A. Correlated Equilibrium

In this section we define the differences between two competing definitions of approximate correlated equilibrium (ϵ-CE) and define the related solution concept approximate coarse correlated equilibrium (ϵ-CCE).

A.1. Correlated Equilibrium

An approximate correlated equilibrium (ϵ-CE) is one where the advantage of a single player unilaterally switching away from a recommended action is no more than ϵ. When $\epsilon = 0$, the standard CE is recovered. Define $A_p(a'_p, a_p, a_{-p}) = G(a'_p, a_{-p}) - G(a_p, a_{-p})$ as the advantage for player p switching action from a_p to a'_p, when other players play a_{-p}. This relationship is described mathematically in Equation 13.

$$\sum_{a_{-p}} \sigma(a_{-p}|a_p) A_p(a'_p, a_p, a_{-p}) \leq \epsilon_p \tag{13}$$

$$\sum_{a_{-p}} \sigma(a_{-p}, a_p) A_p(a'_p, a_p, a_{-p}) \leq \sigma(a_p)\epsilon_p \tag{14}$$

$$\sum_{a_{-p}} \sigma(a_{-p}, a_p) \left(A_p(a'_p, a_p, a_{-p}) - \epsilon_p \right) \leq 0 \tag{15}$$

$$\forall p \in \mathcal{P}, a'_p \neq a_p \in \mathcal{A}_p$$

Together, A_p and ϵ represent the CE linear inequality constraints. Mathematically these are equations of a plane, and separate the mixed joint probability $\sigma(a)$ into half-spaces. Together these half-spaces intersect to form a convex polytope of valid CE solutions.

A.2. Alternate Form Correlated Equilibrium

Sometimes another definition for CEs is used which is not equivalent to the definition above when $\epsilon \neq 0$. We call Equation 16 the alternate approximate CE. In matrix form, we can simply write $A\sigma \leq \epsilon$.

$$\sum_{a_{-p}} \sigma(a_{-p}, a_p) A_p(a'_p, a_p, a_{-p}) \leq \epsilon_p \tag{16}$$

$$\forall p \in \mathcal{P}, a'_p \neq a_p \in \mathcal{A}_p$$

This form is often easier to deal with computationally (particularly with $\min \epsilon$-MG(C)CE) because it does not require dealing with a conditional distribution, and the approximation term is independent of probabilities. We use this definition throughout this work, although MGCE is still well defined using the former definition.

A.3. Coarse Correlated Equilibrium

The coarse correlated equilibrium (CCE) is a looser solution concept where a player must decide if they are going to play the correlation device's recommendation before they receive the recommendation. Define $A_p(a'_p, a) = G(a'_p, a_{-p}) - G(a)$ as the gain from deviating before an action has been recommended.

$$\sum_{a} \sigma(a) A_p(a'_p, a) \leq \epsilon_p \tag{17}$$

$$\sum_{a} \sigma(a) \left(A_p(a'_p, a) - \epsilon_p \right) \leq 0 \tag{18}$$

$$\forall p \in \mathcal{P}, a'_p \in \mathcal{A}_p$$

Note that this can be derived from the correlated equilibrium by mixing over all possible actions, a_p, that an agent can take.

$$\sum_{a_p} \sigma(a_p) \sum_{a_{-p}} \sigma(a_{-p}|a_p) A_p(a'_p, a_p, a_{-p}) \leq \sum_{a_p} \sigma(a_p)\epsilon_p$$

$$\sum_{a_p} \sum_{a_{-p}} \sigma(a_{-p}, a_p) A_p(a'_p, a_p, a_{-p}) \leq \epsilon_p$$

$$\sum_{a} \sigma(a_{-p}, a_p) \left(G(a'_p, a_{-p}) - G(a_p, a_{-p}) \right) \leq \epsilon_p$$

$$\sum_{a} \sigma(a) A_p(a'_p, a) \leq \epsilon_p$$

Note that if one wished to solve for MGCCE, simply substitute the advantage matrix A^{CE} for A^{CCE}, without any other additional changes.

B. Generalized Entropy

Shannon's Entropy (Shannon, 1948), I_S, is a familiar quantity and is described as a measure of "information gain". The Gini Impurity (Breiman et al., 1984; Bishop, 2006) is a measurement of the probability of mis-classifying a sample of a discrete random variable, if that sample were randomly classified according to its own probability mass function, $I_G = \sum_i^N \sigma_i \sum_{j \neq i} \sigma_j = 1 - \sum_i^N \sigma_i^2$.

Both Shannon's entropy and Gini Impurity are maximized when the probability mass function is uniform $\sigma_i = \frac{1}{|\mathcal{A}|}$ and minimized when all mass is on a single outcome. Both metrics are used in decision tree classification algorithms, with Gini being more popular because it is easier to compute (Breiman et al., 1984).

In physics, there has been recent interest in non-extensive entropies which have been found to better model certain physical properties. One such entropy is called the Tsallis entropy, $I_T = \frac{1 - \sum_i \sigma_i^q}{q-1}$, (Tsallis, 1988; Havrda et al., 1967; Wang & Xia, 2017; Kaur & Buttar, 2019) and is parameterized by real q.

A notable property of the Tsallis entropy is that it is non-additive. Assume that we have two independent variables A

and B, with joint probability $P(A, B) = P(A)P(B)$, then the combined Tsallis entropy of this system is $I_T(A, B) = I_T(A) + I_T(B) + (1 - q)I_T(A)I_T(B)$. Therefore it can be seen that the $(1 - q)$ quantity is a measure of the departure from additivity, with additivity being recovered in the limit when $q \to 1$. This corresponds to the additive Shannon's entropy. The Gini impurity is recovered when $q = 2$. Therefore, the Gini impurity is a non-extensive generalized entropy.

C. Proofs of MG(C)CE Properties

C.1. Uniqueness and Existence

Theorem 1 (Uniqueness and Existence)**.** *MG(C)CE provides a unique solution to the equilibrium solution problem and always exists.*

Proof. The problem is a concave maximization problem with linear constraints so therefore has a unique solution. Existence follows from the fact that a CE always exists. □

C.2. Scalable Representation

Theorem 2 (Scalable Representation)**.** *The maximum Gini (C)CE, σ^*, has the following forms:*

$$\textit{General Support:} \quad \sigma^* = CA^T\alpha^* + C\beta^* + b \tag{19}$$

$$\textit{Full Support:} \quad \sigma^* = CA^T\alpha^* + b \tag{20}$$

Where e is a vector of ones, $|\mathcal{A}| = \prod_p |\mathcal{A}_p|$, $C = I - e^Tb$, and $b = \frac{1}{|\mathcal{A}|}e$ are constants. $\alpha^* \geq 0$ and $\beta^* \geq 0$ are the optimal dual variables of the solution, corresponding to the CE and distribution inequality constraints respectively.

Proof. Start with the equation we call the primal Lagrangian form.

$$\begin{aligned} L^\sigma_{\alpha,\beta,\lambda} = &-\frac{1}{2}\sigma^T\sigma - \alpha(A\sigma - \epsilon) - \beta^T\sigma \\ &+ \lambda(e^T\sigma - 1) \end{aligned} \tag{21}$$

To construct the dual Lagrangian we first take derivatives with respect to the primal variables σ, and set them equal to zero.

$$\begin{aligned} \frac{\partial L^\sigma_{\alpha,\beta,\lambda}}{\partial\sigma} &= \sigma^* - \left(A^T\alpha + \beta - \lambda\right) = 0 \implies \\ \sigma^* &= A^T\alpha + \beta - \lambda e \end{aligned} \tag{22}$$

These can be substituted back into the primal Lagrangian.

$$\begin{aligned} L_{\alpha,\beta,\lambda} &= -\frac{1}{2}\left[A^T\alpha + \beta - \lambda e\right]^T\left[A^T\alpha + \beta - \lambda e\right] \\ &\quad + \alpha^T e\epsilon - \lambda \\ &= -\frac{1}{2}\alpha^T AA^T\alpha - \frac{1}{2}\beta^T\beta - \alpha^T A\beta + \epsilon\alpha^T e \\ &\quad - \frac{|\mathcal{A}|}{2}\lambda^2 + \left(e^TA^T\alpha + e^T\beta - 1\right)\lambda \end{aligned}$$

Taking derivatives with respect to λ.

$$\begin{aligned} \frac{\partial L_{\alpha,\beta,\lambda}}{\partial\lambda} &= -|\mathcal{A}|\lambda^* + e^TA^T\alpha_p + e^T\beta - 1 = 0 \implies \\ \lambda^* &= \frac{1}{|\mathcal{A}|}\left(e^TA^T\alpha + e^T\beta - 1\right) \end{aligned} \tag{23}$$

And substituting back in:

$$\begin{aligned} L_{\alpha,\beta} = &-\frac{1}{2}\alpha^T AA^T\alpha - \frac{1}{2}\beta^T\beta - \alpha^T A\beta + \epsilon\alpha^T e \\ &+ \frac{1}{2|\mathcal{A}|}\left[e^TA^T\alpha + e^T\beta - 1\right]^T\left[e^TA^T\alpha + e^T\beta - 1\right] \\ = &-\frac{1}{2}\alpha^T AA^T\alpha - \frac{1}{2}\beta^T\beta - \alpha^T A\beta + \epsilon\alpha^T e \\ &+ \frac{1}{2|\mathcal{A}|}\alpha^T Aee^TA^T\alpha + \frac{1}{2N}\beta^T ee^T\beta \\ &- \frac{1}{2|\mathcal{A}|} + \frac{1}{|\mathcal{A}|}\alpha^T A_p ee^T\beta - \frac{1}{|\mathcal{A}|}e^TA^T\alpha \\ &- \frac{1}{|\mathcal{A}|}e^T\beta \end{aligned}$$

Doing some rearrangement.

$$\begin{aligned} L_{\alpha,\beta} = &\frac{1}{2}\alpha^T A\left(\frac{1}{|\mathcal{A}|}ee^T - I\right)A^T\alpha - \frac{1}{|\mathcal{A}|}e^TA^T\alpha \\ &+ \frac{1}{2}\beta^T\left(\frac{1}{|\mathcal{A}|}ee^T - I\right)\beta - \frac{1}{|\mathcal{A}|}e^T\beta \\ &+ \alpha^T A_p\left(\frac{1}{|\mathcal{A}|}ee^T - I\right)\beta + \epsilon\alpha^T e + \frac{1}{2|\mathcal{A}|} \end{aligned}$$

Remember that there are non-negative constraints on $\alpha \geq 0$ and $\beta \geq 0$. We therefore cannot easily solve for β to reduce this expression further. By defining $C = I - eb^T$, and $b^T = \frac{1}{|\mathcal{A}|}e^T$ (the uniform distribution), we arrive at the general support dual Lagrangian form.

$$\begin{aligned} L_{\alpha,\beta} = &-\frac{1}{2}\alpha^T ACA^T\alpha - b^TA^T\alpha + \epsilon^T\alpha \\ &- \frac{1}{2}\beta^TC\beta - b^T\beta - \alpha^T AC\beta + \frac{1}{2}b^Tb \end{aligned} \tag{24}$$

By combining Equations 22 and 23, we can arrive at an equation that describes the relationship between the primal and dual parameters.

$$\sigma^* = CA^T\alpha^* + C\beta^* + b \tag{25}$$

It is advantageous to try and obtain a more compact representation. We can achieve this if we assume σ has full support. In this case, $\beta = 0$, because none of the $\sigma \geq 0$ constraints are active and we obtain Equation 26 the full support dual Lagrangian form.

$$L_\alpha = -\frac{1}{2}\alpha^T ACA^T\alpha - b^T A^T \alpha + \epsilon^T \alpha + \frac{1}{2}b^T b \tag{26}$$

$$\sigma^* = CA^T\alpha^* + b \tag{27}$$

□

Theorem 3 (Existence of Full-Support ϵ-MG(C)CE)**.** *For all games, there exists an $\epsilon \leq \max(Ab)$ such that a full-support, ϵ-MG(C)CE exists. A uniform solution, b, always exists when $\max(Ab) \leq \epsilon$. When $\epsilon < \max(Ab)$, the solution is non-uniform.*

Proof. Note, $A\sigma \leq \epsilon \iff AC\sigma + Ab \leq \epsilon$, $Cb = 0$ and that b is the uniform distribution with maximum possible Gini impurity. Note that when $\max(Ab) \leq \epsilon$ the inequality will always hold with $\sigma = b$. And the inequality cannot hold with $\sigma = b$ when $\epsilon \leq \max(Ab)$. □

C.3. Family

Theorem 4. *For non-trivial games, the MG(C)CE lies on the boundary of the polytope and hence is a weak equilibrium.*

Proof. MG(C)CE is attempting to be near the uniform distribution. If the uniform distribution is not a (C)CE the MG(C)CE lies on the boundary of the (C)CE polytope, and by definition is weak. If the uniform distribution is a (C)CE, then it is also a NE (because it factorizes). It therefore lies on the polytope if it is a non-trivial game by (Nau et al., 2004). □

Table 1 summarizes the family of solutions that make up MG(C)CE. Note that a similar family can be defined for ME(C)CE.

C.4. Invariance

Theorem 5 (Affine Payoff Transformation Invariance)**.** *If σ^* is the ϵ-MG(C)CE of a game, $\mathcal{G}$, then for each player p independently we can transform the payoff tensors $\tilde{G}_p = c_p G_p + d_p$ and approximation vector $\tilde{\epsilon}_p = a_p \epsilon_p$ for some positive c_p and real d_p scalars, without changing the solution. Furthermore, if a game, $\mathcal{G}$ has (C)CE constraint matrix, A, and bound vector, ϵ, then each row can be scaled independently without changing the MG(C)CE.*

Proof. The only way that a game's payoff, G, influences the solution is via the (C)CE constraint matrices A_p. Recall that these are defined as the difference between action payoffs $a_p \neq a'_p \in \mathcal{A}_p$. It is easy to see that the constant d_p will cancel immediately.

$$\begin{aligned}\tilde{A}_{p,i,j} &= \tilde{G}_p(a'_p, a_{-p}) - \tilde{G}_p(a_p, a_{-p}) \\ &= c(G_p(a'_p, a_{-p}) - G_p(a_p, a_{-p}))\end{aligned} \tag{28}$$

Notice that A always appears alongside the dual variables α. Therefore any scale in $\tilde{A}\tilde{\alpha} = cA\tilde{\alpha}$ can be counteracted by $\tilde{\alpha} = \frac{\alpha}{c}$, without changing the nature of the optimization.

Similar to above, not only does α_p appear alongside A_p, each element appears alongside a particular row of A_p. Therefore not only can a whole A_p be scaled by a positive factor, each row of A_p can be scaled individually. Intuitively, each row of the (C)CE constraint matrix defines an equation of a plane in the simplex, and planes are not altered when scaled by a positive factor. We may exploit this property to better condition our optimization problem. □

D. MGCE Computation

There are several tricks that can be employed to simplify the nature of the computation problem.

D.1. Bounded Gradient Methods

It is easy to formulate gradient algorithms to solve for the MG(C)CE. It is most convenient to work in the reduced dual form of the problem as it enforces the probability equality constraint automatically, allows for making the full-support assumption, and does not require any projection routines. The computations involve sparse matrices, so appropriate sparse data structures should be used. The dual variables have a non-negative constraint, which is also sometimes referred to as a box or bound constraints in the literature.

For gradient ascent, initialize $\alpha^0 = 0$, $\beta^0 = 0$, and update the variables according to their gradient, where $\text{NN}(\sigma) = \max(0, \sigma)$, ensures the variables remain non-negative.

$$\begin{aligned}\alpha^{t+1} &\leftarrow \text{NN}\left[\alpha^t - \gamma(ACA^T\alpha^t + Ab + \epsilon + AC\beta^t)\right] \\ \beta^{t+1} &\leftarrow \text{NN}\left[\beta^t - \gamma(C\beta^t + b + C^T A^T \alpha^t)\right]\end{aligned} \tag{29}$$

If we assume the solution is full-support, we can simplify the dual version even further by dropping the β variable updates.

$$\alpha^{t+1} \leftarrow \text{NN}\left[\alpha^t - \gamma(ACA^T\alpha^t + Ab + \epsilon)\right] \tag{30}$$

Table 1. Family of MG(C)CE solutions.

MG(C)CE	ϵ	Properties
$\max(Ab)\epsilon$-MG(C)CE	$\max(Ab)$	Uniform, highest entropy, lowest payoff
$\frac{1}{2}\max(Ab)\epsilon$-MG(C)CE	$\frac{1}{2}\max(Ab)$	Between uniform and (C)CE
fullϵ-MG(C)CE	$\leq \max(Ab)$	Minimum ϵ such that MG(C)CE is full-support
MG(C)CE	0	Weak (C)CE, NE in two-player constant sum
$\min \epsilon$-MG(C)CE	≤ 0	Strictest (C)CE, lowest entropy, highest payoff

Second order derivatives are also easily computed, allowing use of bounded second order linesearch optimizers, such as L-BFGS-B (Byrd et al., 1995). Other techniques such as momentum (Rumelhart et al., 1986), preconditioning the rows of the A matrix, and iterated elimination of strictly dominated strategies of the payoff matrix will also help. An efficient conjugate gradient method can be adapted from Polyak's algorithm (Polyak, 1969; O'Leary, 1980), which is a conjugate gradient method modified to support solving problems with bounds and is proven to converge in finite iterations.

D.2. Payoff Reductions

There are two methods which could be used to reduce the size of the payoff tensor and hence reduce the complexity of the game that is required to be solved; repeated action elimination, and dominated action elimination.

Repeated Action Elimination: Consider a payoff which has repeated strategies (identical payoffs). This represents a redundancy in the game formulation and we can therefore keep only one of these actions and appropriately modify the objective to account for this alteration. Let r_p be the number of repeats for each action after elimination (i.e. $r_p = e$ if all were unique). Define $r = \otimes_p r_p$ as the flattened repeat count which is the same size as σ and $\tilde{r}_p = \otimes_{p'} \{e \text{ if } p' = p \text{ else } r_{p'}\}$. Then the constraints now become $r^T \sigma = 0$ and $A_p(\sigma \cdot \tilde{r}_p) \leq \epsilon_p$, and the objective becomes $1 - \sigma^T(\sigma \cdot r)$. This has the dual effect of reducing the number of variables and constraints in the problem and, more importantly, breaks the symmetry of repeated terms which several solvers can struggle with. It is important to run this procedure before eliminated dominated actions, because repeated actions by definition do not dominate one another.

Dominated Action Elimination: Strictly dominated strategies can be pruned from the payoff without affecting the results because dominated strategies can never have non-zero support in CEs where $\epsilon \leq 0$. Any CE solution with non-positive ϵ can exploit this reduction.

The nature of JPSRO means that it is common for actions to be repeated (best responders can produce the same output over multiple distributions) and actions to be strictly dominated by others (as the algorithm finds better and better policies).

D.3. Eigenvalue Normalization

Some methods, such as gradient methods, benefit from the eigenvalues of the problem being similar in magnitude. We found empirically that re-normalizing by the L_2 norm of the rows of the constraint matrix resulted in eigenvalues close to 1. By Theorem 5 this is a legal procedure.

D.4. Dual Optimal Learning Rate

For the dual form of the objective there is an optimal constant learning rate we can use which is based on the eigenvalues of the Hessian. Calculating the eigenvalues exactly may be too computationally expensive. We can instead obtain an upper bound. A good choice of learning rate that is guaranteed to converge is $\gamma = \frac{2}{\sigma_{\text{max}} + \sigma^{+}_{\text{min}}} \geq \frac{2}{\max_j \sum_i |D_{ij}| + \min_j \sum_i |D_{ij}|}$, where D is the Hessian of the dual form. A proof follows below.

Proof. C is idempotent and positive semi-definite. For any B, BB^T is positive semi-definite, therefore $(AC)(AC)^T = ACA^T$ is positive semi-definite. This is the first part of the block diagonals of the Hessian, D, which is therefore singular symmetric positive semi-definite.

It is known that the best choice of constant learning rate in this setting is $\gamma = \frac{2}{\sigma_{\text{max+}} + \sigma_{\text{min+}}}$. Because the Hessian is not full rank and positive semi-definite, $\sigma_{\text{min}} = 0$. We need to find the smallest non-zero eigenvalue. One possible upper bound on the maximal eigenvalues of a positive semi-definitive matrix, by the Gerschgorin circle Theorem (Gerschgorin, 1931), is:

$$\sigma_{max} \leq \max_j \sum_i |D_{ij}| = \max_i \sum_j |D_{ij}| \tag{31}$$

$$\sigma^{+}_{min} \leq \min_j \sum_i |D_{ij}| = \min_i \sum_j |D_{ij}| \tag{32}$$

□

D.5. $\min \epsilon$-MG(C)CE

The previous formulations discussed assume that ϵ is given as a hyper-parameter. If we want to directly find the minimum ϵ that produces a valid maximum Gini impurity we must also optimize over ϵ. The insight here is that the derivatives of the objective function with respect to the approximation parameter must always be stronger than the derivatives of the objective function with respect to the distribution.

$$\frac{\partial L}{\partial \epsilon} \geq e^T \frac{\partial L}{\partial \sigma} = -e^T \sigma = -1 \tag{33}$$

Therefore an additional objective with a term of -2ϵ would be sufficient to ensure this condition holds.

$$\begin{aligned} L^{\sigma}_{\alpha,\beta,\lambda,\epsilon} &= -\frac{1}{2}\sigma^T\sigma - 2\epsilon - \alpha^T(A\sigma - \epsilon) \\ &\quad - \beta^T\sigma + \lambda(e^T\sigma - 1) \end{aligned} \tag{34}$$

$$\begin{aligned} L_{\alpha,\beta,\epsilon} &= -2\epsilon - \frac{1}{2}\alpha^T ACA^T\alpha + b^T A^T \alpha + \epsilon^T \alpha \\ &\quad - \frac{1}{2}\beta^T C\beta - b^T\beta - \alpha^T AC\beta + \frac{1}{2}b^T b \end{aligned} \tag{35}$$

$$\sigma^* = CA^T\alpha^* + C\beta^* + b \tag{36}$$

E. Joint PSRO

While the concept of JPSRO is straightforward, careful attention needs to be made around a) formulating best response operators, b) creating suitable MSs, c) defining evaluation metrics, and d) establishing convergence. We discuss these in detail in this section.

E.1. Meta Game Estimation

There are two strategies for estimating the meta-game (a normal form payoff tensor populated by the returns of all the policies); exact sampling and empirical sampling.

Exact Sampling: The exact return is computed for each player by traversing the entire game tree. This is only suitable for small games, or when using deterministic policies that cannot reach the majority of the game tree.

Empirical Sampling: For larger games, or situations where the policy cannot be easily queried (for example when using a policy that depends on internal state like an LSTM) we may have to estimate the return through sampling.

In this work we used exact sampling so we could conduct an exact study into the performance of different MSs without introducing noise form other sources. However, the authors believe this approach can be scaled with empirical sampling, as has been achieved with PSRO.

E.2. Meta-Solvers

Many of the traditional PSRO solvers are factorizable solutions. Equivalently, their joint probabilities can be marginalized without losing any information.

Uniform: This solver places equal probability mass over each policy it has found so far. PSRO using a uniform distribution is also known as Fictitious Self Play (FSP) (Heinrich et al., 2015). A key advantage of this approach is that it is not necessary to compute the metagame to obtain this distribution. It is proven to slowly converge in the two-player, constant-sum setting.

Nash Equilibrium (NE): The well known solution concept (Nash, 1951), when used in PSRO is called Double Oracle (DO) (McMahan et al., 2003). This is difficult to compute for n-player, general-sum, and is equivalent to CE in two-player, constant-sum so we did not benchmark against this MS.

Projected Replicator Dynamics (PRD): An evolutionary method of approximating NE, introduced in (Lanctot et al., 2017).

There are a number of solvers which produce full joint distributions. We describe some we think are relevant here. Note that all factorizable solutions mentioned previously can be trivially promoted to full distributions.

α-Rank: A solution concept based on the stationary distribution of a Markov chain (Omidshafiei et al., 2019). α-Rank has been studied before in the context of PSRO (Muller et al., 2020), however the authors marginalized over the distribution.

Maximum Welfare (C)CE (MW(C)CE): A non-unique linear formulation that maximizes the sum of payoffs over all players. In the case where there are multiple (C)CEs with maximum welfare we can define a maximum entropy version to spread weight, MEMW(C)CE, and a random version to select one at random, RMW(C)CE. We use the latter as a MS baseline in experiments.

Random Vertex (C)CE (RV(C)CE): A linear formulation. In our implementation we formulate the standard linear (C)CE problem and randomly sample a linear cost function from the unit ball. Note that this selects a random vertex on the (C)CE polytope and is not sampling from within the polytope volume or elsewhere on the polytope surface.

Maximum Entropy (C)CE (ME(C)CE): A unique nonlinear convex formulation that maximizes the Shannon Entropy of the resulting distribution (Ortiz et al., 2007).

We do not evaluate this solution concept in this work due to computational difficulties when scaling to large payoff tensors, however we expects its performance to be similar to MG(C)CE.

Maximum Gini (C)CE (MGCE): A unique quadratic convex formulation that maximizes the Gini Impurity (a form of Tsallis Entropy), introduced in this work.

Random Dirichlet: Sample a distribution randomly from a Dirichlet distribution with $\alpha = 1$. This has not been used in the literature before but we believe acts as a good (naive) baseline against RVCE.

Random Joint: Sample a single joint policy from the set. This has not been used in the literature before either but we believe acts as a good (naive) baseline against RV(C)CE.

In previous work joint solvers have been used (Muller et al., 2020), however the authors marginalized the distributions so they could be used in classic PSRO.

E.3. Joint Best Responders

We provide two best response operators for JPSRO. The first is required to converge to a CCE in policy space (when using CCE meta-solvers). The second is required to converge to a CE in policy space (when using CE meta-solvers).

JPSRO(CCE) : At each iteration there is a single BR objective for each player, which expands the player policy set, $\Pi_p^{0:t+1} = \Pi_p^{0:t} \cup \Pi_p^{t+1}$, where $\Pi_p^{t+1} = \{\text{BR}_p^{t+1}\}$, and $\sigma(\pi_{-p}) = \sum_{\pi_p \in \Pi_p^{0:t}} \sigma(\pi_p, \pi_{-p})$.

$$\text{BR}_p^{t+1} \in \underset{\pi_p^* \in \Pi_p^*}{\text{argmax}} \sum_{\pi_{-p} \in \Pi_{-p}^{0:t}} \sigma^t(\pi_{-p}) G_p^*(\pi_p^*, \pi_{-p})$$

Therefore, the CCE BR attempts to exploit the joint distribution with the responder's own policy preferences marginalized out, resulting in a joint policy distribution over the *other* players' policies. This means that a player is best responding to a weighted mixture of up to $\otimes_{-p}|\Pi_p^t|$ joint opponent policies. This is an upper bound because σ is often sparse.

JPSRO(CE): There is a BR for each possible recommendation a player can get, $\Pi_p^{t+1} = \Pi_p^{0:t} \cup \Pi_p^{t+1}$, where $\Pi_p^{t+1} = \{(\text{BR}_p^{t+1}(\pi_p^i))_{i=1..|\Pi_p^{0:t}|}\}$.

$$\text{BR}_p^{t+1}(\pi_p) \in \underset{\pi_p^* \in \Pi_p^*}{\text{argmax}} \sum_{\pi_{-p} \in \Pi_{-p}^{0:t}} \sigma^t(\pi_{-p}|\pi_p) G_p^*(\pi_p^*, \pi_{-p})$$

Therefore the CE BR attempts to exploit each policy conditional "slice". In practice, we only calculate a BR for positive support policies (similar to Rectified Nash (Balduzzi et al., 2019). Computing the argmax of the BRs can be achieved through RL or exactly traversing the game tree. Similarly each BR is responding to a weighted mixture of up to $\otimes_{-p}|\Pi_p^t|$ joint opponent policies.

Notice that if the distribution is factorizable (like NE), then the CE BR is equal for all player policies, and furthermore is equal to the CCE BR, illuminating the connection to PSRO's BR operator.

The best response is independent of the best responding player's policy. We can compute the argmax in a number of ways. Two common ways are exact best response, and reinforcement learning.

Exact Best Response: Maintain exact tabular policies and compute a best response against the joint policies for each player, through maximizing value by traversing the game tree. We employ this approach in this work to allow us to compare meta-solvers without introducing noise from approximate BRs. This method is only suitable for small games, or when using only deterministic policies.

RL: In this setting, the learning algorithms train against randomly sampled joint-policies according to σ, and do standard value maximization. Both on-policy (such as Policy Gradient) and off-policy (such as Q-Learning) are suitable learning algorithms. Function approximation may also be used. This approach has been used extensively in PSRO before.

E.4. Evaluation Metrics

For two-player, constant-sum games there is a clear evaluation metric; how close the players are to the unique Nash Equilibrium (measured by NEGap defined below). However, outside of this narrow setting it is unclear how to fairly evaluate the policies that have been found. This is true for a number of reasons including: there being multiple equilibria, and equilibria not necessarily having good payoff. A combination of high payoff and stability is indicative of a strong set of policies. In this section we describe a number of metrics that could help describe the strength of the resulting joint policies.

Value: This describes the undiscounted return for each player at the root state of a game when following a joint policy, mixed under a joint distribution.

$$\begin{aligned} V_p(\sigma) &= \sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi) = \mathop{\mathbb{E}}_{\pi \sim \sigma} \big[G_p(\pi) \big] \\ V_p(\sigma(\cdot|\pi_p)) &= \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_{-p}|\pi_p) G_p(\pi_p, \pi_{-p}) \\ &= \mathop{\mathbb{E}}_{\substack{\pi_{-p} \sim \\ \sigma(\cdot|\pi_p)}} \big[G_p(\pi_p, \pi_{-p}) \big] \end{aligned}$$

NE Gap: This quantity describes how close joint policies are to an NE (referred to as NashConv in (Lanctot et al., 2017)) under σ. This is only defined for marginal distributions over policies.

$$\begin{aligned} \text{NEGap}_p(\sigma) &= \sum_{\pi \in \Pi} \sigma(\pi) G_p(\text{BR}_p, \pi_{-p}) - V_p(\sigma) \\ &= \mathop{\mathbb{E}}_{\pi \sim \sigma} \big[G_p(\text{BR}_p, \pi_{-p}) \big] - V_p(\sigma) \\ \text{NEGap}(\sigma) &= \sum_p \text{NEGap}_p(\sigma) \end{aligned} \tag{37}$$

CCE Gap: This quantity describes how close joint policies are to a coarse correlated equilibrium (CCE) under σ. The origins of this metric can be deduced from studying the CCE BR operator.

$$\begin{aligned} \text{CCEGap}_p(\sigma) &= \left\lfloor \sum_{\pi \in \Pi} \sigma(\pi) G_p(\text{BR}_p, \pi_{-p}) - V_p(\sigma) \right\rfloor_+ \\ &= \left\lfloor \mathop{\mathbb{E}}_{\pi \sim \sigma} \big[G_p(\text{BR}_p, \pi_{-p}) \big] - V_p(\sigma) \right\rfloor_+ \\ \text{CCEGap}(\sigma) &= \sum_p \text{CCEGap}_p(\sigma) \end{aligned}$$

Where $\lfloor x \rfloor_+ = max(0, x)$, is the non-negative operator. Note that it is possible for a best response over all joint strategies to have lower value than playing according to the joint distribution for a given player (because a BR is blind to the best responding player's correlation with the opponent policies, and deviating from this correlation can hurt performance).

CE Gap: This quantity describes how close joint policies are to a correlated equilibrium (CE) under σ.

$$\begin{aligned} &\text{CEGap}_p(\sigma, \pi_p) \\ &= \left\lfloor \sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_{-p}|\pi_p) G_p(\text{BR}_p(\pi_p), \pi_{-p}) - V_p(\sigma(\cdot|\pi_p)) \right\rfloor_+ \\ &= \left\lfloor \mathop{\mathbb{E}}_{\substack{\pi_{-p} \sim \\ \sigma(\cdot|\pi_p)}} \big[G_p(\text{BR}_p(\pi_p), \pi_{-p}) \big] - V_p(\sigma(\cdot|\pi_p)) \right\rfloor_+ \\ &\text{CEGap}_p(\sigma) = \sum_{\pi_p \in \Pi_p} \sigma(\pi_p) \text{CEGap}_p(\sigma, \pi_p) \\ &\text{CEGap}(\sigma) = \sum_p \text{CEGap}_p(\sigma) \end{aligned}$$

Unique Policy: Each iteration of JPSRO(CCE) produces n new policies (one for each player), and JPSRO(CE) produces up to the number of policies found so far. These are best responses to the joint mixture of existing polices, however, they are not guaranteed to be distinct from previous policies that have been found. The number of unique policies found so far could be a good indicator of how efficiently a meta-solver is producing new policies.

E.5. Proof of JPSRO Convergence

We provide two convergence proofs for JPSRO. Firstly, when using CCE meta-solvers with a CCE best response operator, which we refer to as JPSRO(CCE), and secondly when using CE meta-solvers with a CE best response operator, which we refer to as JPSRO(CE). Note that, in order to ignore possibly undefined values of $\sigma_t(\pi_{-p}|\pi_p)$, we use the formulation of correlated equilibria using joint probabilities instead of conditional ones. The definitions being equivalent, the conclusions are as well. Note that we also assume that $\forall p, t, |\text{BR}_p^t| > 0, \forall \pi_p$ st. $\sigma_t(\pi_p) > 0, |\text{BR}_p^t(\pi_p)| > 0$, i.e. every time a best response should be computed, it is. We discuss a relaxation of these conditions, and why it is useful, in Section E.5.3.

E.5.1. Proof of JPSRO(CCE)

Theorem 6 (CCE Convergence)**.** *When using a CCE meta-solver and CCE best response in JPSRO(CCE) the mixed joint policy converges to a CCE under the meta-solver distribution.*

We recall the definition of coarse correlated equilibria. For joint probability σ, joint policy set $\Pi = \otimes_p \Pi_p$ where Π_p is the set of valid policies of player p and $\otimes$ is the Cartesian product, and payoff function G, such that $G_p(\sigma)$ is the payoff of player p when all player play according to σ, a Coarse Correlated Equilibrium is a joint distribution σ over

Π such that, for any player p and any policy π'_p of player p,

$$\sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi) \tag{38}$$

In other words, a CCE is a distribution from which no player has an incentive to unilaterally deviate *before* being assigned their action. From this definition of CCEs, we derive the definition of CCEGap, which measures the above gap over all players

$$\text{CCEGap}(\sigma) = \sum_p \left\lfloor \max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) \right\rfloor_+$$

where $\lfloor x \rfloor_+ = max(0, x)$, this $\lfloor \rfloor_+$ term being necessary because the gap is potentially negative, as one can see from Equation 38. From this definition, we introduce the following lemma:

Lemma 1 (Game CCE and CCEGap)**.** *We have the following equivalence:*

(i) σ is a CCE of the game

(ii) CCEGap(σ) = 0

Proof. Let us first prove (i) → (ii). Suppose σ is a CCE. Then for any player p and any policy π'_p of player p,

$$\sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi)$$

therefore, by subtracting the right hand-term and taking the maximum over $\pi'_p \in \Pi_p$,

$$\max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) \leq 0$$

and so

$$\left\lfloor \max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(G_p(\pi'_p, \pi_{-p} - G_p(\pi)) \right\rfloor_+ = 0$$

Summing this last inequality over all players yields (ii).

Let us now prove (ii) → (i). Suppose that σ is such that CCEGap(σ) = 0. Then, for all p,

$$\max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) \leq 0 \tag{39}$$

For all $\pi''_p \in \Pi_p$ we have

$$\sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi''_p, \pi_{-p}) \leq \max_{\pi'_p} \sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi'_p, \pi_{-p})$$

and therefore, by subtracting $\sum_{\pi \in \Pi} \sigma(\pi) G_p(\pi)$ and using Equation 39,

$$\sum_{\pi \in \Pi} \sigma(\pi)(G_p(\pi''_p, \pi_{-p}) - G_p(\pi)) \leq 0$$

Rearranging the terms yields the proof. □

The context of JPSRO motivates us to expand and overload the definition CCEGap. Let us denote by Π^* the policies of the extensive form game, and by $\Pi^{0:t}$ all the policies found by JPSRO by iteration t. We immediately have, for all t, $\Pi^{0:t} \subset \Pi^*$. We expand CCEGap via, for all t,

$$\text{CCEGap}(\sigma, \Pi^*, \Pi^{0:t}) = \sum_p \left\lfloor \max_{\pi^*_p \in \Pi^*_p} \sum_{\pi \in \Pi^{0:t}} \sigma(\pi)(G_p(\pi^*_p, \pi_{-p}) - G_p(\pi)) \right\rfloor_+$$

The only difference is the search space of π^*_p, which now lives within Π^*, while the policies used in the sum live in $\Pi^{0:t}$. It is nevertheless easy to see that this new definition characterizes CCEs of Π^* (and not of $\Pi^{0:t}$), albeit a restricted class, since $\Pi^{0:t} \subset \Pi^*$ and one can expand σ to be zero over $\Pi^* \setminus \Pi^{0:t}$. Let us now prove Theorem 6.

Proof. To prove that JPSRO with a CCE meta-solver, JPSRO(CCE), converges to a CCE, we need only prove one thing: that JPSRO(CCE) is unable to produce new policies if and only if it has reached a CCE of the extensive form game. Provided this is true, and since all games have a finite number of deterministic policies, we have that JPSRO(CCE) necessarily cannot produce new policies forever, and therefore eventually can only produce already-discovered policies.

Note that the joint distribution σ_t of JPSRO(CCE) is by construction a CCE over $\Pi^{0:t}$ for all t (when using a CCE meta-solver). It is nevertheless not necessarily a CCE of Π^*.

Let us now suppose that JPSRO(CCE) has not produced any new policy for any player at iteration t. Given the JPSRO(CCE) formulation, we can therefore restrict the search space of policies from Π^* to $\Pi^{0:t}$ in the CCEGap max term, since the max of the expression is reached in $\Pi^{0:t}$, and we thus rewrite the CCEGap definition:

$$\sum_p \left\lfloor \max_{\pi'_p \in \Pi^*_p} \sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) \right\rfloor_+$$
$$= \sum_p \left\lfloor \max_{\pi'_p \in \Pi^{0:t}_p} \sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) \right\rfloor_+$$

But since σ_t is a CCE over $\Pi^{0:t}$, the second term is null. Therefore, CCEGap$(\sigma, \Pi^*, \Pi^{0:t}) = 0$, and according to Lemma 1, σ_t is therefore a CCE over Π^*, which concludes the proof. □

E.5.2. Proof of JPSRO(CE)

Theorem 7 (CE Convergence)**.** *When using a CE meta-solver and CE best response in JPSRO(CE) the mixed joint policy converges to a CE under the meta-solver distribution.*

We recall the definition of correlated equilibria. Keeping the same notations as above, a correlated equilibrium is a joint distribution σ over Π such that, for any player p and any policies π_p, π'_p of player p,

$$\sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi_p, \pi_{-p})$$

In other words, a CE is a distribution from which no player has an incentive to unilaterally deviate even *after* having been assigned their action. They are therefore stronger than CCEs, and the result CEs $\subseteq$ CCEs easily follows from the above inequality. From this definition of CEs, we derive the definition of CEGap, which measures the above gap over all players.

$$\text{CEGap}(\sigma) = \sum_{p, \pi_p \in \Pi_p} \Big\lfloor \max_{\pi'_p} \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \Big\rfloor_+$$

From this definition, we conclude the following lemma:

Lemma 2 (Game CE and CEGap)**.** *We have the following equivalence:*

(i) σ is a CE of the game

(ii) CEGap(σ) = 0

Proof. Let us first prove (i) → (ii). Let σ be a CE of the game. Therefore, for all p, for all $\pi_p, \pi'_p \in \Pi_p$,

$$\sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi_p, \pi_{-p})$$

therefore

$$\sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \leq 0$$

which is true for all $\pi'_p \in \Pi_p$, so also true for the max over them

$$\max_{\pi'_p \in \Pi_p} \sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \leq 0$$

$$\Big\lfloor \max_{\pi'_p \in \Pi_{-p}} \sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \Big\rfloor_+ = 0$$

Therefore (i) → (ii).

Let us now suppose that σ is such that $\text{CEGap}(\sigma) = 0$. Thus

$$\sum_{p, \pi_p \in \Pi_p^{0:t+}} \Big\lfloor \max_{\pi'_p} \sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_p, \pi_{-p}) (G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \Big\rfloor_+ = 0$$

Given the presence of the positivity operator $\lfloor . \rfloor_+$, we deduce that for all p, for all $\pi_p, \pi'_p \in \Pi_p^{0:t}$,

$$\sum_{\substack{\pi_{-p} \in \\ \Pi_{-p}}} \sigma(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \leq 0$$

We therefore deduce

$$\sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi'_p, \pi_{-p}) \leq \sum_{\pi_{-p} \in \Pi_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi_p, \pi_{-p})$$

which concludes the proof. □

Once again, the CEGap definition is extended

$$\text{CEGap}(\sigma, \Pi^*, \Pi^{0:t}) = \sum_{p, \pi_p \in \Pi_p^{0:t}} \Big\lfloor \max_{\pi^*_p \in \Pi^*_p} \sum_{\pi_{-p} \in \Pi^t_{-p}} \sigma(\pi_p, \pi_{-p})(G_p(\pi^*_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \Big\rfloor_+$$

It is once again easy to see that $\text{CEGap}(\sigma, \Pi^*, \Pi^{0:t})$ characterizes CEs of Π^*.

This lemma proven, we prove Theorem 7.

Proof. Once again, it is sufficient to prove that JPSRO(CE) stops producing new policies if and only if it has reached a CE of the extensive form game, the rest of the argument being supplied by the finiteness of the game forcing JPSRO(CE) to eventually stop producing new policies.

Let us now suppose that JPSRO(CE) has not produced any new policy for any new player at iteration t. This means that for all $\pi_p \in \Pi_p^t$,

$$\max_{\pi^*_p \in \Pi^*_p} \sum_{\substack{\pi_{-p} \in \\ \Pi^{0:t}_{-p}}} \sigma(\pi_p, \pi_{-p}) G_p(\pi^*_p, \pi_{-p}) = \max_{\pi'_p \in \Pi^t_p} \sum_{\substack{\pi_{-p} \in \\ \Pi^{0:t}_{-p}}} \sigma(\pi_p, \pi_{-p}) G_p(\pi'_p, \pi_{-p})$$

We subtract $\sum_{\pi_{-p}\in\Pi^t_{-p}} \sigma(\pi_p, \pi_{-p}) G_p(\pi_p, \pi_{-p})$ to both expressions, apply $\lfloor . \rfloor_+$ and sum over $\pi_p \in \Pi^t_p$ and p, and finally apply the fact that σ is a CE of the restricted game to obtain that

$$\text{CEGap}(\sigma, \Pi^*, \Pi^{0:t}) = \sum_{p, \pi_p \in \Pi_p} \Bigg\lfloor \max_{\pi'_p \in \Pi^t_p} \sum_{\pi_{-p}\in\Pi_{-p}} \sigma(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \Bigg\rfloor_+ = 0$$

which, by extension, is also true for the CEGap over the extensive form game. By Lemma 2, σ is therefore a CE of the extensive form game, which concludes the proof. □

E.5.3. Relaxation on Proof Requirements

Our definition of Best Responses (BRs) is that they are functions that return a set of policies which maximize their value against a given objective. There are two reasons to add a set of policies. Firstly, the max of a given objective can be reached at different points, thus returning a set of policies enables us to potentially include them all. Secondly, using sets also enables us to potentially set some of the BR outputs to $\emptyset$. Concretely, this means that no policy is computed by the BR in that case, which saves compute time and memory. The proofs shown so far rely on each BR having cardinality greater than or equal to 1, which means that one should compute at least one new policy every time the BR operator is called. We can relax this condition into the following conditions, which we prove are sufficient (but not necessary) for convergence.

CCE-Condition:

$$\forall T > 0, p, \exists t > T, |\text{BR}^t_p| \geq 1$$

i.e. each player receives an infinity of best responses.

CE-Condition:

$$\forall T > 0, p, \pi_p, \exists t > T, \text{ either } \forall t' \geq t, \sigma_{t'}(\pi_p) = 0 \text{ or } |\text{BR}^t_p(\pi_p)| \geq 1$$

i.e. any policy of any player is either never selected by the CE meta-solver after some time, or is considered for a best response an infinite number of times.

Solver-Condition: $\forall t, \forall t' \geq t$, if $\forall p, \forall \pi_p \in \Pi^{0:t'}_p, \pi_p \in \Pi^{0:t}_p$, then $\forall \pi \in \Pi^{0:t}$ (or $\pi \in \Pi^{t'}$), $\sigma_t(\pi) = \sigma_{t'}(\pi)$: if no new policy has been added to the pool between t and t', the amount of mass granted to each policy by the solver does not change, i.e. repeating policies does not affect solver outputs, and the solver's outputs are constant given the same pools.

The rest of this section presents the relaxed theorems, their proofs, and discusses why such a relaxation is of interest.

Relaxed Theorems and Proofs

Theorem 8 (Relaxed CCE-Convergence)**.** *When using a CCE meta-solver and CCE best response in JPSRO(CCE), under CCE-Condition and Solver-Condition, the mixed joint policy converges to a CCE under the meta-solver distribution.*

Proof. Let us suppose CCE-Condition and Solver-Condition. We have that JPSRO(CCE) will necessarily be able to produce new policies until it reaches a CCE. Let us prove this: while $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$, JPSRO(CCE) is able to add at least one new policy to its pool. Indeed, let $t > 0$ be such that $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$. Then there exists at least one p such that

$$\max_{\pi'_p \in \Pi^*_p} \sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) > 0.$$

Let us select one of these p with minimal $t' \geq t$, $|\text{BR}^t_p| \geq 1$, i.e. the first best response with positive CCEGap to be added to the pool after and including t. t' exists because we suppose CCE-Condition. Let us suppose that no new policies have been added to the pool between t and t'. Then, since no new best response has been added to the pool between t and t', $\sigma_t = \sigma_{t'}$ since we suppose Solver-Condition, and therefore $\forall \pi' \in \text{BR}^{t'}_p$,

$$\sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) > 0.$$

We have that necessarily, $\text{BR}^{t'}_p \cap \Pi^{0:t}_p = \emptyset$, as otherwise σ_t would not be a CCE of $\Pi^{0:t}$: indeed, since σ_t is a CCE of $\Pi^{0:t}$, $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) = 0$, and thus $\forall p, \pi'_p \in \Pi^{0:t}_p$,

$$\sum_{\pi \in \Pi^{0:t}} \sigma_t(\pi)(G_p(\pi'_p, \pi_{-p}) - G_p(\pi)) \leq 0,$$

thus new best responses can be added to the pool. We therefore have that $\text{CCEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$ implies that at least one new policy can be found by JPSRO.

Thus a new best response can always be added, and will always be added since we have CCE-Condition, to the pool while σ_t is not a CCE of the extensive form game. Therefore, if JPSRO(CCE) is unable to add any new policy to the pool (which has to be verified over all players, or measured through CCEGap), then it must be at a CCE, which concludes the proof. □

Theorem 9 (Relaxed CE-Convergence)**.** *When using a CE meta-solver and CE best response in JPSRO(CE), under CE-Condition and Solver-Condition, the mixed joint policy converges to a CE under the meta-solver distribution.*

Proof. Let us suppose CE-Condition and Solver-Condition. We have that JPSRO(CE) will necessarily be able to produce new policies until it reaches a CE. Let us prove this: while $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$, JPSRO(CE) is able to add at least one new policy to its pool. Indeed, let $t > 0$ be such that $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$. Then there exists at least one p, π_p st. $\sigma_t(\pi_p) > 0$ such that

$$\max_{\pi'_p \in \Pi^*_p} \sum_{\pi_{-p} \in \Pi^t_{-p}} \sigma_t(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) > 0.$$

By CE-Condition, we have that either new policies have been added to the pool before any such p, π_p has been selected, or that there exists t' such that $t' \geq t$, $|\text{BR}^t_p(\pi_p)| \geq 1$. Indeed, if no new best response has been added to the pool by $t' \geq t$, the Solver-Condition implies that for all these p, π_p st. $\sigma_t(\pi_p) > 0$, we also have $\sigma_{t'}(\pi_p) > 0$, hence there exists t', $|\text{BR}^t_p(\pi_p)| > 1$. Let us select the minimal t' over all p, π_p such that $\text{CEGap}_p(\sigma_t, \Pi^*, \Pi^{0:t})(\pi_p) > 0$.

Let us suppose that no new policies have been added to the pool between t and t'. Then, since no new best response has been added to the pool between t and t', $\sigma_t = \sigma_{t'}$ since we suppose Solver-Condition, and therefore $\forall \pi' \in \text{BR}^{t'}_p(\pi_p), \sum_{\pi_{-p} \in \Pi^t_{-p}} \sigma_t(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) > 0$. We have that necessarily, $\text{BR}^{t'}_p(\pi_p) \cap \Pi^{0:t}_p = \emptyset$, as otherwise σ_t would not be a CE of $\Pi^{0:t}$: indeed, since σ_t is a CE of $\Pi^{0:t}$, $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) = 0$, and thus $\forall p, \pi_p \in \Pi^{0:t}_p, \pi'_p \in \Pi^{0:t}_p$,

$$\sum_{\pi_{-p} \in \Pi^{0:t}_{-p}} \sigma_t(\pi_p, \pi_{-p})(G_p(\pi'_p, \pi_{-p}) - G_p(\pi_p, \pi_{-p})) \leq 0.$$

Thus new best responses can be added to the pool. We therefore have that $\text{CEGap}(\sigma_t, \Pi^*, \Pi^{0:t}) > 0$ implies that at least one new policy can be found by JPSRO.

Thus a new best response can always be added, and will always be added since we have CE-Condition, to the pool while σ_t is not a CE of the extensive form game. Therefore, if JPSRO(CE) is unable to add any new policy to the pool (Which has to be verified over all players, or measured through CEGap), then it must be at a CE, which concludes the proof. □

Discussion on Relaxation

These relaxed conditions matter especially for JPSRO(CE), which has potentially exponential complexity in term of number of policies to keep (if the solver spreads mass on all policies at each iteration, then the number of policies in each players' pools at iteration t is $\geq 1 + \sum_{k=1}^{t} 2^k = 2^{t+1} - 1$).

Given that the policies produced for one player at the same iteration are potentially similar (even identical), a number of modifications could be imagined to keep JPSRO(CE) tractable. For example: a) randomly select only one π_p from which to best respond for each player, b) only compute a best response for one randomly chosen π_p, or c) compute all BRs, but only add the BR with the largest gap to the pool.

It could make sense to randomly select only one π_p from which to best respond for each player, at each iteration, or even to only compute a best response for one randomly chosen π_p for one randomly-chosen p at each iteration.

Note that it is necessary to impose a condition on the solver (although an alternate Solver-Condition could be formulated). To illustrate this, let us imagine modes between the best response chooser and the solver. Namely, let us imagine a two-player game, for which on even t, in JPSRO(CCE), the best response operator only computes one best response for player 1 (and on odd t, the best response is computed only for player 2). Let us also infer that the current restricted game has two CCEs. The first of these (CCE1) is not "expandable" for player 1, but is for player 2 (i.e. the best response for player 1 is already in the pool, but player 2's best response is not). The second (CCE2) is expandable for player 1, but not for player 2. If the CCE solver outputs CCE1 on even t, and CCE2 on odd t, then the algorithm never produces new policies, and therefore never converges.

Of course, the conditions provided are sufficient, but not necessary, and in the case where best response and meta-solver outputs' randomizations are decorrelated, it makes intuitive sense that the algorithm should also converge with probability 1, which one can prove with a more involved argument.

F. Games

We study several games with JPSRO; Kuhn Poker, Trade Comm, and Sheriff. These cover three-player, general-sum, and common-payoff games. Implementations of all the games are available in OpenSpiel (Lanctot et al., 2019).

Kuhn Poker: A simplified n-player, zero-sum, sequential, imperfect information version of poker. It consists of $n + 1$ playing cards. In each round of the game, every player remaining *antes* one chip. One card is dealt to each player. Each player has two choices, *bet* one chip or *check*. If a player bets other players have the option to *call* or *fold*. Out of the players that bet, the one with the highest card wins. If all players check the player with the highest card wins. The original two-player game is described in (Kuhn, 1950). An n-player extension is described in (Lanctot, 2014). Additional information about the game (such as equilibrium) can be found in (Hoehn et al.).

Trade Comm: A simple two-player, common-payoff trading game (Sokota et al., 2021). In this game each player (in secret) receives one of I different items. The first player can then make one of I utterances to the second agent, and vice versa. Then each agent chooses one of I^2 trades in private, if the trade is compatible both agents receive 1 reward, otherwise both receive 0. The goal of the agents is therefore to find a bijection between the items and utterances and the trade proposal. There are I^4 deterministic policies per player, and good learning algorithms will be be able to search over these policies. Because the game is common-payoff, it is very transitive, and has many dominated strategies, however there are multiple strategies with equal payoff, and therefore many equilibria in partially explored policy space. It is for this reason many learning algorithms get stuck exploiting sub-optimal policies they have already found.

Sheriff: A simplified two-player, general-sum version of the board game Sheriff of Nottingham (Farina et al., 2019b). The game consists of a smuggler, who is motivated to import contraband without getting caught, and a sheriff, who is motivated to either find contraband or accept bribes. The players negotiate a bribe over several rounds after which the bribe if accepted or rejected. If the sheriff finds contraband, the smuggler pays a fine, otherwise if no contraband is found the sheriff must pay compensation to the smuggler. The smuggler also gets value from smuggling goods. The game has different optimal values for NFCCE, EFCCE, EFCE, and NFCE solutions concepts.

G. JPSRO Hyper-parameters

There are a number of ways of implementing JPSRO in practice through various hyper-parameters.

Best Response: We use an exact best response calculation that assigns uniform probability over valid actions for states with zero reach probability. However, other best response approaches will also work including reinforcement learning (which we will leave to future work).

Pool Type: The data structure used to store the policies found so far can either be a set or a multi-set. Using a set ensures that all policies are unique and only appear once even if multiple iterations produce the same best response policy. Some meta-solvers rely on repeated policies being present for convergence convergence (for example, the uniform meta-solver can converge in two-player, zero-sum because the repeated policies trend to a NE over repeats). In this case using a multi-set is more suitable. This parameterization is only relevant when using tabular policies which can be checked for equality.

Player Updates Per Iteration: It is not necessary to find the best response for all players at every iteration. Other strategies such as cycling through players or randomly selecting a player will work too. It is sufficient that over time all players should be updated. Updating a single player at a time is more efficient when minimizing the number of best responses necessary for convergence, however updating all can be done in parallel.

Best Responses Per Iteration: When computing the CE best response, each player has several best responses to calculate. It is not necessary to compute them all and, even if they are all computed, it is not necessary to add them all to the pool of policies. The best responses can be calculated at random. And only best responses with nonzero gap need be added, or perhaps only the one with largest gap. In order to measure convergence to a CE, all best responses (and their gaps) must be computed.

Policy Initialization: Policies can be initialized in any manner and the algorithm will converge to an equilibrium under any initial condition. However, the initial policies does determine the space of equilibrium reachable (so for example is may not be possible to find the MWCE from all initial policies). JPSRO works, without limitation, using only deterministic policies, however stochastic policies are supported too. A stochastic uniform policy over valid actions is a reasonable setting.

Best Response Type: The most important parameterization is picking one of the two best response types: CE and CCE. The resulting algorithm is named either JPSRO(CE) or JPSRO(CCE) respectively.

Meta-Solvers: The second most important parameterization is the type of meta-solver to use (Table 2). An important constraint is that JPSRO(CE) is only guaranteed to converge under CE meta-solvers. JPSRO(CCE) must use CCE meta-solvers (noting that CEs are a subset of CCEs).

H. Experiments

We conduct experiments over three extensive form games to demonstrate the versatility of the algorithm over n-player general-sum games. For each game we run on both JPSRO(CCE) and JPSO(CE) algorithms under all suitable meta-solvers and baselines.

For JPSRO(CCE), we initialize using uniform policies, update all players at every iteration, and use multi-sets for the pool. For JPSRO(CE), we initialize using uniform policies,

Table 2. Summary of meta-solvers used during experiments and their properties. We use the normalized ϵ for naming, for example $\frac{1}{100}\epsilon$-MGCE means $\frac{1}{100}\max(Ab)\epsilon$-MGCE.

Meta-Solver	Joint	CCE	CE	Max Val	Max Ent	Rand
Uniform					✓	
PRD						
α-Rank	✓					
Rand Dirichlet	✓					✓
Rand Joint	✓					✓
RMWCCE	✓	✓		✓		✓
RVCCE	✓	✓				✓
$\frac{1}{100}\epsilon$-MGCCE	✓	ϵ			✓	
MGCCE	✓	✓			✓	
min ϵ-MGCCE	✓	✓			✓	
RMWCE	✓	✓	✓	✓		✓
RVCE	✓	✓	✓			✓
$\frac{1}{100}\epsilon$-MGCE	✓	ϵ	ϵ		✓	
MGCE	✓	✓	✓		✓	
min ϵ-MGCE	✓	✓	✓		✓	

update all players at every iteration, only add the highest-gap BR to the pool for each player at each iteration, and use multi-sets for the pool. For random meta-solvers we repeat the experiment five times and show the average, otherwise the experiment is deterministic. The experiments were run for 6 hours, after which any that had not finished were truncated.

In order to measure performance, we track five metrics:

1. The gap to equilibrium under a maximum welfare equilibrium (MW(C)CE) distribution. This describes how close the algorithm is to finding a set of joint policies that are in exact equilibrium in the extensive form game.

2. The gap to equilibrium under the meta-solver's distribution. This is the gap that JPSRO theoretically converges to when using (C)CEs.

3. The value of the game to the players under the MW(C)CE distribution.

4. The value of the game to the players under the meta-solver's distribution.

5. The number of unique policies found so far.

Ultimately, the algorithm should be finding high-value joint policies that are in equilibrium, over a variety of games. The first game is a purely competitive, three-player game called Kuhn Poker (Figure 3). The second game is a purely cooperative, common-payoff game called Trade Comm (Figure 4). The final game is a general-sum game called Sheriff (Figure 5).

I. Open Source Code

An open source implementation of JPSRO is available in OpenSpiel (Lanctot et al., 2019) under https://github.com/deepmind/open_spiel/blob/master/open_spiel/python/examples/jpsro.py.

J. Necessity of Population Based Training

In the absence of a correlating signal, a single joint policy is, in general, insufficient to represent a correlated equilibrium. To see this, let us consider the Traffic Light game (Figure 1b). One possible correlated equilibrium consists in recommending (G, W) half of the time, and (W,G) the other half.

Let us now consider this game as an extensive-form, partial-information game, where the row player first chooses their action, and the column player then chooses their own without knowing the action chosen by the row player. In the absence of a correlating signal, it is impossible for the column player to know which action the row player has played, and therefore playing (G, W) or (W, G) becomes impossible, as the column player is unable to change their action as a function of the action taken by the row player.

Therefore, without modifying the game and observation space to add a correlating signal, convergence to a correlated equilibrium necessarily requires a distribution over joint policies. Population Based Training (PBT), a set of methods that slowly grow the space of (joint) policies, therefore appears to be the appropriate framework to converge to (C)CEs without adding correlating signals to the considered game.

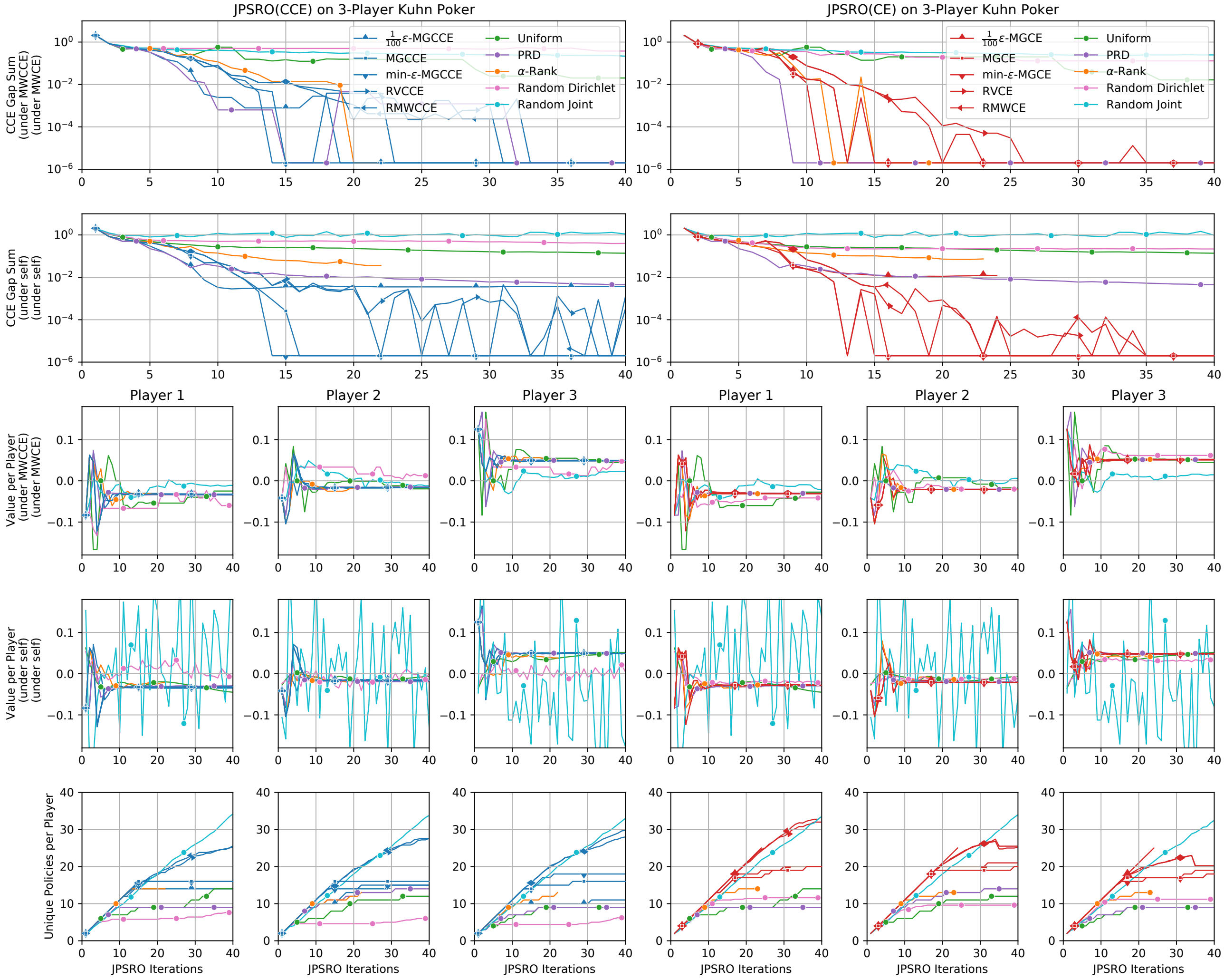

Figure 3. JPSRO(CCE) and JPSRO(CE) on three-player Kuhn Poker. All (C)CE MSs, PRD and α-Rank find joint policies capable of supporting equilibrium (although α-Rank was slow and was terminated after 6 hours). This is some evidence that classic MSs designed for the two-player, zero-sum setting can generalize well to the three-player, zero-sum.

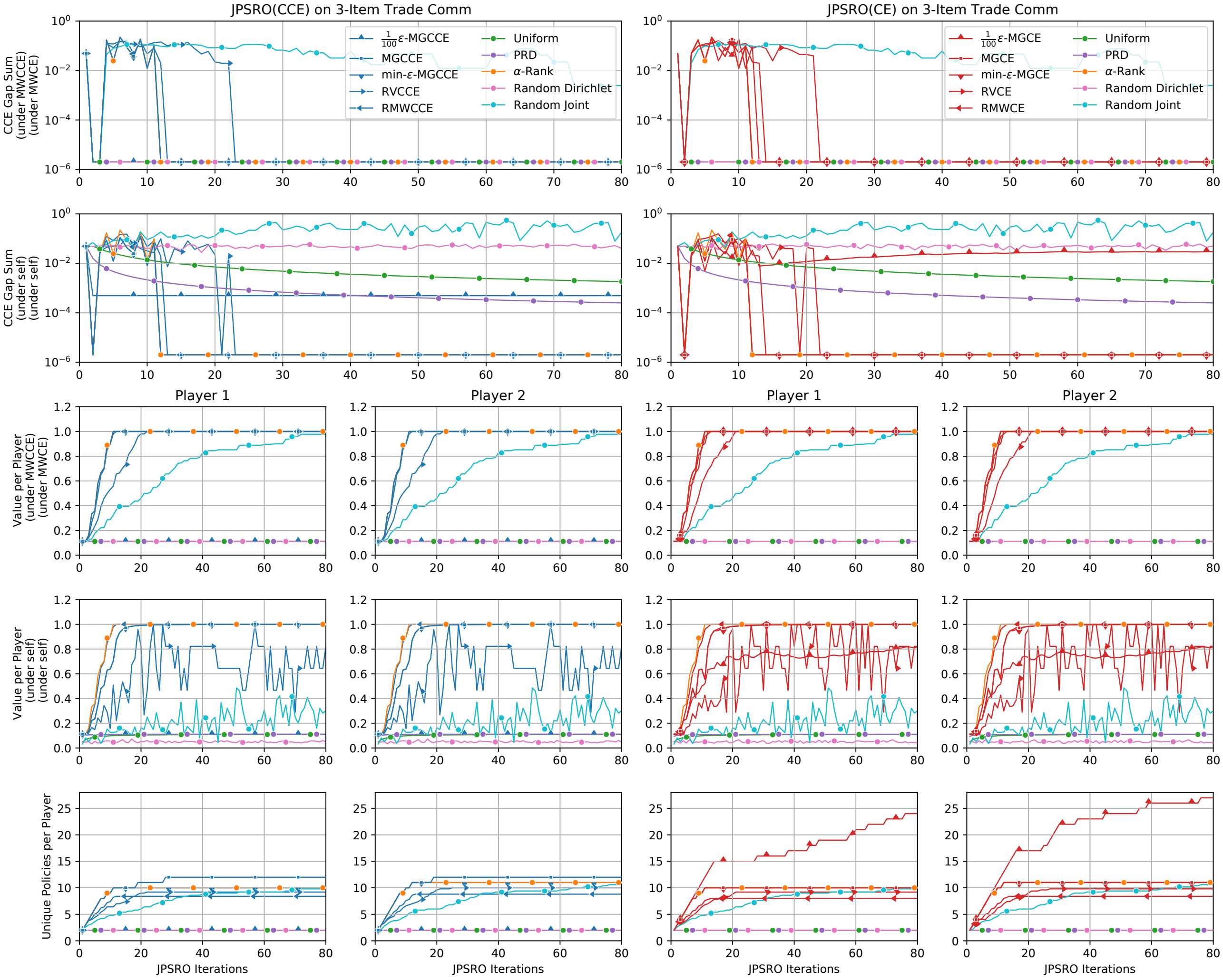

Figure 4. JPSRO(CCE) and JPSRO(CE) on three-item Trade Comm. In JPSRO(CCE), $\frac{1}{100}$ min-MGCCE fails find the maximum welfare equilibrium, however, all other (C)CE MSs find the maximum welfare equilibrium. Unexpectedly, α-Rank performs well on this game, while all other classic MSs fail to make progress on this purely cooperative game. Performing well on this game requires exploration, so the random joint MS is able to make progress, albeit naively and slowly.

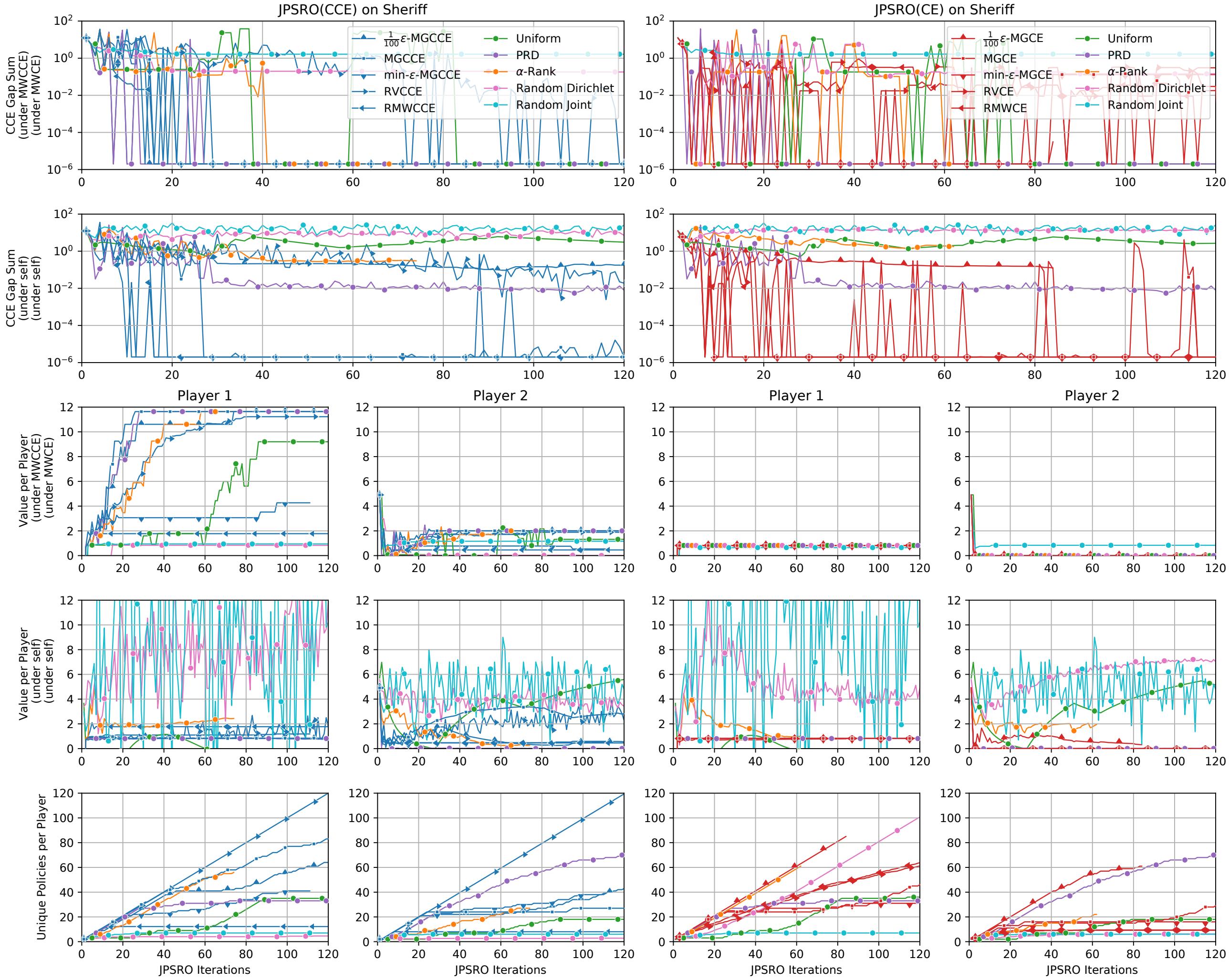

Figure 5. JPSRO(CCE) and JPSRO(CE) on Sheriff. This game is interesting because it is general-sum and different solution concepts have different optimal maximum welfare values. The maximum welfare NFCCE is 13.64 for the smuggler and 2.0 for the sheriff which JPSRO(CCE) successfully finds, while the maximum welfare NFCE is 0.82 for the smuggler and 0.0 for the sheriff which JPSRO(CE) successfully finds. This demonstrates the appeal of using NFCCE as a target equilibrium. Interestingly, for this game, $\frac{1}{100}\epsilon$-MG(C)CE was able to produce BRs of high enough quality to converge which is evidence that scaled methods that only approximate (C)CEs may be enough in some settings. RMWCCE converged to an equilibrium, but not the welfare maximizing one, providing evidence that greedy MSs are not always suitable. In a similar argument, min-ϵ-MGCCE did not reach the maximum welfare solution within the allocated number of iterations. RV(C)CE is efficient at finding novel policies but ones of limited utility. PRD and α-Rank perform well and find the maximum welfare (C)CE equilibria.