\mathcal{E}_{t,4}| \le \tfrac{504}{5}\widehat{L}\gamma(\eta\widehat{\tau})^2 \cdot \eta^3\|V(x_t)\|_{y_t}^2 \le 110\widehat{L}\eta^3\|V(x_t)\|_{y_t}^2 \,.$$

Here, the final inequality comes from the fact that $\frac{504}{5} \cdot \gamma(\eta\widehat{\tau})^2 \le 110$ for $\eta$ as in (84).

**3. Combining the pieces.**

Together, for $\eta$ as in (84), we conclude that the magnitude of $C_t$ is bounded above by

$$|C_t| \le |\mathcal{E}_{t,1}| + |\mathcal{E}_{t,2}| + |\mathcal{E}_{t,3}| + |\mathcal{E}_{t,4}| \le \tfrac{24}{5}\nu(\eta\widehat{\tau}) \cdot \eta^2\|V(x_t)\|_{y_t}^2 + 110\widehat{L} \cdot \eta^3\|V(x_t)\|_{y_t}^2 \,,$$

which completes the proof of the proposition. ■

### H.2.3 Proof of Lemma H.5

Here, we prove Lemma H.5, which establishes the stability of the dual iterates of (OFTRL). Note that this lemma is the analog of (Lazarsfeld et al., 2026, Proposition F.6); but the present proof is generalized to handle general non-linear operators.

**Proof.** For readability, we will use the following shorthand notation: for $t \ge 2$, define

$$\Delta_t := y_t - y_{t-1} = \begin{cases} -\eta V(x_0) & \text{for } t = 1 \\ -2\eta V(x_{t-1}) + \eta V(x_{t-2}) & \text{for } t \ge 2 \,. \end{cases}$$

Thus, our goal is to prove for all $t \ge 2$ that

$$\text{(i)}\ \ \|\Delta_{t-1}\|_{y_t} \le 2\|\Delta_t\|_{y_t} \quad \text{and} \quad \text{(ii)}\ \ \|\Delta_t\|_{y_t} \le 3 \cdot \|\Delta_{t+1}\|_{y_t} \,. \tag{104}$$

We start by proving inequality (i) from which inequality (ii) will follow.

**Proof of inequality (i).** First, observe by the triangle inequality that

$$\|\Delta_{t-1}\|_{y_t} = \|\Delta_t + (\Delta_{t-1} - \Delta_t)\|_{y_t} \le \|\Delta_t\|_{y_t} + \|\Delta_t - \Delta_{t-1}\|_{y_t} \,. \tag{105}$$

Now if we can establish for some $B > 0$ that

$$\|\Delta_t - \Delta_{t-1}\|_{y_t} \le \eta C\|\Delta_{t-1}\|_{y_t} \tag{106}$$

then assuming $\eta < 1/(2C)$, it will follow from (105) that

$$\|\Delta_{t-1}\|_{y_t} \le \frac{1}{1 - \eta C} \cdot \|\Delta_t\|_{y_t} \le 2\|\Delta_t\|_{y_t} \,. \tag{107}$$

We will prove the inequality of (106) via induction on $t \ge 2$. For this, we first define several constants:

**Definition of constants.** Fix $\eta > 0$ satisfying (84). For the remainder of the proof, we will slightly overload notation and let $\phi$, $\gamma$, and $B$ be the ($\eta$-dependent) constants

$$\phi := \exp(\eta\widehat{\omega}\widehat{\tau}) \ , \quad \gamma := \tfrac{\exp(3\eta\widehat{\omega}\widehat{\tau})-1}{3\eta\widehat{\omega}\widehat{\tau}} \ , \quad \text{and} \ \ C := 4\widehat{L}\phi^4\gamma \ . \tag{108}$$

Here, $\widehat{\omega} > 0$, $\widehat{\tau} > 0$, and $\widehat{L} \geq 1$ are the constants from (82), where we recall that $\|y_t - y_{t-1}\|_* \leq \eta\widehat{\tau}$ for all $t \geq 0$. Further observe for all $\eta > 0$ that both $\phi > 0$ and $\gamma > 0$ are positive and increasing in $\eta$. Moreover, for $\eta$ as in (84), it is also straightforward to check that $\eta C \leq 1/2$.

We now proceed with the inductive proof of expression (106) with $C$ as defined in (108):

**Base case.** At $t = 2$, we have by definition that

$$\Delta_2 - \Delta_1 = -2\eta V(x_1) + \eta V(x_0) - (-\eta V(x_0)) = -2\eta(V(x_1) - V(x_0)) \ . \tag{109}$$

Then taking local seminorms and applying Lemma G.7, we have by definition of $\gamma$ that

$$\|\Delta_2 - \Delta_1\|_{y_1} = 2\eta\|V(x_1) - V(x_0)\|_{y_1} \leq 2\eta \cdot \hat{L}\gamma \cdot \|y_1 - y_0\|_{y_1} \ . \tag{110}$$

Then, using two applications of Lemma G.2, we conclude that

$$\|\Delta_2 - \Delta_1\|_{y_2} \leq \phi \cdot \|\Delta_2 - \Delta_1\|_{y_1} \leq \eta \cdot 2L\phi\gamma \cdot \|y_1 - y_0\|_{y_1} \leq \eta \cdot 2L\phi^2\gamma \cdot \|y_1 - y_0\|_{y_2} \ . \tag{111}$$

It then follows by definition of $\Delta_1$ and using the fact that $\widehat{L}, \phi, \gamma > 0$ that

$$\|\Delta_2 - \Delta_1\|_{y_2} \leq \eta \cdot 2\widehat{L}\phi^2\gamma \cdot \|\Delta_1\|_{y_2} \leq \eta \cdot 4\widehat{L}\phi^4\gamma \cdot \|\Delta_1\|_{y_2} \ . \tag{112}$$

**Inductive case.** Now assume for all $2 \leq k \leq t-1$ that

$$\|\Delta_k - \Delta_{k-1}\|_{y_k} \leq \eta \cdot 4\widehat{L}\phi^4\gamma \cdot \|\Delta_{k-1}\|_{y_k} \ .$$

Our goal is to establish this inequality also holds at time $t$. For this observe by definition that

$$\begin{aligned}\Delta_t - \Delta_{t-1} &= -2\eta V(x_{t-1}) + \eta V(x_{t-2}) - (-2\eta V(x_{t-2}) + \eta V(x_{t-3})) \\ &= -2\eta(V(x_{t-1}) - V(x_{t-2})) + \eta(V(x_{t-2}) - V(x_{t-3})) \ .\end{aligned}$$

Then taking local semi-norms and using the triangle inequality, we can further write and bound:

$$\begin{aligned}\|\Delta_t - \Delta_{t-1}\|_{y_t} &\leq 2\eta\|V(x_{t-1}) - V(x_{t-2})\|_{y_t} + \eta\|V(x_{t-2}) - V(x_{t-3})\|_{y_t} \\ &\leq 2\eta\phi\|V(x_{t-1}) - V(x_{t-2})\|_{y_{t-1}} + \eta\phi^2\|V(x_{t-2}) - V(x_{t-3})\|_{y_{t-2}} \quad \text{(Lemma G.2)} \\ &\leq 2\eta\phi\widehat{L}\gamma\|y_{t-1} - y_{t-2}\|_{y_{t-1}} + \eta\phi^2\widehat{L}\gamma\|y_{t-2} - y_{t-3}\|_{y_{t-2}} \quad \text{(Lemma G.7)} \\ &= \eta 2\widehat{L}\phi\gamma\|\Delta_{t-1}\|_{y_{t-1}} + \eta\widehat{L}\phi^2\gamma\|\Delta_{t-2}\|_{y_{t-2}} \\ &\leq \eta \cdot 2\widehat{L}\phi^2\gamma\|\Delta_{t-1}\|_{y_t} + \eta L\phi^3\gamma\|\Delta_{t-2}\|_{y_{t-1}} \ . \quad \text{(Lemma G.2)}\end{aligned} \tag{113}$$

Now recall by the inductive hypothesis that $\|\Delta_{t-1} - \Delta_{t-2}\|_{y_{t-1}} \leq \eta 4\widehat{L}\phi^4\gamma\|\Delta_{t-2}\|_{y_{t-1}}$. By the triangle inequality, it then follows that

$$\|\Delta_{t-2}\|_{y_{t-1}} \leq \|\Delta_{t-1}\|_{y_{t-1}} + \|\Delta_{t-1} - \Delta_{t-2}\|_{y_{t-1}} \leq \|\Delta_{t-1}\|_{y_{t-1}} + \eta \cdot 4\widehat{L}\phi^4\gamma \cdot \|\Delta_{t-2}\|_{y_{t-1}} \ .$$

Recall from (108) that we let $C := 4\widehat{L}\phi^4\gamma$. Then rearranging and using the fact that $C\eta \leq 1/2$ under the constraint on $\eta$, we then obtain the bound

$$\|\Delta_{t-2}\|_{y_{t-1}} \leq \frac{1}{1-\eta C} \cdot \|\Delta_{t-1}\|_{t-1} \leq \frac{\phi}{1-\eta C} \cdot \|\Delta_{t-1}\|_t \leq 2\phi \cdot \|\Delta_{t-1}\|_t \ , \tag{114}$$

where the second inequality follows from Lemma G.2. Substituting (114) into (113) then yields

$$\|\Delta_t - \Delta_{t-1}\|_{y_t} \leq \eta \cdot (2\widehat{L}\phi^2\gamma + 2\widehat{L}\phi^4\gamma) \cdot \|\Delta_{t-1}\|_t \leq \eta \cdot 4\widehat{L}\phi^4\gamma \cdot \|\Delta_{t-1}\|_t \ , \tag{115}$$

where the final inequality is due to the fact that $\widehat{L}, \phi, \gamma > 0$. Thus we conclude the inductive hypothesis also holds at time $t$.

**Combining the pieces for inequality (i).** Having established that expression (106) holds for all $t \geq 2$ with $C = 4\widehat{L}\phi^4\gamma$, it follows by expression (107) and the fact that $\eta C \leq 1/2$ for $\eta$ as in (84) that $\|\Delta_{t-1}\|_{y_t} \leq 2\|\Delta_t\|_{y_t}$, which yields the desired inequality (i) from (104).

**Proof of inequality (ii).** To prove the second inequality (ii) from (104), we apply inequality (i) at time $t+1$. Together with two applications of Lemma G.2, we then obtain

$$\|y_t - y_{t-1}\|_{y_t} \leq \phi \cdot \|y_t - y_{t-1}\|_{y_{t+1}} \leq 2\phi \cdot \|y_{t+1} - y_t\|_{y_{t+1}} \leq 2\phi^2 \cdot \|y_{t+1} - y_t\|_{y_t} \ , \tag{116}$$

where $\phi > 0$ is as defined in (108). Under the setting of $\eta > 0$ as in (84), it is straightforward to check that $\phi^2 = \exp(6\eta\widehat{\omega}\widehat{B}) \leq 3/2$. Thus we conclude from (116) that $\|y_t - y_{t-1}\|_{y_{t+1}} \leq 3\|y_{t+1} - y_t\|_{y_t}$. This yields the desired inequality (ii) from (104) and completes the proof of the proposition. ■

### H.2.4 Bound on initial change in energy

For (OFTRL), note that the dissipation of Lemma 5.4 holds only for times $t \geq 1$. At time $t = 0$, the shifted energy $F_{x^\star}(y_1) - F_{x^\star}(y_0)$ (for $x^\star \in \mathrm{MEQ}(V)$) may increase. In the following lemma, we derive an upper bound on this maximum increase in shifted energy after the first iteration:

**Lemma H.6** (OFTRL energy startup)**.** *Let $\{(x_t, y_t)\}$ be the iterates of (OFTRL) with $\eta$ as in Lemma 5.4. Then for any $x^\star \in \mathrm{MEQ}(V)$, it holds at time $t = 1$ that*

$$F_{x^\star}(y_1) - F_{x^\star}(y_0) \leq \eta^2 \|V(x_0)\|_{y_0}^2 \ .$$

**Proof.** Recall for (OFTRL) that $y_1 - y_0 = -\eta V(x_0)$. Thus at the first step, OFTRL takes a forward, vanilla FTRL step. Then, fix $x^\star \in \mathrm{MEQ}(V)$, and using the fact that the initializations satisfy $x_0 = \nabla F(y_0)$, observe that

$$\begin{aligned} F_{x^\star}(y_1) - F_{x^\star}(y_0) &= F(y_1) - F(y_0) - \langle x^\star, y_1 - y_0\rangle \\ &= -\eta\langle x_0 - x^\star, V(x_0)\rangle + \mathsf{D}_F(y_1, y_0) = \mathsf{D}_F(y_1, y_0) \ . \end{aligned}$$

Here, the final equality is due to $x^\star \in \mathrm{MEQ}(V)$. As $\|y_1 - y_0\|_* \leq \eta\widehat{\tau}$, we apply Lemma G.4 to obtain

$$F_{x^\star}(y_1) - F_{x^\star}(y_0) = \mathsf{D}_F(y_1, y_0) \leq \mu(\eta\widehat{\tau}) \cdot \|y_1 - y_0\|_{y_0}^2 \leq \mu(\eta\widehat{\tau}) \cdot \eta^2 \|V(x_0)\|_{y_0}^2 \ ,$$

where $\mu(\cdot)$ is the scalar function from (85). For $\eta$ as in Lemma 5.4 (and in (84)), it is straightforward to verify the crude bound $\mu(\eta\widehat{\tau}) \leq 1$. Substituting this into the prior expression yields the desired bound and completes the proof. ■

We also derive the following crude upper bound on $\mathsf{D}_R(x^\star, x_1)$ which will be used later:

**Lemma H.7** (OFTRL initial Bregman)**.** *Let $\{(x_t, y_t)\}$ be the iterates of (OFTRL) with $\eta$ as in Lemma 5.4. Then for every $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$, it holds that $\mathsf{D}_R(x^\star, x_1) \leq 2\mathsf{D}_R(x^\star, x_0)$.*

**Proof.** First, we will derive a general (algorithm-agnostic) upper bound on the term $\|V(x)\|_y^2$ that holds for all primal-dual pairs $x$ and $y$. For this, fix $y \in \mathbb{R}^D$ and $x = \nabla F(y) \in \mathrm{ri}(\mathcal{X})$. Further fix $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$, and recall by Lemma B.5 that $V(x^\star) \in \mathcal{S}$. Thus, as $\mathrm{Null}(\nabla^2 F(y)) = \mathcal{S}$ for all $y \in \mathbb{R}^D$, we have $\|V(x)\|_y^2 = \|V(x) - V(x^\star)\|_y^2$. Then, applying Lemma G.1, using the $\widehat{L}$-Lipschitzness of $V$, and by 1-strong-convexity of $R$ with respect to $\|\cdot\|$, we obtain:

$$\|V(x)\|_y^2 = \|V(x) - V(x^\star)\|_y^2 \leq \|V(x) - V(x^\star)\|_*^2 \leq \widehat{L}^2 \cdot \|x - x^\star\|^2 \leq 2\widehat{L}^2 \cdot \mathsf{D}_R(x^\star, x) \ . \tag{117}$$

Now, for the (OFTRL) iterates, we have by Lemma H.6 that $F_{x^\star}(y_1) - F_{x^\star}(y_0) \leq \eta^2\|V(x_0)\|_{y_0}^2$ for all $x^\star \in \mathrm{MEQ}(V)$. Then applying (117) and to the OFTRL pair $(x_0, y_0)$ and using Proposition 5.3, we obtain for any $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$:

$$\mathsf{D}_R(x^\star, x_1) - \mathsf{D}_R(x^\star, x_0) \leq \eta^2\|V(x_0)\|_{y_0}^2 \leq 2\widehat{L}^2\eta^2\mathsf{D}_R(x^\star, x_0) \leq \mathsf{D}_R(x^\star, x_0) \ . \tag{118}$$

Here, the final inequality comes from the fact that (conservatively) $2\widehat{L}^2\eta^2 \leq 1$ for $\eta$ as in (84). Rearranging (118) then yields the statement of the lemma. ■

### H.3 Instantiating Coupling for EGFTRL

Here we give details on instantiating of Lemma 5.4 along the iterates of (EGFTRL), which requires slightly less work than for the OFTRL case.

First, the following proposition relates the dual increments $y_{t+1} - y_t$ to the operator value $V(x_t)$:

**Proposition H.8.** *For* (EGFTRL) *with* $\eta > 0$ *as in* (84)*, it holds for all* $t \geq 1$*:*

$$\tfrac{5}{6}\eta\|V(x_t)\|_{y_t} \leq \|y_{t+1} - y_t\|_{y_t} \leq \tfrac{7}{6}\eta\|V(x_t)\|_{y_t} \ .$$

The second proposition derives a bound on the coupling error terms $C_t = \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle$ from the implicit-step coupling of Proposition 5.5.

**Proposition H.9.** *For* (EGFTRL) *with* $\eta > 0$ *as in* (84)*, it holds for all* $t \geq 1$ *and* $x^\star \in \mathrm{MEQ}(V)$*:*

$$|C_t| \leq 9\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 \ .$$

We give the proof of Propositions H.8 and H.9 for (EGFTRL) below in Section H.3.1 and Section H.3.2, respectively. Compared to the proofs for (OFTRL), for the (EGFTRL), we require slightly different machinery given the differences in the dual updates. In particular, we require the following stability bound on the dual (EGFTRL) iterates in terms of the halfstep sequence $\{y_{t+1/2}\}$:

**Lemma H.10** (EGFTRL Dual Stability)**.** *Under* (EGFTRL) *with* $\eta$ *satisfying* (84)*, for all* $t \geq 0$*:*

$$\tfrac{5}{6} \cdot \|y_{t+1/2} - y_t\|_{y_t} \ \leq \ \|y_{t+1} - y_t\|_{y_t} \ \leq \ \tfrac{7}{6} \cdot \|y_{t+1/2} - y_t\|_{y_t} \ .$$

The proof of Lemma H.10 is given in Section H.3.3. For now, we grant the result as true for use in the proofs of Propositions H.3 and H.4.

#### H.3.1 Proof of Proposition H.8

Here, we give the short proof of Proposition H.8, which relates for (EGFTRL) the magnitude of the dual increments $y_{t+1} - y_t$ to the vector $V(x_t)$ under the local seminorm $\|\cdot\|_{y_t}$

**Proof.** Recall under (EGFTRL) that $y_{t+1/2} - y_t = -\eta V(x_t)$ for all $t \geq 0$. Thus substituting this difference into the statement of Lemma H.10 yields the claim. ■

#### H.3.2 Proof of Proposition H.9

In this section, we give the proof of Proposition H.9. For this, recall for any $x^\star \in \mathrm{MEQ}(V)$ that we define $C_t := \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle$.

**Proof.** Similar to Proposition H.4, the proof proceeds by first expanding $C_t$ under the (EGFTRL) iterates, and then bounding the resulting constituent error terms.

**Expanding the inner product term.**

Under (EGFTRL), the difference $y_{t+1} - \widehat{y}_{t+1}$ is given by

$$y_{t+1} - \widehat{y}_{t+1} = y_t - \eta V(x_{t+1/2}) - (y_t - \eta V(x_{t+1})) = \eta\big(V(x_{t+1}) - V(x_{t+1/2})\big) \ .$$

Fixing $x^\star$ satisfying (MEQ), it follows that we can simplify

$$\begin{aligned} C_t = \langle x_{t+1} - x^\star, y_{t+1} - \widehat{y}_{t+1}\rangle &= \eta\langle x_{t+1} - x^\star, V(x_{t+1}) - V(x_{t+1/2})\rangle \\ &= -\eta\langle x_{t+1} - x^\star, V(x_{t+1/2})\rangle \\ &= -\eta\langle x_{t+1/2} - x^\star + x_{t+1} - x_{t+1/2}, V(x_{t+1/2})\rangle \\ &= -\eta\langle x_{t+1} - x_{t+1/2}, V(x_{t+1/2})\rangle \ . \end{aligned} \tag{119}$$

Here, the second and fourth equalities follow from the fact that $x^\star$ satisfies (MEQ). To further simplify the inner product in (119), we use the definition of the (EGFTRL) iterates and the Hessian remainder term $\mathsf{G}_F(y_{t+1/2}, y_t)$ to write

$$\begin{aligned} x_{t+1} - x_{t+1/2} &= \nabla F(y_{t+1}) - \nabla F(y_{t+1/2}) \\ &= \nabla^2 F(y_{t+1/2})(y_{t+1} - y_{t+1/2}) + \mathsf{G}_F(y_{t+1}, y_{t+1/2}) \\ &= \eta\nabla^2 F(y_{t+1/2})(V(x_t) - V(x_{t+1/2})) + \mathsf{G}_F(y_{t+1}, y_{t+1/2}) \ . \end{aligned} \tag{120}$$

Due to the symmetry of the Hessian $\nabla^2 F(y_{t+1/2})$, substituting (120) into (119) yields

$$C_t = \underbrace{\eta^2 \big\langle V(x_{t+1/2}) - V(x_t), V(x_{t+1/2}) \big\rangle_{y_{t+1/2}}}_{\mathcal{E}_{t,1}} \underbrace{- \eta \big\langle \mathsf{G}_F(y_{t+1}, y_{t+1/2}), V(x_{t+1/2}) \big\rangle}_{\mathcal{E}_{t,2}} . \tag{121}$$

**2. Bounds on magnitude of coupling error terms.**

We now derive bounds on the magnitudes of the two error terms $\mathcal{E}_{t,1}$ and $\mathcal{E}_{t,2}$ from (121). For this, we use the scalar functions $\phi(\cdot)$, $\gamma(\cdot)$, and $\nu(\cdot)$ from (85). Recall also that under (EGFTRL), we have $\|y_{t+1} - y_t\|_* , \|y_{t+1/2} - y_t\|_* \leq \eta\widehat{\tau}$ for all $t \geq 0$. Thus, for readability, we will slightly abuse notation and overload $\phi := \phi(\eta\widehat{\tau})$, $\gamma := \gamma(\eta\widehat{\tau})$, and $\nu := \nu(\eta\widehat{\tau})$. We then proceed with the bounds on $\mathcal{E}_{t,1}$, and $\mathcal{E}_{t,2}$:

First, for the term $\mathcal{E}_{t,1}$, recall that $y_{t+1} - y_t = -\eta V(x_{t+1/2})$. Then we can bound:

$$\begin{aligned}
|\mathcal{E}_{t,1}| &= \big|\langle \eta(V(x_{t+1/2}) - V(x_t)), \eta(V(x_{t+1/2})\rangle_{y_{t+1/2}}\big| \\
&\leq \eta \|V(x_{t+1/2}) - V(x_t)\|_{y_{t+1/2}} \cdot \|\eta V(x_{t+1/2})\|_{y_{t+1/2}} && \text{(Cauchy-Schwarz)} \\
&\leq \phi^2 \cdot \eta \|V(x_{t+1/2}) - V(x_t)\|_{y_t} \cdot \|y_{t+1} - y_t\|_{y_t} && \text{(Lemma G.2, twice)} \\
&\leq \widehat{L}\gamma\phi^2 \cdot \eta \|y_{t+1/2} - y_t\|_{y_t} \cdot \|y_{t+1} - y_t\|_{y_t} && \text{(Lemma G.7)} \\
&\leq 2\widehat{L}\gamma\phi^2 \cdot \eta \|y_{t+1} - y_t\|_{y_t}^2 && \text{(Lemma H.10)} \\
&\leq \tfrac{9}{2}\widehat{L}\gamma\phi^2 \cdot \eta^3 \|V(x_t)\|_{y_t}^2 && \text{(Proposition H.8) .}
\end{aligned} \tag{122}$$

For the second term $\mathcal{E}_{t,2}$, first observe under (EGFTRL) that $y_{t+1} - y_{t+1/2} = \eta(V(x_t) - V(x_{t+1/2}))$. Then it follows that we can bound the term

$$\begin{aligned}
\|\mathsf{G}_F(y_{t+1}, y_{t+1/2})\|_{y_t,*} &\leq \phi \cdot \|\mathsf{G}_F(y_{t+1}, y_{t+1/2})\|_{y_{t+1/2},*} && \text{(Lemma G.2)} \\
&\leq \phi\nu \cdot \|y_{t+1} - y_{t+1/2}\|_{y_{t+1/2}} && \text{(Lemma G.6)} \\
&\leq \phi^2\nu \cdot \|y_{t+1} - y_{t+1/2}\|_{y_t} && \text{(Lemma G.2)} \\
&= \phi^2\nu \cdot \eta \|V(x_{t+1/2}) - V(x_t)\|_{y_t} \\
&\leq \widehat{L}\gamma\phi^2\nu \cdot \eta \|y_{t+1/2} - y_t\|_{y_t} && \text{(Lemma G.7)} \\
&\leq 2\widehat{L}\gamma\phi^2\nu \cdot \eta \|y_{t+1} - y_t\|_{y_t} && \text{(Lemma H.10) .}
\end{aligned} \tag{123}$$

To bound $|\mathcal{E}_{t,2}|$, we can then use the generalized Cauchy-Schwarz inequality to write

$$\begin{aligned}
|\mathcal{E}_{t,2}| &\leq \|\mathsf{G}_F(y_{t+1}, y_{t+1/2})\|_{y_t,*} \cdot \|\eta V(x_{t+1/2})\|_{y_t} \\
&\leq 2\widehat{L}\gamma\phi^2\nu \cdot \eta \|y_{t+1} - y_t\|_{y_t}^2 && \text{(substituting (123))} \\
&\leq \tfrac{9}{2}\widehat{L}\gamma\phi^2\nu \cdot \eta^3 \|V(x_t)\|_{y_t}^2 && \text{(Proposition H.8) .}
\end{aligned} \tag{124}$$

**3. Combining the pieces.**

Substituting the bounds on $|\mathcal{E}_{t,1}|$ and $|\mathcal{E}_{t_2}|$ into (121), we obtain

$$|C_t| \leq |\mathcal{E}_{t,1}| + |\mathcal{E}_{t,2}| \leq \tfrac{9}{2}\widehat{L}\gamma\phi^2 \cdot (1+\nu) \cdot \eta^3 \|V(x_t)\|_{y_t}^2 . \tag{124}$$

For $\eta$ as in (84), it is straightforward to verify that $\gamma\phi^2(1+\nu) \leq 2$. Thus we obtain

$$|C_t| \leq 9\widehat{L} \cdot \eta^3 \|V(x_t)\|_{y_t}^2 ,$$

which concludes the proof. ■

### H.3.3 Proof of Lemma H.10.

This section gives the proof of Lemma H.10, which establishes the stability of the dual iterates of (EGFTRL) in terms of the half-step sequence $\{y_{t+1/2}\}$:

**Proof.** For all $t \geq 0$, let $\Delta_{t+1/2} = y_{t+1/2} - y_t$ and $\Delta_{t+1} = y_{t+1} - y_t$. Then under (EGFTRL), observe that

$$\Delta_{t+1} - \Delta_{t+1/2} = \eta(V(x_t) - V(x_{t+1/2})) \, .$$

Further recall for all $t \geq 0$ that $\|y_t - y_{t+1/2}\|_* \leq \eta\widehat{\tau}$ (for $\widehat{\tau}$ as in (82)). Similar to the proof of Lemma H.5, we again slightly overload notation and let $\gamma := \frac{\exp(\eta\widehat{\omega}\widehat{\tau})-1}{\eta\widehat{\omega}\widehat{\tau}}$. For $\eta$ as in (84), it is then straightforward to verify that $\eta\widehat{L}\gamma \leq 1/6$. Then, taking local semi-norms and applying Lemma G.7, we obtain the bound

$$\|\Delta_{t+1} - \Delta_{t+1/2}\|_{y_t} = \eta\|V(x_{t+1/2}) - V(x_t)\|_{y_t} \leq \eta \cdot \widehat{L}\gamma \cdot \|y_{t+1/2} - y_t\|_{y_t} \, . \tag{125}$$

We now prove the upper and lower inequalities of the lemma statement separately:

**Upper inequality.** We have that $y_{t+1} - y_t = \Delta_{t+1} = \Delta_{t+1/2} + (\Delta_{t+1} - \Delta_{t+1/2})$. Then by the triangle inequality and substituting the bound from (125), we find

$$\begin{aligned}\|y_{t+1} - y_t\|_{y_t} &\leq \|\Delta_{t+1/2}\|_{y_t} + \|\Delta_{t+1} - \Delta_{t+1/2}\|_{y_t} \\ &\leq (1 + \eta L\gamma) \cdot \|\Delta_{t+1/2}\|_{y_t} \leq \tfrac{7}{6} \cdot \|\Delta_{t+1/2}\|_{y_t} = \tfrac{7}{6} \cdot \|y_{t+1/2} - y_t\|_{y_t} \, .\end{aligned}$$

Here, the final inequality comes from $\eta\widehat{L}\gamma \leq 1/6$ for $\eta$ as in (84). This proves the upper inequality.

**Lower inequality.** We similarly have that $y_{t+1/2} - y_t = \Delta_{t+1/2} = \Delta_{t+1} + (\Delta_{t+1/2} - \Delta_{t+1})$. Again by the triangle inequality and substituting the bound from (125), we then obtain

$$\|\Delta_{t+1/2}\|_{y_t} \leq \|\Delta_{t+1}\|_{y_t} + \|\Delta_{t+1} - \Delta_{t+1/2}\|_{y_t} \leq \|\Delta_{t+1}\|_{y_t} + \eta\widehat{L}\gamma \cdot \|\Delta_{t+1/2}\|_{y_t} \, .$$

Rearranging and using the fact that $\eta\widehat{L}\gamma \leq 1/6$ then yields

$$\|y_{t+1} - y_t\|_{y_t} = \|\Delta_{t+1}\|_{y_t} \geq (1 - \eta\widehat{L}\gamma) \cdot \|\Delta_{t+1/2}\|_{y_t} \geq \tfrac{5}{6} \cdot \|y_{t+1/2} - y_t\|_{y_t} \, .$$

This proves the lower inequality and completes the proof of the lemma. ■

# I Details on Equilibrium Selection

In this section, we prove for the symmetrized system that OFTRL and EGFTRL converge asymptotically to a unique limit point, and we further give an exact characterization of the limit. Formally:

**Lemma I.1** (Equilibrium selection)**.** *Assume the setting of Lemma 5.4. Further, let Assumption 5 hold, and assume that* $\mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$*. Then for any initialization satisfying* $x_0 \in \mathrm{ri}(\mathcal{X})$*, it holds for both* (OFTRL) *and* (EGFTRL) *that:*

*(i) There exists a unique limit point* $x_\infty := \lim_{t\to\infty} x_t \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$.

*(ii) The point* $x_\infty$ *satisfies* $x_\infty = \mathrm{argmin}_{x^\star \in \mathrm{MEQ}(V)} \mathsf{D}_R(x^\star, x_0)$.

## I.1 Proof of Lemma I.1

This section gives the proof of Lemma I.1. We prove the two claims in sequence:

**Proof of part (i) – unique MEQ limit point.**

First, we show that the iterates $\{x_t\}_{t\geq 1}$ converge asymptotically to a unique limit point $x_\infty \in \mathrm{MEQ}(V)$: We proceed in several steps:

*1. Iterates lie in compact sublevel set.*

First, using Proposition 5.3 and the upper bound on shifted energy dissipation from Lemma 5.4, we have for all $t \geq 1$ and $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$ that:

$$\mathsf{D}_R(x^\star, x_{t+1}) - \mathsf{D}_R(x^\star, x_t) = F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \leq -\tfrac{1}{8}\eta^2\|V(x_t)\|_{y_t}^2 \, . \tag{126}$$

Now, fix one point $\widetilde{x}^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$, which exists by assumption of the lemma. Further, let $v := \mathsf{D}_R(\widetilde{x}^\star, x_1)$, where $v$ is finite due to $x_1, \widetilde{x}^\star \in \mathrm{ri}(\mathcal{X})$. By Proposition C.7, recall that

$\mathsf{D}_R(\widetilde{x}^\star, x) \to \infty$ as $x$ approaches the relative boundary of $\mathcal{X}$. Together with the continuity of $\mathsf{D}_R(\widetilde{x}^\star, \cdot)$ over $\mathrm{ri}(\mathcal{X})$, this implies compactness of the sublevel set

$$\mathcal{C} := \{x \in \mathrm{ri}(\mathcal{X}) : \mathsf{D}_R(\widetilde{x}^\star, x) \le v\} \subset \mathrm{ri}(\mathcal{X}) \; . \tag{127}$$

Together with (126), this implies that $x_t \in \mathcal{C}$ for all $t \ge 1$.

*2. Properties of dissipation terms.*

First, let $\Phi : \mathrm{ri}(\mathcal{X}) \to \mathbb{R}$ be given by $\Phi(x) := \frac{1}{8}\eta^2 \langle V(x), \nabla^2 F(y) V(x)\rangle$ for all $x \in \mathrm{ri}(\mathcal{X})$, where $y = \nabla R(x) \in \mathbb{R}^D$. Using the fact that $y_t - \nabla R(x_t) \in \mathcal{S}$ from Proposition C.2 and the shift-invariance identity of Proposition C.5 (i), it holds that $\Phi(x_t) = \frac{1}{8}\eta^2 \|V(x_t)\|_{y_t}^2$ for $x_t = \nabla F(y_t)$ as in (126). By definition of the local norm $\|\cdot\|_y$ and continuity of $V$, $\nabla R$, and $\nabla^2 F(y)$, we have that $\Phi$ is continuous on $\mathrm{ri}(\mathcal{X})$. Moreover, recall from Section G.1 that $\Phi(x) \ge 0$, and $\Phi(x) = 0$ iff $V(x) \in \mathrm{Null}(\nabla^2 F(y)) = \mathcal{S}$. Thus by Proposition B.2, we have $\Phi(x) = 0$ iff $x \in \mathrm{NE}(V)$.

*3. Subsequence convergence to NE.*

Now fix the point $\widetilde{x}^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$ from Step 1. Then summing (126) from $t = 1$ to $T$ and using the definition of $\Phi$ yields

$$\mathsf{D}_R(\widetilde{x}^\star, x_{T+1}) - \mathsf{D}_R(\widetilde{x}^\star, x_1) = \textstyle\sum_{t=1}^T \big(\mathsf{D}_R(\widetilde{x}^\star, x_{t+1}) - \mathsf{D}_R(\widetilde{x}^\star, x_t)\big) \le -\sum_{t=1} \Phi(x_t) \; .$$

Rearranging, using the nonnegativity of $\mathsf{D}_R(\widetilde{x}^\star, x_{T+1})$ and each $\Phi(x_t)$, it follows that

$$0 \le \textstyle\sum_{t=1}^T \Phi(x_t) \le \mathsf{D}_R(\widetilde{x}^\star, x_1) = v < \infty \;\; \text{for all } T \ge 1 \; .$$

Taking $T \to \infty$, this means that $\sum_{t=1}^\infty \Phi(x_t) < \infty$ is also finite, which implies $\lim_{t\to\infty} \Phi(x_t) = 0$. Now, since all $x_t \in \mathcal{C}$, it follows by compactness of $\mathcal{C}$ that there exists a convergent subsequence $\{t_k\}_{k\ge1}$ such that $x_{t_k} \to \bar{x}$. By compactness, we then have $\bar{x} \in \mathcal{C} \subset \mathrm{ri}(\mathcal{X})$, and by continuity of $\Phi$ on $\mathrm{ri}(\mathcal{X})$, it thus follows that $\lim_{k\to\infty} \Phi(x_{t_k}) = \Phi(\bar{x}) = 0$. Here, the second equality holds due to $\Phi(x_t) \to 0$ along the entire sequence. By Step 2, this means that the limit $\bar{x}$ of any convergent subsequence $\{x_{t_k}\}_{k\ge1}$ satisfies $\bar{x} \in \mathrm{NE}(V)$.

*4. Global convergence to unique MEQ point.*

Finally, we establish that the entire sequence $\{x_t\}_{t\ge1}$ converges to a unique MEQ point. First, we show that the limit point $\bar{x}$ of any convergent subsequence from Step 3 is in the reachable primal set $\mathcal{R}(x_0)$. Let $\bar{x}$ be such a limit point, and recall that $x_t \in \mathcal{R}(x_0)$ for all $t \ge 0$. Then, by Step 3, we also have that $\bar{x} \in \mathrm{NE}(V) \cap \mathcal{C} \subset \mathrm{ri}(\mathcal{X})$. Recall from Proposition E.3 that for the initialization pair $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0)$, it holds for any $x \in \mathrm{ri}(\mathcal{X})$ that $x \in \mathcal{R}(x_0)$ if and only if $\Pi_{\mathcal{S}^\perp}(\nabla R(x) - y_0) \in \mathcal{U}$, where $\mathcal{U} = \Pi_{\mathcal{S}^\perp}\mathrm{Span}(V(\mathcal{X}))$. By continuity of $\nabla R$ on $\mathrm{ri}(\mathcal{X})$, and since $\mathcal{U}$ is a closed, linear subspace, it then follows that if $x_{t_k} \in \mathcal{R}(x_0)$ for all points in the convergent subsequence $\{x_{t_k}\}$, then also $\bar{x} \in \mathcal{R}(x_0)$.

Now, under Assumption 5, recall by Corollary E.5 that if $x \in \mathrm{NE}(V) \cap \mathcal{R}(x_0)$, then $x \in \mathrm{MEQ}(V)$. Thus $\bar{x} \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$. However, we have by Proposition E.2 that $|\mathrm{MEQ}(V) \cap \mathcal{R}(x_0)| \le 1$ (which holds independently of Assumption 5). Since the preceding arguments held for the limit of any convergent subsequence, it follows that all convergent subsequences have the same limit, which we denote by $x_\infty \in \mathrm{MEQ}(V)$. Finally, by compactness, the entire sequence $\{x_t\}$ must also converge to this unique point $x_\infty := \lim_{t\to\infty} x_t \in \mathrm{MEQ}(V)$. ■

**Proof of part (ii) – characterization of limit point.**

By Part (i) of Lemma I.1, we have that $x_\infty := \lim_{t\to\infty} x_t \in \mathrm{MEQ}(V)$. For Part (ii), we further give an exact characterization of $x_\infty$ in terms of the initialization $x_0$. We proceed in several steps:

*1. Change in shifted energy is equivalent for all MEQ points.*

Recall for $x^\star \in \mathrm{MEQ}(V)$ and $y \in \mathbb{R}^D$ that $F_{x^\star}(y) = F(y) - \langle x^\star, y\rangle$. Then the following helper lemma establishes that the change in shifted energy with respect to any MEQ point is invariant:

**Lemma I.2.** *Assume the setting of Lemma I.1. Then for any $x_1^\star, x_2^\star \in \mathrm{MEQ}(V)$ and all $t \ge 0$:*

$$F_{x_1^\star}(y_{t+1}) - F_{x_1^\star}(y_t) = F_{x_2^\star}(y_{t+1}) - F_{x_2^\star}(y_t) \; .$$

*Proof.* Let $\Delta F_{x_1^\star}(y_t) := F_{x_1^\star}(y_{t+1}) - F_{x_1^\star}(y_t)$ and $\Delta F_{x_2^\star}(y_t) := F_{x_2^\star}(y_{t+1}) - F_{x_2^\star}(y_t)$. Now, recall that for both (OFTRL) and (EGFTRL), we have $y_{t+1} - y_t \in \mathrm{Span}(V(\mathcal{X}))$ for all $t \geq 1$. Thus, in both cases, there exists some $k \geq 1$, $\{x_i\}$, and $\{c_i\}$, where each $x_i \in \mathcal{X}$ and $c_i \in \mathbb{R}$ such that $y_{t+1} - y_t = \sum_{i=1}^k c_i V(x_i)$. Taking the difference between $\Delta F_{x_1^\star}(y_t)$ and $\Delta F_{x_2^\star}(y_t)$ then yields:

$$\Delta F_{x_1^\star}(y_t) - \Delta F_{x_2^\star}(y_t) = -\langle y_{t+1} - y_t, x_1^\star - x_2^\star \rangle = -\textstyle\sum_{i=1}^k c_i \langle V(x_i), x_1^\star - x_2^\star \rangle \,. \tag{128}$$

Moreover, since $x_1^\star, x_2^\star \in \mathrm{MEQ}(V)$ it follows by definition of (MEQ) that, for each $i \in [k]$:

$$\langle V(x_i), x_i - x_1^\star \rangle = \langle V(x_i), x_i - x_2^\star \rangle = 0 \quad \Longrightarrow \quad \langle V(x_i), x_1^\star - x_2^\star \rangle = 0 \,. \tag{129}$$

Substituting (129) into (128) then gives $\Delta F_{x_1^\star}(y_t) - \Delta F_{x_2^\star}(y_t) = 0$. □

*2. Change in Bregman divergence is equivalent for all MEQ points.*

Using Lemma I.2, we can further establish the following invariant between the total change in Bregman divergence from two interior MEQ points after $t$ steps:

**Lemma I.3.** *Let $x_1^\star, x_2^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$. Then for any $t \geq 1$, it holds that*

$$\mathsf{D}_R(x_1^\star, x_t) - \mathsf{D}_R(x_1^\star, x_0) = \mathsf{D}_R(x_2^\star, x_t) - \mathsf{D}_R(x_2^\star, x_0) \,.$$

*Proof.* First, observe for any $x_1^\star, x_2^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$ and for any $t \geq 0$ that

$$\begin{aligned} \mathsf{D}_R(x_1^\star, x_{t+1}) - \mathsf{D}_R(x_1^\star, x_t) &= F_{x_1^\star}(y_{t+1}) - F_{x_1^\star}(y_t) \\ &= F_{x_2^\star}(y_{t+1}) - F_{x_2^\star}(y_t) = \mathsf{D}_R(x_2^\star, x_{t+1}) - \mathsf{D}_R(x_2^\star, x_t) \end{aligned} \tag{130}$$

The first and last equalities follow from Proposition 5.3, and the second equality by Lemma I.2. Telescoping then yields $\sum_{k=1}^t \mathsf{D}_R(x_1^\star, x_k) - \mathsf{D}_R(x_1^\star, x_{k-1}) = \sum_{k=1}^t \mathsf{D}_R(x_2^\star, x_k) - \mathsf{D}_R(x_2^\star, x_{k-1})$ for all $t \geq 1$. Thus for $x_1^\star, x_2^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$, we conclude by simplifying the two telescoping sums that

$$\mathsf{D}_R(x_1^\star, x_t) - \mathsf{D}_R(x_1^\star, x_0) = \mathsf{D}_R(x_2^\star, x_t) - \mathsf{D}_R(x_2^\star, x_0) \,, \tag{131}$$

which yields the desired equality. □

*3. Limit point is Bregman projection from initialization onto* $\mathrm{MEQ}(V)$.

As part (i) of the lemma proves $x_\infty = \lim_{t\to\infty} x_t \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$, we substitute $x_2^\star = x_\infty$ in the right hand side of Lemma I.3. This then yields for all $t \geq 1$ and $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$ that

$$\mathsf{D}_R(x^\star, x_t) - \mathsf{D}_R(x^\star, x_0) = \mathsf{D}_R(x_\infty, x_t) - \mathsf{D}_R(x_\infty, x_0) \tag{132}$$

Then by continuity of $\mathsf{D}_R$, taking $t \to \infty$ yields for all $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$:

$$\mathsf{D}_R(x^\star, x_\infty) - \mathsf{D}_R(x^\star, x_0) = \mathsf{D}_R(x_\infty, x_\infty) - \mathsf{D}_R(x_\infty, x_0) = -\mathsf{D}_R(x_\infty, x_0) \,.$$

Rearranging then yields: $\mathsf{D}_R(x^\star, x_0) = \mathsf{D}_R(x^\star, x_\infty) + \mathsf{D}_R(x_\infty, x_0)$ for all $x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$. As $\mathsf{D}_R(x^\star, x_\infty) \geq 0$ for all $x^\star \in \mathrm{MEQ}(V)$ by convexity, we further obtain

$$\mathsf{D}_R(x^\star, x_0) \geq \mathsf{D}_R(x_\infty, x_0) \;\; \text{for all } x^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X}) \,. \tag{133}$$

To extend (133) to also hold for *non-interior* $x^\star \in \mathrm{MEQ}(V)$, observe that if $R(x^\star) = \infty$ is not finite, then also $\mathsf{D}_R(x^\star, x_0) = \infty$, and thus (133) trivially holds. Now suppose $x^\star \in \mathrm{MEQ}(V)$ but $x^\star \notin \mathrm{ri}(\mathcal{X})$ and $R(x^\star) < \infty$ is finite. Then for $s \in (0,1)$, let $x_s^\star := (1-s)x^\star + s x_\infty$. By convexity of $\mathrm{MEQ}(V)$ and since $x_\infty \in \mathrm{ri}(\mathcal{X})$, it follows that $x_s^\star \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$. Thus, applying (133) to $x_s^\star$ and by Jensen's inequality (since $\mathsf{D}_R(\cdot, x_0)$ is convex in its first argument), we have

$$\mathsf{D}_R(x_\infty, x_0) \leq \mathsf{D}_R(x_s^\star, x_0) \leq (1-s) \cdot \mathsf{D}_R(x^\star, x_0) + s \cdot \mathsf{D}_R(x_\infty, x_0) \,.$$

Rearranging, we then obtain $(1-s) \cdot \mathsf{D}_R(x_\infty, x_0) \leq (1-s) \cdot \mathsf{D}_R(x^\star, x_0)$, and thus dividing by $1 - s > 0$ we obtain $\mathsf{D}_R(x_\infty, x_0) \leq \mathsf{D}_R(x^\star, x_0)$. Thus it holds also that

$$\mathsf{D}_R(x^\star, x_0) \geq \mathsf{D}_R(x_\infty, x_0) \;\; \text{for all } x^\star \in \mathrm{MEQ}(V) \tag{134}$$

Moreover, since $\mathrm{MEQ}(V)$ is convex and $\mathsf{D}_R(\cdot, x_0)$ is strictly convex, (134) implies that $x_\infty$ is the unique minimizer

$$x_\infty = \mathrm{argmin}_{x^\star \in \mathrm{MEQ}(V)} \mathsf{D}_R(x^\star, x_0) \,,$$

which concludes the proof. ■

# J Proof of Proposition 5.6: Operator Domination Inequality

In this section, we develop the proof of the operator domination inequality of Proposition 5.6 that was introduced in Section 5. This inequality is restated here.

**Proposition 5.6** (Operator domination on reachable set)**.** *Let Assumption 5 hold. Fix $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0)$ such that $\mathcal{R}(x_0) \not\subseteq \mathrm{NE}(V)$. Let $x^\star \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$, and let $\mathcal{K} \subset \mathrm{ri}(\mathcal{X})$ be any compact set such that $x^\star \in \mathcal{K}$. Then there exists a positive constant $\widetilde{\gamma}(x_0, \mathcal{K}) > 0$ such that*

$$\|V(x)\|_y^2 \;\geq\; \widetilde{\gamma}(x_0, \mathcal{K}) \cdot \mathsf{D}_R(x^\star, x)\,, \quad \forall\, y \in \mathbb{R}^D \text{ s.t. } x = \nabla F(y) \in \mathcal{K} \cap \mathcal{R}(x_0)\,.$$

## J.1 Proof of Proposition 5.6

We now give the proof of Proposition 5.6. For readability, we first recall several definitions for the symmetrized system from Setting D.1.

**Recap of spaces and maps.**

Recall the subspaces $\mathcal{Y} = \mathrm{Span}(V(\mathcal{X}))$ and $\mathcal{U} = \Pi_{\mathcal{S}^\perp}\mathcal{Y}$. For the fixed initialization pair $y_0 \in \mathbb{R}^D$ and $x_0 = \nabla F(y_0)$, recall that $\mathcal{R}(x_0) = \{\nabla F(y_0 + y) : y \in \mathcal{Y}\} \subset \mathrm{ri}(\mathcal{X})$. For $u \in \mathcal{U}$, recall that:

$$x_0(u) = \nabla F(y_0 + u) \in \mathcal{R}(x_0)\,, \quad \pi_0(u) = \Pi_{\mathcal{S}^\perp} V(x_0(u))\,, \quad u_0(x) = \Pi_{\mathcal{S}^\perp}(\nabla R(x) - y_0)\,.$$

By Proposition E.3, recall the mapping $u \mapsto x_0(u)$ is a bijection from the effective dual space $\mathcal{U}$ onto the reachable primal set $\mathcal{R}(x_0)$, with inverse given by $u_0(x_0(u)) = u$ for $u \in \mathcal{U}$. Thus, fix $x^\star \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$. We will then let $u^\star = u_0(x^\star)$, meaning $x_0(u^\star) = x^\star$.

The proof now proceeds via the following steps:

**1. Compactness in effective dual space.**

Let $\mathcal{K} \subset \mathrm{ri}(\mathcal{X})$ be the compact set from the proposition statement, where we assume that $x^\star \in \mathcal{K}$, and recall that we are interested in establishing a bound on $\|V(x)\|_y^2$ that holds for $y \in \mathbb{R}^D$ and $x = \nabla F(y)$ where $x \in \mathcal{K} \cap \mathcal{R}(x_0)$. Given the inverse relationship between $u_0(\cdot)$ and $x_0(\cdot)$ between $\mathcal{U}$ and $\mathcal{R}(x_0)$, it is convenient to consider the subset of $\mathcal{U}$ corresponding to $\mathcal{K} \cap \mathcal{R}(x_0)$. For this, let $\mathcal{K}_{\mathcal{U}}$ be this subset of the effective dual space given by

$$\mathcal{K}_{\mathcal{U}} := \{u_0(x) : x \in \mathcal{K} \cap \mathcal{R}(x_0)\} \subseteq \mathcal{U}\,. \tag{135}$$

By Proposition E.3, recall that $x \in \mathcal{R}(x_0)$ if and only if $u_0(x) \in \mathcal{U}$. As the map $u_0$ is continuous on $\mathrm{ri}(\mathcal{X})$ and $\mathcal{U}$ is a closed linear subspace, it follows that $\mathcal{R}(x_0)$ is relatively closed in $\mathrm{ri}(\mathcal{X})$. Thus, $\mathcal{K} \cap \mathcal{R}(x_0)$ is a closed subset of the compact set $\mathcal{K}$ and hence is compact. Moreover, $\mathcal{K}_{\mathcal{U}}$ and its convex hull $\mathrm{conv}(\mathcal{K}_{\mathcal{U}})$ are then both also compact. Observe further that for $x^\star \in \mathcal{K} \cap \mathcal{R}(x_0)$ the corresponding $u^\star = u_0(x^\star) \in \mathcal{K}_{\mathcal{U}}$.

**2. Relating dissipation term to Bregman divergence.**

Fix $u \in \mathcal{K}_{\mathcal{U}} \subset \mathcal{U}$, and recall that $x_0(u) = \nabla F(y_0 + u)$. It then suffices to derive a lower bound on $\|V(x_0(u))\|_{y_0+u}^2$ in terms of $\mathsf{D}_R(x^\star, x_0(u))$. For this, we have the following components:

*(i). Lower-bounding $\|V(x_0(u))\|_{y_0+u}^2$ in terms of $\|\pi_0(u)\|_2^2$.*

First, recall that $\nabla^2 F(y) \succeq 0$ and $\mathrm{Null}(\nabla^2 F(y)) = \mathcal{S}$ for all $y \in \mathbb{R}^D$. Then it holds that $\nabla^2 F(y)$ is positive definite when restricted to $\mathcal{S}^\perp$. Recall by definition that $\mathcal{U} \subseteq \mathcal{S}^\perp$. Then by the compactness of the convex hull $\mathrm{conv}(\mathcal{K}_{\mathcal{U}})$ and the continuity of $\nabla^2 F(\cdot)$, it follows that there exist positive constants $0 < c_\ell \leq c_h < \infty$ such that

$$c_\ell \|v\|_2^2 \leq \langle v, \nabla^2 F(y_0 + u) v\rangle \leq c_h \|v\|_2^2 \quad \forall\, u \in \mathrm{conv}(\mathcal{K}_{\mathcal{U}}), v \in \mathcal{U}\,. \tag{136}$$

For $u \in \mathcal{K}_{\mathcal{U}}$, it then follows that

$$\begin{aligned}
\|V(x_0(u))\|_{y_0+u}^2 &= \langle V(x_0(u)), \nabla^2 F(y_0 + u) V(x_0(u))\rangle && (137)\\
&= \langle \Pi_{\mathcal{S}^\perp} V(x_0(u)), \nabla^2 F(y_0 + u) \Pi_{\mathcal{S}^\perp} V(x_0(u))\rangle && (138)\\
&= \langle \pi_0(u), \nabla^2 F(y_0 + u) \pi_0(u)\rangle && (139)\\
&\geq c_\ell \|\pi_0(u)\|_2^2\,. && (140)
\end{aligned}$$

Here, the second equality follows from the fact that $\text{Null}(\nabla^2 F(y_0 + u)) = \mathcal{S}$, the third equality comes from the definition of $\pi_0(u)$, and the final inequality comes from (136).

*(ii). Upper-bounding $\mathsf{D}_R(x^\star, x_0(u))$ in terms of $\|u^\star - u\|_2^2$.*

Next, recall that $x^\star = \nabla F(y_0 + u^\star)$ for $u^\star \in \mathcal{K}_\mathcal{U}$, and using Proposition C.6, it follows that

$$\mathsf{D}_R(x^\star, x_0(u)) = F(y_0 + u) - F(y_0 + u^\star) - \langle \nabla F(y_0 + u^\star), u - u^\star \rangle \, .$$

By Taylor's theorem with remainder, it then follows that

$$\mathsf{D}_R(x^\star, x_0(u)) = \tfrac{1}{2}\langle u - u^\star, \nabla^2 F(y_0 + u^\star + u_s)(u - u^\star) \rangle \, , \tag{141}$$

where $u_s := s(u - u^\star)$ for some $s \in (0, 1)$.

As $u \in \mathcal{K}_\mathcal{U}$, then also $u^\star + u_s \in \text{conv}(\mathcal{K}_\mathcal{U})$. Moreover, we have $u - u^\star \in \mathcal{U}$. Thus, we can apply the upper bound from (136) to obtain

$$\mathsf{D}_R(x^\star, x_0(u)) \leq c_h \cdot \tfrac{1}{2}\|u - u^\star\|_2^2 \, . \tag{142}$$

Combining (140) and (142), we have the inequalities:

$$\|V(x_0(u))\|_{y_0+u}^2 \geq c_\ell \|\pi_0(u)\|_2^2 \quad \text{and} \quad \|u - u^\star\|_2^2 \geq \tfrac{1}{c_h} \cdot 2\mathsf{D}_R(x^\star, x_0(u)) \, . \tag{143}$$

The remaining task is thus to etablish that there exists a positive constant $b > 0$ such that

$$\|\pi_0(u)\|_2 \geq b\|u - u^\star\|_2 \quad \text{for all } u \in \mathcal{K}_\mathcal{U} \, . \tag{144}$$

We proceed to establish existence of such a constant:

**3. Positivity of constant in (144).**

*(i). Unique zero of $\pi_0$.*

First, we show that $u^\star = u_0(x^\star)$ is the unique zero of $\pi_0$. For this, we have by assumption that $x^\star \in \text{MEQ}(V) \cap \text{ri}(\mathcal{X})$, and thus by Proposition B.2 and Proposition E.1 we have that $V(x^\star) \in \mathcal{S}$ and thus $\pi_0(u^\star) = 0$. For the other direction, observe that if $\pi_0(u) = 0$, then we must have $x_0(u) \in \text{NE}(V) \cap \mathcal{R}(x_0)$ by Proposition E.1. Under the regularity condition of Assumption 5, it then follows from Corollary E.5 that $x_0(u) \in \text{MEQ}(V)$. However, by Proposition E.2, $|\text{MEQ}(V) \cap \mathcal{R}(x_0)| \leq 1$, and thus $x_0(u) = x^\star$. Because the map $u \mapsto x_0(u)$ is bijective, it then follows that $u = u_0(x_0(u)) = u_0(x^\star) = u^\star$. Thus $u^\star \in \mathcal{K}_\mathcal{U}$ is the unique zero of $\pi_0(u)$.

*(ii). Local error bound.*

By Assumption 5, we thus have that the Jacobian $J_{\pi_0(u^\star)} : \mathcal{U} \to \mathcal{U}$ of $\pi_0$ at $u^\star$ is non-singular. Then let $\sigma > 0$ denote the smallest singular value of $J_{\pi_0(u^\star)}$ (with respect to $\|\cdot\|_2$ on $\mathcal{U}$). As $\pi_0(u^\star) = 0$, differentiability of $\pi_0$ at $u^\star$ then implies that for $u \in \mathcal{U}$:

$$\pi_0(u) = J_{\pi_0(u^\star)}(u - u^\star) + r(u) \, , \quad \text{where} \quad \frac{\|r(u)\|_2}{\|u - u^\star\|_2} \to 0 \quad \text{as } u \to u^\star \, .$$

This implies *locally* that there exists some $\delta > 0$ such that $\|r(u)\|_2 \leq \frac{\sigma}{2}\|u - u^\star\|_2$ for all $u \in \mathcal{U}$ where $\|u - u^\star\|_2 \leq \delta$. Taking norms, it then follows for all such $u \in \mathcal{U} : \|u - u^\star\|_2 \leq \delta$ that

$$\|\pi_0(u)\|_2 \geq \|J_{\pi_0(u^\star)}(u - u^\star)\|_2 - \|r(u)\|_2 \geq \sigma \cdot \tfrac{1}{2}\|u - u^\star\|_2 \, . \tag{145}$$

*(iii) Non-local error bound.*

For $\delta > 0$ that leads to (145), let $\mathcal{K}_{\mathcal{U},\delta} := \{u \in \mathcal{K}_\mathcal{U} : \|u - u^\star\|_2 \geq \delta\}$. Suppose first that $\mathcal{K}_{\mathcal{U},\delta} \neq \emptyset$, which implies $u^\star \notin \mathcal{K}_{\mathcal{U},\delta}$. Moreover, $\mathcal{K}_{\mathcal{U},\delta}$ is a closed subset of the compact set $\mathcal{K}_\mathcal{U}$ and is thus also compact. Also, since $u^\star$ is the unique zero of $\pi_0$ from part (i), the map $u \mapsto \|\pi_0(u)\|_2 / \|u - u^\star\|_2$ is continuous and strictly positive on $\mathcal{K}_{\mathcal{U},\delta}$. Together, this implies

$$a := \min_{u \in \mathcal{K}_{\mathcal{U},\delta}} \frac{\|\pi_0(u)\|_2}{\|u - u^\star\|_2} > 0 \, . \tag{146}$$

If $\mathcal{K}_{\mathcal{U},\delta} = \emptyset$, the local bound (145) already holds throughout $\mathcal{K}_\mathcal{U}$, so taking $b = \sigma/2$ yields the desired (144). Combining the local inequality of (145) and the non-local inequality of (146), it then follows that

$$\|\pi_0(u)\|_2 \geq \min(\sigma/2, a) \cdot \|u - u^\star\|_2 \quad \text{for all } u \in \mathcal{K}_\mathcal{U} \, . \tag{147}$$

We thus let $b := \min(\sigma/2, a) > 0$.

**4. Combining the pieces.**

Finally, we combine the bounds of (143) and (147) to obtain for every $u \in \mathcal{K}_{\mathcal{U}}$ that

$$\|V(x_0(u))\|_{y_0+u}^2 \geq c_\ell \|\pi_0(u)\|_2^2 \geq c_\ell b^2 \|u - u^\star\|_2^2 \geq \frac{2c_\ell b^2}{c_h} \cdot \mathsf{D}_R(x^\star, x_0(u)) \;, \tag{148}$$

where $c_\ell, b, c_h > 0$. Then set $\widetilde{\gamma}(x_0, \mathcal{K}) := (2c_\ell b^2)/c_h > 0$, and consider $y \in \mathbb{R}^D$ such that $x = \nabla F(y) \in \mathcal{K} \cap \mathcal{R}(x_0)$. Then instantiate $u = u_0(x) \in \mathcal{K}_{\mathcal{U}}$, which implies that $x = x_0(u)$. Moreover, since by definition we have $u = \Pi_{\mathcal{S}^\perp}(\nabla R(x) - y_0)$ and also $y - \nabla R(x) \in \mathcal{S}$, we have $y - (y_0 + u) \in \mathcal{S}$. Thus by Proposition C.5, it follows that $\nabla^2 F(y) = \nabla^2 F(y_0 + u)$. Then instantiating (148), we finally obtain

$$\|V(x)\|_y^2 = \|V(x_0(u))\|_{y_0+u}^2 \geq \widetilde{\gamma}(x_0, \mathcal{K}) \cdot \mathsf{D}_R(x^\star, x) \;,$$

which completes the proof of the proposition. ■

## K Proof of Theorem 4.1: Last-Iterate Convergence Rates

In this section, we formally give the proof of Theorem 4.1, restated here:

**Theorem 4.1** (Linear last-iterate convergence rate)**.** *Let $G : \mathcal{W} \to \mathbb{R}^D$ be a loss operator such that $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weight matrix $M$, and let $\kappa_M = \frac{m_{\max}}{m_{\min}}$. Let Assumptions 1, 2, 3, 4, and 5 hold. Then for (OFTRL) or (EGFTRL) with initialization $w_0 \in \mathrm{ri}(\mathcal{W})$ and with stepsize $\eta \in (0, 1/(880LB\omega\kappa_M)]$, there exists a constant $\gamma(w_0) > 0$ such that the following hold:*

*(i) There exists a unique limit $w_\infty := \lim_{t\to\infty} w_t = \mathrm{argmin}_{w^\star \in \mathrm{WMEQ}_M(G)} \mathsf{D}_{Q_M}(w^\star, w_0)$ .*

*(ii) For all $t \geq 0$: $\mathsf{D}_Q(w_\infty, w_{t+1}) \leq 2\kappa_M \cdot \mathsf{D}_Q(w_\infty, w_0) \cdot \exp(-\gamma(w_0)\eta^2 \cdot t)$ .*

**Proof.** The proof follows by combining the components introduced in the overview of Section 5. In particular, recall that our analysis is primarily based on the $M$-symmetrized system summarized in Setting D.1, The proof then follows by studying the iterates of the symmetrized system, and using (i) energy dissipation bounds of Lemma 5.4, (ii) the asymptotic convergence and equilibrium selection characterization of Lemma I.1, and (iii) by applying the operator domination inequality of Proposition 5.6 over the OFTRL and EGFTRL iterates. In more details:

**1. Asymptotic convergence and equilibrium selection.**

Let $\{(x_t, y_t)\}$ denote the symmetrized iterates. By Lemma I.1, we have that there is a unique limit $x_\infty := \lim_{t\to\infty} x_t = \mathrm{argmin}_{x^\star \in \mathrm{MEQ}(V)} \mathsf{D}_R(x^\star, x_0) \in \mathrm{ri}(\mathcal{X})$. By the relationships between the original and $M$-symmetrized systems of Proposition 5.2, this immediately implies the existence of a unique limit $w_\infty := \lim_{t\to\infty} w_t$ satisfying $w_\infty = \mathrm{argmin}_{w^\star \in \mathrm{WMEQ}_M(G)} D_{Q_M}(w^\star, w_0)$. Moreover, it thus holds by definition that $x_\infty = M^{1/2} w_\infty$. This proves part (i) of the Theorem.

**2. Convergence rate for symmetrized system.**

For part (ii) of the theorem, we first establish a linear last-iterate convergence rate for the symmetrized system. In particular, combining the equivalence of Proposition 5.3 and using the energy dissipation upper bounds of Lemma 5.4, we have for all $t \geq 1$ that, since $x_\infty \in \mathrm{MEQ}(V) \cap \mathrm{ri}(\mathcal{X})$:

$$\mathsf{D}_R(x_\infty, x_{t+1}) - \mathsf{D}_R(x_\infty, x_t) = F_{x_\infty}(y_{t+1}) - F_{x_\infty}(y_t) \leq -\tfrac{1}{8} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \tag{149}$$

Now we have by definition of the iterates that $x_t \in \mathcal{R}(x_0) \subset \mathrm{ri}(\mathcal{X})$ for all $t \geq 1$. Also as established in the proof of Lemma I.1, there exists a compact set $\mathcal{C} \subset \mathrm{ri}(\mathcal{X})$ such that $x_t \in \mathcal{C}$ for all $t \geq 1$, and moreover that $x_\infty \in \mathcal{C} \cap \mathcal{R}(x_0)$. Then applying the operator domination inequality of Proposition 5.6 with the compact set $\mathcal{C}$, the reachable primal set $\mathcal{R}(x_0)$, for the point $x_\infty \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$, and for the iterates $y_t \in \mathbb{R}^D$ and $x_t = \nabla F(y_t) \in \mathcal{C} \cap \mathcal{R}(x_0)$ it then follows from (149) for $t \geq 1$ that

$$\mathsf{D}_R(x_\infty, x_{t+1}) - \mathsf{D}_R(x_\infty, x_t) \leq -\tfrac{1}{8}\widetilde{\gamma}(\mathcal{C}, x_0)\eta^2 \cdot \mathsf{D}_R(x_\infty, x_t) \;, \tag{150}$$

where $\widetilde{\gamma}(\mathcal{C}, x_0) > 0$ is positive and uniform over all $t \geq 1$.

Now let $\eta_{\max} := 1/(880LB\omega\kappa_M)$, and let $\widetilde{\gamma} := \min(\widetilde{\gamma}(\mathcal{C}, x_0), 4/\eta_{\max}^2) > 0$. Then observe that for all $\eta$ as in the theorem statement, it holds that $\widetilde{\gamma}\eta^2/8 \le 1/2$, and $\widetilde{\gamma}(\mathcal{C}, x_0)\eta^2 \ge \widetilde{\gamma}\eta^2$.

Then rearranging (150), it then follows recursively for all $t \ge 0$ that

$$\begin{aligned}\mathsf{D}_R(x_\infty, x_{t+1}) &\le \mathsf{D}_R(x_\infty, x_1) \cdot (1 - \tfrac{1}{8}\widetilde{\gamma}\eta^2)^t \\ &\le \mathsf{D}_R(x_\infty, x_1) \cdot \exp\big(-\tfrac{1}{8}\widetilde{\gamma}\eta^2 \cdot t\big) \end{aligned} \tag{151}$$

Now for (EGFTRL), it is straightforward to check that the preceding arguments also directly yield $\mathsf{D}_R(x_\infty, x_1) \le \mathsf{D}_R(x_\infty, x_0)$. On the other hand, for (OFTRL), we have by Lemma H.7 that $\mathsf{D}_R(x_\infty, x_1) \le 2\mathsf{D}_R(x_\infty, x_0)$. Taking the larger of the two, it follows for both algorithms that

$$\mathsf{D}_R(x_\infty, x_{t+1}) \le 2\mathsf{D}_R(x_\infty, x_0) \cdot \exp\big(-\tfrac{1}{8}\widetilde{\gamma}\eta^2 \cdot t\big) \quad \text{for all } t \ge 0\,. \tag{152}$$

**3. Convergence rate for original system.**

To complete the proof of part (ii) of the theorem, we translate the bound of (152) back to the original system. For this, we use the relationships $x_{t+1} = M^{1/2}w_{t+1}$, $x_0 = M^{1/2}w_0$, and $x_\infty = M^{1/2}w_\infty$ and the inequalities of (13) to obtain

$$m_{\min} \cdot \mathsf{D}_Q(w_\infty, w_{t+1}) \le \mathsf{D}_R(x_\infty, x_{t+1}) \quad \text{and} \quad \mathsf{D}_R(x_\infty, x_0) \le m_{\max}\mathsf{D}_Q(w_\infty, w_0)\,. \tag{153}$$

Let $\gamma(w_0) := \frac{1}{8}\widetilde{\gamma} > 0$ for the positive constant from Step 2. Then substituting the bounds of (153) into (152) and rearranging, we obtain

$$\mathsf{D}_Q(w_\infty, w_{t+1}) \le 2\tfrac{m_{\max}}{m_{\min}} \cdot \mathsf{D}_Q(w_\infty, w_0) \cdot \exp\big(-\gamma(w_0) \cdot \eta^2 t\big) \quad \text{for all } t \ge 0\,, \tag{154}$$

which concludes the proof of the theorem. ■

# L Proof of Theorem 4.5: Divergence of FTRL

In this section, we develop the proof of Theorem 4.5, restated here:

**Theorem 4.5** (Divergence of FTRL). *Let $G : \mathcal{W} \to \mathbb{R}^D$ be a linear loss operator such that $\mathrm{WMEQ}_M(G) \cap \mathrm{ri}(\mathcal{W}) \neq \emptyset$ for a weight matrix $M$, and let $\kappa_M = \frac{m_{\max}}{m_{\min}}$. Consider (FTRL) with $\eta \in (0, 1/(10B\omega\kappa_M)]$ and $w_0 \notin \mathrm{NE}(G)$. Then for $w^\star \in \mathrm{NE}(G) \cap \mathrm{ri}(\mathcal{W})$ and $\alpha > \mathsf{D}_{Q_M}(w^\star, w_0)$, letting $T_\alpha(w^\star) = \sup\{t \ge 0 : \mathsf{D}_{Q_M}(w^\star, w_t) \le \alpha\}$, there exists $\beta(w_0, w^\star, \alpha) > 0$ such that*

$$\mathsf{D}_Q(w^\star, w_t) \ge \kappa_M^{-1} \cdot \mathsf{D}_Q(w^\star, w_0) \cdot (1 + \beta(w_0, w^\star, \alpha) \cdot \eta^2)^t \ \text{ for all } 0 \le t \le T_\alpha(w^\star)\,.$$

The proof of the theorem first requires establishing a bound on the shifted energy *expansion* under vanilla FTRL (for the symmetrized system). We present this result in Section L.1, and we then give the full proof of Theorem 4.5 in Section L.2.

## L.1 Energy Expansion under FTRL

We prove the following bounds on *expansion* of shifted-energy along the vanilla (FTRL) iterates for the symmetrized system from Setting D.1.

**Lemma L.1** (Energy expansion). *Let $\{(x_t, y_t)\}$ be the iterates of (FTRL) for the symmetrized system with $\eta \in (0, 1/(10\widehat{B}\widehat{\omega})]$. Then for all $x^\star \in \mathrm{MEQ}(V)$ and all $t \ge 0$:*

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \ge \tfrac{1}{4} \cdot \eta^2 \|V(x_t)\|_{y_t}^2\,.$$

**Proof.** Reusing Proposition F.1, we have for all $x^\star \in \mathrm{MEQ}(V)$ and $t \ge 0$ that:

$$F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) = \mathsf{D}_F(y_{t+1}, y_t) \tag{155}$$

Now for all $t \ge 0$, recall that $\|y_{t+1} - y_t\|_* \le \eta\widehat{B}$. Then by the lower bound of Lemma G.4, we have

$$\mathsf{D}_F(y_{t+1}, y_t) \ge \big(\tfrac{\exp(-\widehat{\omega}\eta\widehat{B}) + \widehat{\omega}\eta\widehat{B} - 1}{(\widehat{\omega}\eta\widehat{B})^2}\big) \cdot \|y_{t+1} - y_t\|_{y_t}^2 = \big(\tfrac{\exp(-\widehat{\omega}\eta\widehat{B}) + \widehat{\omega}\eta\widehat{B} - 1}{(\widehat{\omega}\eta\widehat{B})^2}\big) \cdot \eta^2 \|V(x_t)\|_{y_t}^2\,. \tag{156}$$

The multiplicative factor in (156) is positive and strictly decreasing in $\eta$. Moreover, for $\eta$ as in the lemma statement, it is straightforward to check that this term is bounded below by $1/4$. Substituting this bound into (156) and back into (155) then concludes the proof. ■

### L.2 Proof of Theorem 4.5

We now give the full proof of Theorem 4.5, which proceeds in several steps.

**1. Overview and setup.**

Similar to the proof of Theorem 4.1, we analyze the FTRL iterates under symmetrized system of Setting D.1 and then translate the guarantees back to the original system. Specifically, let $\{(x_t, y_t)\}$ denote the vanilla (FTRL) iterates for the symmetrized system.

Then, fix $w^\star \in \mathrm{NE}(G) \cap \mathrm{ri}(\mathcal{W})$, and let $x^\star = M^{1/2} w^\star \in \mathrm{NE}(V) \cap \mathrm{ri}(\mathcal{X})$. Fix $\alpha$ satisfying the statement of the theorem, and define $T_\alpha(x^\star) = \sup\{t \geq 0 : \mathsf{D}_R(x^\star, x_t) \leq \alpha\}$. Then by definition of the symmetrized system, it holds that $T_\alpha(x^\star) = T_\alpha(w^\star)$, for $T_\alpha(w^\star)$ as defined in the theorem statement. Moreover, we fix initialization pair $(w_0, z_0)$ (respectively, $(x_0, y_0)$) such that $w_0 \in \mathrm{ri}(\mathcal{W}) \setminus \mathrm{NE}(G)$ (respectively, $x_0 \in \mathrm{ri}(\mathcal{X}) \setminus \mathrm{NE}(V)$).

**2. Convergence rate to sublevel set boundary for symmetrized system.**

First, recall that the operator $G$ is linear by assumption, which also implies $V$ is linear. In this case, we have by Proposition B.6 that $\mathrm{NE}(V) = \mathrm{MEQ}(V)$. Now recall for $t \geq 0$ that $x_t \in \mathcal{R}(x_0) \subseteq \mathrm{ri}(\mathcal{X})$. Moreover, observe that combining Proposition E.2 and the result of Lemma I.1 establishes the existence and uniqueness of a point $\widetilde{x} \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$. In particular, note that it suffices to apply Lemma I.1 to an auxiliary OFTRL trajectory initialized at the same pair $(x_0, y_0)$, where it then follows that its limit is a point $\widetilde{x} \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$, which establishes existence. Proposition E.2 then implies this point is unique.

Now let $x^\star \in \mathrm{NE}(V) \cap \mathrm{ri}(\mathcal{X})$ be the point from the setup above. As $\mathrm{NE}(V) = \mathrm{MEQ}(V)$, it follows also that $x^\star \in \mathrm{MEQ}(V)$. Then following arguments identical to part (ii) of Lemma I.1, we further obtain for all $t \geq 0$ that

$$\mathsf{D}_R(x^\star, x_t) = \mathsf{D}_R(x^\star, \widetilde{x}) + \mathsf{D}_R(\widetilde{x}, x_t)\,. \tag{157}$$

For convenience, we will define

$$a_t := \mathsf{D}_R(x^\star, x_t)\,, \quad b_t := \mathsf{D}_R(\widetilde{x}, x_t)\,, \quad c := \mathsf{D}_R(x^\star, \widetilde{x})\,. \tag{158}$$

Then (157) says that $a_t = b_t + c$, where $c \geq 0$ by convexity of $R$.

Now, by Lemma L.1 and Proposition 5.3, we obtain for all $t \geq 0$ and $x^\star \in \mathrm{MEQ}(V) = \mathrm{NE}(V)$ that

$$\mathsf{D}_R(x^\star, x_{t+1}) - \mathsf{D}_R(x^\star, x_t) = F_{x^\star}(y_{t+1}) - F_{x^\star}(y_t) \geq \tfrac{1}{4} \cdot \eta^2 \|V(x_t)\|_{y_t}^2\,. \tag{159}$$

As both $\widetilde{x} \in \mathrm{MEQ}(V)$ and $x^\star \in \mathrm{MEQ}(V)$, it follows that we can write for all $t \geq 0$ that

$$a_{t+1} - a_t = b_{t+1} - b_t \geq \tfrac{1}{4} \cdot \eta^2 \|V(x_t)\|_{y_t}^2 \geq 0\,. \tag{160}$$

For $\alpha$ as in the statement of the theorem, we then define the sublevel set

$$\mathcal{C}_\alpha := \{x \in \mathrm{ri}(\mathcal{X}) : \mathsf{D}_R(x^\star, x) \leq \alpha\} \subset \mathrm{ri}(\mathcal{X})\,.$$

As $x^\star \in \mathrm{ri}(\mathcal{X})$ by assumption, it holds from Proposition C.7 that $\mathcal{C}_\alpha \subset \mathrm{ri}(\mathcal{X})$ is compact. Also, for the assumptions on $x_0$ (as implied by the condition on $w_0$) in the theorem statement, we have $x_0 \in \mathcal{C}_\alpha$. Moreover, we have from the identity of (157) at time $t = 0$ that

$$\mathsf{D}_R(x^\star, \widetilde{x}) \leq \mathsf{D}_R(x^\star, \widetilde{x}) + \mathsf{D}_R(\widetilde{x}, x_0) = \mathsf{D}_R(x^\star, x_0) \leq \alpha\,,$$

which also implies that $\widetilde{x} \in \mathcal{C}_\alpha$. By definition of $T_\alpha(x^\star)$, and using the monotonicity of $\{a_t\}$ established in (160), we then further have that $x_t \in \mathcal{C}_\alpha \cap \mathcal{R}(x_0)$ for all $0 \leq t \leq T_\alpha(x^\star)$.

Together, this allows us to apply the operator domination inequality of Proposition 5.6 for the point $\widetilde{x} \in \mathrm{MEQ}(V) \cap \mathcal{R}(x_0)$, for the compact set $\mathcal{C}_\alpha \subset \mathrm{ri}(\mathcal{X})$, and for the FTRL iterates $(x_t, y_t)$ satisfying $x_t = \nabla F(y_t) \in \mathcal{R}(x_0) \cap \mathcal{C}_\alpha$ for all $0 \leq t \leq T_\alpha(x^\star)$. This yields for all such $t$ that

$$\|V(x_t)\|_{y_t}^2 \geq \widetilde{\gamma} \cdot \mathsf{D}_R(\widetilde{x}, x_t) = \widetilde{\gamma} \cdot b_t \tag{161}$$

for a positive constant $\widetilde{\gamma} := \widetilde{\gamma}(x_0, \mathcal{C}_\alpha) > 0$.

Now, to relate (161) back to the term $a_t = \mathsf{D}_R(x^\star, x_t)$, observe that $b_t \geq b_0$ and $a_t \geq a_0$ due to (160), and moreover $b_0 = \mathsf{D}_R(\widetilde{x}, x_0) > 0$ and $a_0 = \mathsf{D}_R(x^\star, x_0) > 0$ by strict convexity of $R$ and due to the assumption that $x_0 \notin \mathrm{NE}(V)$. As also $c \geq 0$ by convexity of $R$, we have the relationships:

$$\frac{b_t}{a_t} = \frac{b_t}{b_t + c} \geq \frac{b_0}{b_0 + c} = \frac{b_0}{a_0} \,, \tag{162}$$

where the final equality is due to $b_0 + c = a_0$ from (157). Then combining the inequalities (160), (161), and (162), we obtain the bound

$$a_{t+1} \geq a_t + \tfrac{1}{4} \cdot \widetilde{\gamma}\eta^2 \cdot b_t \geq \left(1 + \tfrac{\widetilde{\gamma} b_0}{4 a_0}\eta^2\right) \cdot a_t \,, \tag{163}$$

for all $0 \leq t \leq T_\alpha(x^\star)$. Then define the constant $\beta := \frac{\widetilde{\gamma} b_0}{4 a_0} > 0$. From (163), it then follows recursively for all $0 \leq t \leq T_\alpha(x^\star)$ that

$$\mathsf{D}_R(x^\star, x_t) = a_t \geq a_0 \cdot (1 + \beta\eta^2)^t = \mathsf{D}_R(x^\star, x_0) \cdot (1 + \beta\eta^2)^t \,. \tag{164}$$

**3. Translating back to original system.**

It remains to translate the bound in (164) back to the original system. For this, recall that we have $x_t = M^{1/2} w_t$, $x_0 = M^{1/2} w_0$, and $x^\star = M^{1/2} w^\star$. Then using the relationships of (13):

$$m_{\max} \cdot \mathsf{D}_Q(w^\star, w_t) \geq \mathsf{D}_R(x^\star, x_t) \quad \text{and} \quad \mathsf{D}_R(x^\star, x_0) \geq m_{\min} \mathsf{D}_Q(w^\star, w_0) \,. \tag{165}$$

for the constant $\beta$ defined in step 2. Then substituting (165) into (164) and rearranging yields for all $0 \leq t \leq T_\alpha(x^\star)$ that

$$\mathsf{D}_Q(w^\star, w_t) \geq \tfrac{m_{\min}}{m_{\max}} \cdot \mathsf{D}_Q(w^\star, w_0) \cdot (1 + \beta\eta^2)^t \,.$$

Setting $\beta(w_0, w^\star, \alpha) := \beta > 0$ and recalling that $T_\alpha(x^\star) = T_\alpha(w^\star)$ then yields the statement of the theorem. ■