Regret Decomposition in Sequential Games with Convex Action Spaces and Losses[*]

Gabriele Farina, Christian Kroer, Tuomas Sandholm
Computer Science Department
Carnegie Mellon University
{gfarina,ckroer,sandholm}@cs.cmu.edu

Abstract

We derive a new framework for regret minimization on sequential decision problems and extensive-form games with general compact convex sets at each decision point and general convex losses, as opposed to prior work which has been for simplex decision points and linear losses. We call our framework *laminar regret decomposition*. It generalizes the CFR algorithm to this more general setting. Furthermore, our framework enables a new proof of CFR even in the known setting, which is derived from a perspective of decomposing polytope regret, thereby leading to an arguably simpler interpretation of the algorithm. Our generalization to convex compact sets and convex losses allows us to develop new algorithms for several problems: regularized sequential decision making, regularized Nash equilibria in extensive-form games, and computing approximate extensive-form perfect equilibria. Our generalization also leads to the first regret-minimization algorithm for computing reduced-normal-form quantal response equilibria based on minimizing local regrets.

Introduction

Counterfactual regret minimization (CFR) [Zinkevich *et al.*, 2007] is a methodology for setting up regret minimization for sequential decision problems (whether single- or multi-agent), where each decision point requires selecting either an action or a point from the probability distribution over actions. The crux of CFR is counterfactual regret, which leads to a definition of regret local to each decision point. CFR can then be viewed as the observation, and proof, that bounds on counterfactual regret, which can be minimized locally, lead to bounds on the overall regret. To minimize local regret, the framework relies on regret minimizers that operate on a simplex (typically of probabilities over the available actions), such as *regret matching* (RM) [Blackwell, 1956] or the newer variant *regret matching*$^+$ (RM$^+$) [Tammelin *et al.*, 2015].

In this extended abstract we consider the more general problem of how to minimize regret over a sequential decision-making (SDM) polytope, where we allow arbitrary compact convex subsets of simplexes at each decision point (as opposed to only simplexes in CFR), and general convex loss functions (as opposed to only linear losses in CFR). This allows us to model a form of online convex optimization over SDM polytopes. We derive a decomposition of the polytope regret into local regret at each decision point. This allows us to minimize regret locally as with CFR, but for general compact convex decision points and convex losses. We call our decomposition *laminar regret decomposition (LRD)*. We call our overall framework for convex losses and compact convex decision points *laminar regret minimization (LRM)*. As a special case, our framework provides an alternate view of why CFR works—one that may be more intuitive for those with a background in online convex optimization.

[*] Parts of this abstract will appear at AAAI'19.

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada.

Our generalization to general compact convex sets allows us to model entities such as ϵ-perturbed simplexes [Farina and Gatti, 2017; Farina *et al.*, 2017; Kroer *et al.*, 2017a], and thus yields new algorithms for computing approximate equilibrium refinements for EFGs.

General convex losses in SDM and EFG contexts have, to the best of our knowledge, not been considered before. This generalization enables fast algorithms for many new settings. One is to compute regularized zero-sum equilibria. If we apply a convex regularization function at each simplex, we can apply our framework to solve the resulting game. For the negative entropy regularizer this is equivalent to the dilated entropy distance function used for solving EFGs with first-order methods [Hoda *et al.*, 2010; Kroer *et al.*, 2015, 2017b]. Ling *et al.* [2018] show that dilated-entropy-regularized EFGs are equivalent to quantal response equilibria (QRE) in the corresponding reduced normal-form game. Thus our result yields the first regret-minimization algorithm for computing reduced-normal-form quantal response equilibria in EFGs.

Regret Minimization and Sequential Decision Making

We work the online learning framework called *online convex optimization* [Zinkevich, 2003]. In this setting, a decision maker repeatedly plays against an unknown environment by making a sequence of decisions $x^1, x^2, \ldots$. As customary, we assume that the set $X \subseteq \mathbb{R}^n$ of all possible decisions for the decision maker is convex and compact. The outcome of each decision x^t is evaluated as $\ell^t(x^t)$, where ℓ^t is a convex function *unknown* to the decision maker until the decision is made. Abstractly, a *regret minimizer* is a device that supports two operations: (1) it gives a *recommendation* for the next decision $x^{t+1} \in X$; (2) it receives/observes the convex loss function ℓ^t used to "evaluate" x^t.

It turns out that the results of this paper can be proven in a general setting which we call a sequential decision making. Formally, we assume that we have a set of decision points $\mathcal{J}$. Each decision point $j \in \mathcal{J}$ has a set of actions A_j of size n_j. The decision space at each decision point j is represented by a convex set $X_j \subseteq \Delta_{n_j}$. A point $x_j \in X_j$ represents a probability distribution over A_j. When a point x_j is chosen, an action is sampled randomly according to x_j. Given a specific action at j, the set of possible decision points that the agent may next face is denoted by $\mathcal{C}_{j,a}$. It can be an empty set if no more actions are taken after j, a. We assume that the decision points form a tree, that is, $\mathcal{C}_{j,a} \cap \mathcal{C}_{j',a'} = \emptyset$ for all other convex sets and action choices j', a'. This condition is equivalent to the perfect-recall assumption in extensive-form games, and to conditioning on the full sequence of actions and observations in a finite-horizon partially-observable decision process.

Regret in Sequential Decision Making

We assume that we are playing a sequence of T iterations of a sequential decision process. At each iteration t we choose a strategy $x \in X$ and are then given a loss function of the form

$$\ell^t(x) := \sum_{j \in \mathcal{J}} \pi_j(x) \ell^t_j(x_j), \tag{1}$$

where $\ell^t_j : X_j \to \mathbb{R}$ is a convex function for each $j \in \mathcal{J}$. We coin loss functions of this form *separable*, and they will play an important role in our results. Our goal is to compute a new strategy vector x^t such that the regret across all T iterations is as low as possible against any sequence of loss functions.

We now summarize definitions for the value and regret associated with convex sets and strategies. First we have the value of convex set j at iteration t when following strategy $\hat{x}$:

$$\hat{V}^t_{\triangle_j}(\hat{x}_{\triangle_j}) := \ell^t_j(\hat{x}^t_j) + \sum_{a \in A_j} \sum_{j' \in \mathcal{C}_{j,a}} \hat{x}_{j,a} \hat{V}^t_{\triangle_{j'}}(\hat{x}_{\triangle_{j'}}).$$

This definition denotes the utility associated with starting at convex set X_j rather than at the root. Thus we have exchanged the term $\pi_j(x)$ with one for ℓ^t_j and with $\hat{x}_{j,a}$ for $V^t_{\triangle_{j'}}$; this allows us to write the value as a recurrence. We will be particularly interested in the value of x^t, which we denote $V^t_{\triangle_j} := \hat{V}^t_{\triangle_j}(x^t_{\triangle_j})$. Now we can define the cumulative regret at convex set j across all T iterations as

$$R^T_{\triangle_j} := \sum_{t=1}^{T} V^t_{\triangle_j} - \min_{\hat{x}_{\triangle_j}} \sum_{t=1}^{T} \hat{V}^t_{\triangle_j}(\hat{x}_{\triangle_j}). \tag{2}$$

Saddle Point Problems

In an extensive-form game with perfect recall each player faces a sequential decision-making problem, of the type described in the previous section. The set of next potential decision points $\mathcal{C}_{j,a}$ is based on observations of stochastic outcomes and actions taken by other players. Here, we will focus on two-player zero-sum EFGs with perfect recall. In particular, we assume that we are solving the convex-concave saddle-point problem

$$\min_{x\in X}\max_{y\in Y}\left\{\mu(x)^\top A\mu(y) + d_1(\mu(x)) - d_2(\mu(y))\right\}, \tag{3}$$

where X is the SDM polytope for Player 1, Y is the SDM polytope of Player 2, and μ is the function that maps a strategy to its sequence-form representation [von Stengel, 1996]. Each d_i is assumed to be a dilated convex function of the form given in (1).

In standard EFGs, the loss function for each player at each iteration t is defined to be the negative payoff vector associated with the sequence-form strategy of the other player at that iteration; since we additionally allow a *regularization* term we also get a nonlinear convex term. More formally, at each iteration t, the loss functions $\ell^t_X : X \to \mathbb{R}$ and $\ell^t_Y : Y \to \mathbb{R}$ for player 1 and 2 respectively are defined as $\ell^t_X : x \mapsto \langle -A\mu(y^t), \mu(x)\rangle + d_1(x)$, $\ell^t_Y : y \mapsto \langle A^\top \mu(x^t), \mu(y)\rangle + d_2(y)$, where A is the sequence-form payoff matrix of the game [von Stengel, 1996]. Some simple algebra shows that ℓ^t_X and ℓ^t_Y are indeed separable (that is, they can be written in the form of Equation 1), where each decision-point-level loss $\ell^t_{j,X}$ and $\ell^t_{j,Y}$ is a convex function.

A folk theorem explains the tight connection between low-regret strategies and approximate Nash equilibria. We will need a more general variant of that theorem generalized to (3). The convergence criterion we are interested in is the *saddle-point residual (or gap)* ξ of $(\bar{x}, \bar{y})$, defined as $\xi = \max_{\hat{y}}\{d_1(\bar{x}) - d_2(\hat{y}) + \langle \bar{x}, A\hat{y}\rangle\} - \min_{\hat{x}}\{d_1(\hat{x}) - d_2(\bar{y}) + \langle \hat{x}, A\bar{y}\rangle\}$. We show that playing the average of a sequence of regret-minimizing strategies leads to a bounded saddle-point residual. This result is probably known, but it is unclear whether it has been stated in the form here. by Nemirovski [2004].

Theorem 1. *If the average regret accumulated on X and Y by the two sets of strategies $\{x_t\}_{t=1}^T$ and $\{y_t\}_{t=1}^T$ is ϵ_1 and ϵ_2, respectively, then any strategy profile $(\bar{x}, \bar{y})$ such that $\mu(\bar{x}) = \frac{1}{T}\sum_{t=1}^T \mu(x^t), \quad \mu(\bar{y}) = \frac{1}{T}\sum_{t=1}^T \mu(y^t)$ has a saddle-point residual bounded by $\epsilon_1 + \epsilon_2$.*

The above averaging is performed in the sequence-form space, which works because that space is also convex. After averaging we can easily compute $\bar{x}$ in linear time. Hence, by applying LRD to the decision spaces X and Y, we converge to a small saddle-point residual. The fact that the averaging of the strategies is performed in sequence form explains why the traditional CFR presentation requires averaging with weights based on the player's reaches π_j at each decision point j.

Laminar Regret Decomposition

We now define a new parameterized class of loss functions for each subtree X_j which we will show can be used to minimize regret over X by minimizing that loss function independently at each convex set X_j. The loss function is

$$\hat{\ell}^t_j(x_j) := \ell^t_j(x_j) + \sum_{a\in A_j}\sum_{j'\in\mathcal{C}_{j,a}} x_{j,a} V^t_{\triangle_{j'}}. \tag{4}$$

It is convex since ℓ^t_j is convex by hypothesis and we are only adding a linear term to it. Strict convexity is also preserved, and for strongly convex losses, the strong convexity parameter remains unchanged.

We now prove that the regret at information set j decomposes into regret terms depending on $\hat{\ell}^t_j$ and a sum over the regret at child convex sets:

Theorem 2. *The cumulative regret at a decision point j can be decomposed as $R^T_{\triangle_j} = \sum_{t=1}^T \hat{\ell}^t_j(x^t) - \min_{\hat{x}_j\in X_j}\left\{\sum_{t=1}^T \hat{\ell}^t_j(\hat{x}_j) - \sum_{a\in A_j}\sum_{j'\in\mathcal{C}_{j,a}} \hat{x}_{j,a} R^T_{\triangle_{j'}}\right\}$.*

Theorem 2 justifies the introduction of the concept of *laminar regret* at each decision point $j \in \mathcal{J}$: $\hat{R}^T_j := \sum_{t=1}^T \hat{\ell}^t_j(x^t_j) - \min_{\hat{x}_j\in X_j}\sum_{t=1}^T \hat{\ell}^t_j(\hat{x}_j)$. With this, we can write the cumulative subtree regret at decision point j as a sum of laminar regret at j plus a recurrence term for each child decision point. Applying this inductively gives the following theorem which tells us how one can apply regret minimization locally on laminar regrets in order to minimize regret in SDMs:

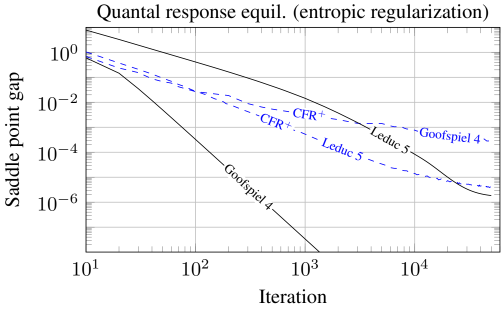

Figure 1: The QRE saddle-point gap as a function of the number of iterations for each game. The convergence rates of CFR^{+} for Nash equilibrium is shown for reference.

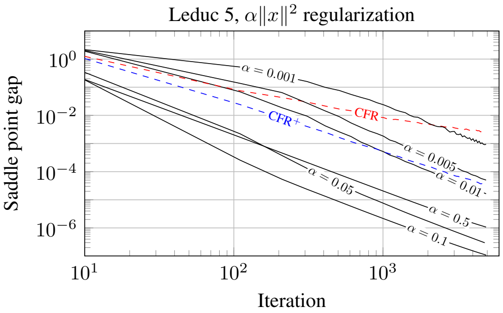

Figure 2: The saddle-point gap as a function of the number of iterations for ℓ_2-regularized Leduc 5 for varying regularization amounts. The convergence rates of CFR^{+} for Nash equilibrium is shown for reference.

Theorem 3. *The cumulative regret on X satisfies $R^T \leq \max_{\hat{x}\in X} \sum_{j\in\mathcal{J}} \pi_j(\hat{x})\hat{R}_j^T$. Hence, if each individual laminar regret $\hat{R}_j^T$ on each of the convex domains X_j grows sublinearly, overall regret on X grows sublinearly.*

Theorem 3 shows that overall regret can be minimized by minimizing each laminar regret separately. In particular, this means that if we have a regret minimizer for each decision point j that can handle the structure of the convex set X_j and the convex loss from (4), then we can apply those regret minimizers individually at each information set. Our result gives an alternative proof of CFR. This is arguably simpler than existing proofs, because we show directly why regret over a sequential decision-making space decomposes into individual regret terms, as opposed to bounding terms in order to fit the CFR framework. Finally, our result also generalizes CFR to new settings, as it can be implemented on arbitrary convex subsets of simplexes and with convex losses rather than linear.

Experiments

We conducted multiple kinds of experiments on two EFG settings. The first game is Leduc 5 poker [Southey *et al.*, 2005], a standard benchmark in imperfect-information game solving. The second game is a variant of Goofspiel [Ross, 1971], a bidding game where each player has a hand of cards numbered 1 to N (we use $N = 4$ in our experiments).

First we investigate a setting where no previous regret-minimization algorithms based on minimizing regret locally existed: the computation of QREs via LRM and our more general convex losses. Ling *et al.* [2018] use Newton's method for this setting, but, as with standard Nash equilibrium, second-order algorithms do not scale to large games (this is why CFR^{+} has been so successful for creating human-level poker AIs). We compare how quickly we can compute QREs compared to how quickly Nash equilibria can be computed, in order to understand how large games we can expect to find QREs for with our approach. To do this we run LRM with online gradient descent (OGD) at each decision point. Because OGD is not guaranteed to stay within the simplex at each iteration we need to project; this can be implemented via binary search for decision points with large dimension [Duchi *et al.*, 2008], and via a constant-size decision tree for low-dimension decision points. The results are shown in Figure 1. We see that LRM performs extremely well; in Goofspiel it converges vastly faster than CFR^{+}, and in Leduc 5 it converges at a rate comparable to CFR^{+} and eventually becomes faster. This shows that QRE computation via LRM likely scales to extremely large EFGs, such as real-world-sized poker games (since CFR^{+} is known to scale to such games).

In the second set of experiments we investigate the speed of convergence for solving ℓ_2-regularized EFGs. Again we include the convergence rate of standard CFR and CFR^{+} for Nash equilibrium computation as a benchmark. The results for Leduc 5 are in Figure 2. Solving the regularized game is much faster than computing an Nash equilibrium via CFR^{+} except for extremely small amounts of regularization.

In the full paper [Farina *et al.*, 2019] we also investigate the performance of LRM in a single-agent-learning setting: learning how to exploit a static opponent where we observe repeated samples from their strategy. We consider a setting where the exploiter wishes to maximally exploit subject to staying near a pre-computed Nash equilibrium in order to avoid opening herself up to exploitability [Ganzfried and Sandholm, 2011]. Our experiments show that this model can indeed be used as a scalable proxy for trading off exploitation and exploitability.

Acknowledgments

This material is based on work supported by the National Science Foundation under grants IIS-1718457, IIS-1617590, and CCF-1733556, and the ARO under award W911NF-17-1-0082. Christian Kroer is supported by a Facebook Fellowship.

References

David Blackwell. An analog of the minmax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

Gabriele Farina and Nicola Gatti. Extensive-form perfect equilibrium computation in two-player games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Regret minimization in behaviorally-constrained zero-sum games. In *International Conference on Machine Learning (ICML)*, 2017.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Online convex optimization for sequential decision processes and extensive-form games. Available at `https://arxiv.org/abs/1809.03075`, 2019.

Sam Ganzfried and Tuomas Sandholm. Game theory-based opponent modeling in large imperfect-information games. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2011.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2), 2010.

Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster first-order methods for extensive-form game solving. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2015.

Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Smoothing method for approximate extensive-form perfect equilibrium. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Theoretical and practical advances on smoothing for extensive-form games. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2017.

Chun Kai Ling, Fei Fang, and J Zico Kolter. What game are we playing? End-to-end learning in normal and extensive form games. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

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), 2004.

Sheldon M Ross. Goofspiel–the game of pure strategy. *Journal of Applied Probability*, 8(3):621–625, 1971.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2005.

Oskari Tammelin, Neil Burch, Michael Johanson, and Michael Bowling. Solving heads-up limit Texas hold'em. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

Martin Zinkevich, Michael Bowling, Michael Johanson, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML)*, pages 928–936, Washington, DC, USA, 2003.