# On the Optimality of Dilated Entropy and Lower Bounds for Online Learning in Extensive-Form Games

Zhiyuan Fan [1] Christian Kroer [2] Gabriele Farina [1]

[1]MIT [2]Columbia University

## Main question

Distance-generating functions provide generalized notions of distances
Constrained optimization require a DGF to move appropriately in the feasible set

**What is the optimal DGF for extensive-form games?**

## Key results

- Developed a new analysis for OMD based on the diameter/convexity ratio results of the DilEnt DGF.
- Established a matching regret lower bound, confirming the implied optimality of DilEnt.
- Achieved a new state-of-the-art convergence rate for Clairvoyant OMD to CCE.

| Regularizer | Norm pair | $\lvert\mathcal{D}\rvert/\mu$ ratio | Max gradient norm |
|---|---|---|---|
| Dilated Entropy [Kroer et al., 2020] | $\ell_1$ and $\ell_\infty$ norms | $\mathcal{O}(2^D\lVert\mathcal{Q}\rVert_1^2\log\lvert\mathcal{A}\rvert)$ | $\le 1$ |
| Dilated GI. Entropy [Farina et al., 2021] | $\ell_1$ and $\ell_\infty$ norms | $\mathcal{O}(\lVert\mathcal{Q}\rVert_1^2\log\lvert\mathcal{A}\rvert)$ | $\le 1$ |
| **DilEnt (this paper)** | treeplex norms | $\ln\lvert\mathcal{V}\rvert$ | $\le 1$ |

Table 1. Comparison of diameter/convexity ($\lvert\mathcal{D}\rvert/\mu$) ratio with prior results. It holds that $\ln\lvert\mathcal{V}\rvert \le \mathcal{O}(\lVert\mathcal{Q}\rVert_1\log\lvert\mathcal{A}\rvert)$

## Tree-Form Sequential Decision Problem (TFSDP)

**TFSDP** models the decision-making process of players in extensive-form games.

- A **decision point** $j \in \mathcal{J}$ corresponds to an information set for the player.
- An **observation point** $ja \in \Sigma$ represents the state that follows an action.

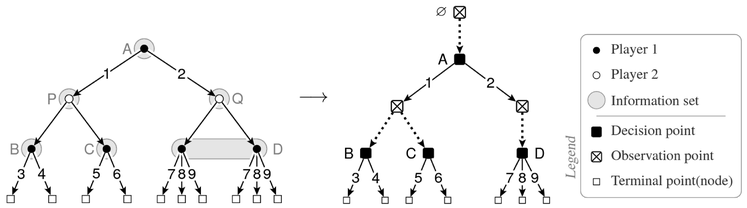


Figure 1. An example of an EFG and its corresponding TFSDP. Notably, $p_{\mathsf{B}} = \mathsf{A1}$.

**Sequence-form strategy space** $\mathcal{Q} \subseteq \mathbb{R}^\Sigma$ structures strategies with linearity. Each entry $\mathbf{x}[ja]$ represents the probability of "reaching" the intermediate state $ja$, satisfying

$$\mathbf{x}[\varnothing] = 1, \qquad \sum_{a\in\mathcal{A}_j} \mathbf{x}[ja] = \mathbf{x}[p_j] \quad \forall j \in \mathcal{J}.$$

Here, $p_j$ denote the preceding observation point of decision point $j$.

## Problem Setting

**Online learning** on sequence-form strategy space $\mathcal{Q}$ with full-information feedback:

- In each round, the agent picks $\mathbf{x}_t \in \mathcal{Q}$.
- They observe the adversarial reward vector $\mathbf{w}_t$ and the reward $\langle \mathbf{x}_t, \mathbf{w}_t\rangle \in [-1, 1]$.

**Goal:** Minimize the cumulative regret:

$$\mathrm{Regret}(T) := \max_{\mathbf{x}_*\in\mathcal{Q}} \sum_{t=1}^{T} \langle \mathbf{x}_* - \mathbf{x}_t, \mathbf{w}_t\rangle.$$

## Proximal Method

**Proximal step** generalizes the notion of gradient ascent to restricted space $\mathcal{Q}$:

$$\mathbf{x}_{t+1} \leftarrow \Pi_\varphi(\eta\mathbf{g}, \mathbf{x}_t) := \operatorname*{argmax}_{\widehat{\mathbf{x}}\in\mathcal{Q}} \{\langle \eta\mathbf{g}, \widehat{\mathbf{x}}\rangle - \mathcal{D}_\varphi(\widehat{\mathbf{x}}\|\mathbf{x}_t)\}.$$

Here, $\varphi$ is a **distance-generating function (DGF)** and $\mathcal{D}_\varphi$ is the *Bregman divergence*.

Given the primal-dual norms $\|\cdot\|$ and $\|\cdot\|_*$ with $\|\mathbf{w}\|_* \le 1$, the performance of OMD is determined by:

- The strong **convexity** $\mu$ of DGF over the primal norm $\|\cdot\|$
  ▷ A larger convexity ⇒ enables a larger step size ⇒ **easier to learn**
- The **diameter** $\lvert\mathcal{D}\rvert$ of the strategy space measured by DGF
  ▷ A smaller diameter ⇒ a smaller search space ⇒ **easier to learn**

⇒ The **diameter/convexity ratio** $\lvert\mathcal{D}\rvert/\mu$ is the key factor in the performance of mirror descent.

We inspect the **weight-one dilated entropy (DilEnt)** in TFSDP:

$$\varphi_{\mathbf{1}} : \mathcal{Q} \ni \mathbf{x} \mapsto \sum_{j\in\mathcal{J}}\sum_{a\in\mathcal{A}} \mathbf{x}[ja] \ln\left(\frac{\mathbf{x}[ja]}{\mathbf{x}[p_j]}\right)$$

▷ When TFSDP only has one decision point, $\varphi_{\mathbf{1}}$ reduces to entropy.

## Primal-Dual Treeplex Norms

We introduce a pair of primal-dual norms: the **treeplex** $\ell_1$**-norm** $\|\cdot\|_{\mathcal{H},1}$ and the **treeplex** $\ell_\infty$**-norm** $\|\cdot\|_{\mathcal{H},\infty}$:

$$\|\mathbf{u}\|_{\mathcal{H},1} := \sup_{\mathbf{y}\in\mathcal{Q}^*} \langle |\mathbf{u}|, \mathbf{y}\rangle, \qquad \|\mathbf{u}\|_{\mathcal{H},\infty} := \sup_{\mathbf{x}\in\mathcal{Q}} \langle |\mathbf{u}|, \mathbf{x}\rangle.$$

Here, $\mathcal{Q}^*$ is the **dual polytope** of $\mathcal{Q}$. Both treeplex norms can be computed via **recursion**:

$$\|\mathbf{u}\|_{\mathcal{H}_p,1} := \max_{j:p_j=p} \sum_{a\in\mathcal{A}_j} \|\mathbf{u}\|_{\mathcal{H}_{ja},1}, \qquad \|\mathbf{u}\|_{\mathcal{H}_p,\infty} := \sum_{j:p_j=p} \max_{a\in\mathcal{A}_j} \|\mathbf{u}\|_{\mathcal{H}_{ja},\infty}.$$

## Key Structural Results

**[Lemma 4.4]** The **dual norm** of any feasible reward vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_{\mathcal{H},\infty} \le 1$.

**[Lemma 5.1]** DilEnt is 1-strongly **convex** with respect to the treeplex $\ell_1$-norm: $\|\cdot\|^2_{\nabla^2\varphi_{\mathbf{1}}(\mathbf{x})} \ge \|\cdot\|^2_{\mathcal{H},1}$.

**[Lemma 5.2]** The **diameter** of the strategy space is upper bounded as $\max_{x_*}\mathcal{D}_{\varphi_{\mathbf{1}}}(x_*, x_1) \le \ln|\mathcal{V}|$.

## Main Results on Regret Upper Bounds

**[Theorem 5.4]** Running OMD with DilEnt during the proximal steps achieves a regret bound of

$$\mathrm{Regret}(T) \le \sqrt{2\ln|\mathcal{V}|}\sqrt{T},$$

where $|\mathcal{V}| := \mathcal{Q}\cap\{0,1\}^\Sigma$ is the number of pure strategies.

**[Theorem 5.6]** If every player runs Clairvoyant OMD with DilEnt in an $n$-player game, the average joint policy converges to a *Coarse Correlated Equilibrium* at the rate

$$\varepsilon \le \mathcal{O}(n\log|\mathcal{V}|\log T/T).$$

## Lower Bound Implies Optimality of DilEnt

**[Theorem 6.1 + Theorem 6.2]** Any algorithm incurs an expected regret of at least

$$\mathrm{Regret}(T) \ge \widetilde{\Omega}(\sqrt{\ln|\mathcal{V}|}\sqrt{T})$$

*Under certain structural assumptions, the bound is free from hidden logarithmic factors.*

DilEnt achieves an nearly-optimal diameter/convexity ratio:

- **Better DGF ⇒ better regret upper bound** in Theorem 5.4
- A contradiction arises since superior regret upper bounds violate the lower bound above.

Checkout our paper on openreview!

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Better regularization for sequential decision spaces: Fast convergence rates for nash, correlated, and team equilibria. In *Proceedings of the 2021 ACM Conference on Economics and Computation*, 2021.

Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, pages 1–33, 2020.