D3C: REDUCING THE PRICE OF ANARCHY IN MULTI-AGENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Even in simple multi-agent systems, fixed incentives can lead to outcomes that are poor for the group and each individual agent. We propose a method, D3C, for online adjustment of agent incentives that reduces the loss incurred at a Nash equilibrium. Agents adjust their incentives by learning to mix their incentive with that of other agents, until a compromise is reached in a distributed fashion. We show that D3C improves outcomes for each agent and the group as a whole on several social dilemmas including a traffic network with Braess's paradox, a prisoner's dilemma, and several reinforcement learning domains.

1 INTRODUCTION

We consider a setting composed of multiple interacting artificially intelligent agents. These agents will be instantiated by humans, corporations, or machines with specific individual incentives. However, it is well known that the interactions between individual agent goals can lead to inefficiencies at the group level, for example, in environments exhibiting social dilemmas (Braess, 1968; Hardin, 1968; Leibo et al., 2017). In order to resolve these inefficiencies, agents must reach a compromise.

Any arbitration mechanism that leverages a central coordinator[1] faces challenges when attempting to scale to large populations. The coordinator's task becomes intractable as it must both query preferences from a larger population and make a decision accounting for the exponential growth of agent interactions. If agents or their designers are permitted to modify their incentives over time, the principal must collect all this information again, exacerbating the computational burden. A central coordinator represents a single point of failure for the system whereas one motivation for multi-agent systems research inspired by nature (e.g., humans, ants, the body, etc.) is robustness to node failures (Edelman and Gally, 2001). Therefore, we focus on decentralized approaches.

A trivial form of decentralized compromise is to require every agent to minimize group loss (maximize welfare). Leaving the optimization problem aside, this removes inefficiency, but similar to a mechanism with a central coordinator, requires communicating all goals between all agents, an expensive step and one with real consequences for existing distributed systems like wireless sensor networks (Kulkarni et al., 2010) where transmitting a signal saps a node's energy budget. There is also the obvious issue that this compromise may not appeal to an individual agent, especially one that is expected to trade its low-loss state for a higher average group loss. One additional, more subtle consequence of optimizing group loss is that it cannot distinguish between behaviors in environments with a group loss that is constant sum, for instance, in zero-sum games. But zero-sum games have rich structure to which we would like agents to respond. Electing a team leader (or voting on a decision) implies one candidate (decision) wins while another loses. Imagine two agents differ on their binary preference with each trying to minimize their probability of losing. A group loss is indifferent; we prefer the agents play the game (and in this, case argue their points).

Design Criteria: We seek an approach to compromise in multi-agent systems that applies to the setting just described. The celebrated Myerson-Satterthwaite theorem (Arrow, 1970; Satterthwaite, 1975; Green and Laffont, 1977; Myerson and Satterthwaite, 1983) states that no mechanism exists that simultaneously achieves optimal efficiency (welfare-maximizing behavior), budget-balance (no taxing agents and burning side-payments), appeals to rational individuals (individuals want to opt-in to the mechanism), and is incentive compatible (resulting behavior is a Nash equilibrium). Given

[1] For example, the VCG mechanism (Clarke, 1971).

this impossibility result, we aim to design a mechanism that approximates weaker notions of these criteria. In addition, the mechanism should should be decentralized, extensible to large populations, and adapt to learning agents with evolving incentives in possibly non-stationary environments.

Design: We formulate compromise as agents mixing their incentives with others. In other words, an agent may become incentivized to minimize a mixture of their loss and other agents' losses. We design a decentralized meta-algorithm to search over the space of these possible mixtures.

We model the problem of efficiency using *price of anarchy*. The price of anarchy, $\rho \in [1, \infty)$, is a measure of inefficiency from algorithmic game theory with lower values indicating more efficient games (Nisan et al., 2007). Forcing agents to minimize a group (average) loss with a single local minimum results in a "game" with $\rho = 1$. Note that any optimal group loss solution is also Pareto-efficient. Computing the price of anarchy of a game is intractable in general. Instead, we derive a differentiable upper bound on the price of anarchy that agents can optimize incrementally over time. Differentiability of the bound makes it easy to pair the proposed mechanism with, for example, deep learning agents that optimize via gradient descent (Lerer and Peysakhovich, 2017; OpenAI et al., 2019). Budget balance is achieved exactly by placing constraints on the allowable mixtures of losses. We appeal to individual rationality in three ways. One, we initialize all agents to optimize only their own losses. Two, we include penalties for agents that deviate from this state and mix their losses with others. Three, we show empirically on several domains that opting into the the proposed mechanism results in better individual outcomes. We also provide specific, albeit narrow, conditions under which agents may achieve a Nash equilibrium, i.e. the mechanism is incentive compatible, and demonstrate the agents achieving a Nash equilibrium under our proposed mechanism in a traffic network problem.

The approach we propose divides the loss mixture coefficients among the agents to be learned individually; critically, the agents do not need to observe or directly differentiate with respect to the other agent strategies. In this work, we do not tackle the challenge of scaling communication of incentives to very large populations; we leave this to future work. Under our approach, scale can be achieved through randomly sharing incentives according to the learned mixture weights or sparse optimization over the simplex (Pilanci et al., 2012; Kyrillidis et al., 2013; Li et al., 2016).

Our Contribution: We propose a differentiable, local estimator of game inefficiency, as measured by price of anarchy. We then present two instantiations of a single decentralized meta-algorithm, one 1st order (gradient-feedback) and one 0th order (bandit-feedback), that reduce this inefficiency. This meta-algorithm is general and can be applied to any group of individual agent learning algorithms.

This paper focuses on how to enable a group of agents to respond to an unknown environment and minimize overall inefficiency. Agents with distinct losses may find their incentives well aligned to the given task, however, they may instead encounter a social dilemma (Sec. 3). We also show that our approach leads to interesting behavior in scenarios where agents may need to sacrifice team reward to save an individual (Sec. F.4) or need to form parties and vote on a new team direction (Sec. 3.4). Ideally, one meta-algorithm would allow a multi-agent system to perform sufficiently well in all these scenarios. The approach we propose, D3C (Sec. 2), is not that meta-algorithm, but it represents a holistic effort to combine critical ingredients that we hope takes a step in the right direction.[2]

2 Dynamically Changing the Game

In our approach, agents may consider slight re-definitions of their original losses, thereby changing the definition of the original game. Critically, this is done in a way that conserves the original sum of losses (budget-balanced) so that the original group loss can still be measured. In this section, we derive our approach to minimizing the price of anarchy in several steps. First we formulate minimizing the price of anarchy via compromise as an optimization problem. Second we specifically consider compromise as the linear mixing of agent incentives. Next, we define a *local* price of anarchy and derive an upper bound that agents can differentiate. Then, we decompose this bound into a set of differentiable objectives, one for each agent. Finally, we develop a gradient estimator to minimize the agent objectives in settings with bandit feedback (e.g., RL) that enables scalable decentralization.

[2] D3C is agnostic to any action or strategy semantics. We are interested in rich environments where high level actions with semantics such as "cooperation" and "defection" are not easily extracted or do not exist.

2.1 Notation and Transformed Losses

Let agent i's loss be $f_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X} \to \mathbb{R}$ where $\boldsymbol{x}$ is the joint strategy of all agents. Let $f_i^A(\boldsymbol{x})$ denote agent i's transformed loss which mixes losses among agents. Let $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})]^\top$ and $\boldsymbol{f}^A(\boldsymbol{x}) = [f_1^A(\boldsymbol{x}), \ldots, f_n^A(\boldsymbol{x})]^\top$ where $n \in \mathbb{Z}$ denotes the number of agents. In general, we require $f_i^A(x) > 0$[3] and $\sum_i f_i^A(x) = \sum_i f_i(x)$ so that total loss is conserved; note that the agents are simply exploring the space of possible non-negative group loss decompositions. We consider transformations of the form $\boldsymbol{f}^A(\boldsymbol{x}) = A^\top \boldsymbol{f}(\boldsymbol{x})$ where each agent i controls row i of A with each row constrained to the simplex, i.e. $A_i \in \Delta^{n-1}$. $\mathcal{X}^*$ denotes the set of Nash equilibria. $[a; b] = [a^\top, b^\top]^\top$ signifies row stacking of vectors.

2.2 Price of Anarchy

Nisan et al. (2007) define price of anarchy as the worst value of an equilibrium divided by the best value in the game. Here, value means sum of player losses, best means lowest, and Nash is the equilibrium. It is well known that Nash can be arbitrarily bad from both an individual agent and group perspective; Appendix B presents a simple example and demonstrates how opponent shaping is not a balm for these issues (Foerster et al., 2018; Letcher et al., 2018). With the above notation, the price of anarchy, ρ, is defined as

$$\rho_{\mathcal{X}}(\boldsymbol{f}^A) \stackrel{\text{def}}{=} \frac{\max_{\mathcal{X}^*} \sum_i f_i^A(\boldsymbol{x}^*)}{\min_{\mathcal{X}} \sum_i f_i^A(\boldsymbol{x})} \geq 1. \tag{1}$$

Note that computing the price of anarchy precisely requires solving both an optimization problem and a Nash equilibrium problem, which is PPAD-complete (Daskalakis et al., 2009; Gilboa and Zemel, 1989). We explain how we circumvent this issue with a local approximation in §2.4.

2.3 Compromise as an Optimization Problem

Given a game, we want to minimize the price of anarchy by perturbing the original agent losses:

$$\min_{\substack{\boldsymbol{f}' = \psi_A(\boldsymbol{f}) \\ \mathbf{1}^\top \boldsymbol{f}' = \mathbf{1}^\top \boldsymbol{f}}} \rho_{\mathcal{X}}(\boldsymbol{f}') + \nu \mathcal{D}(\boldsymbol{f}, \boldsymbol{f}') \tag{2}$$

where $\boldsymbol{f}$ and $\boldsymbol{f}' = \psi_A(\boldsymbol{f})$ denote the vectors of original and perturbed losses respectively, $\psi_A : \mathbb{R}^n \to \mathbb{R}^n$ is parameterized by weights A, ν is a regularization hyperparameter, and $\mathcal{D}$ penalizes deviation of the perturbed losses from the originals or represents constraints through an indicator function. To ensure minimizing the price of anarchy of the perturbed game improves on the original, we incorporate the constraint that the sum of perturbed losses equals the sum of original losses. We refer to this approach as ρ-minimization.

Our agents reconstruct their losses using the losses of all other agents as a basis. For simplicity, we consider linear transformations of their loss functions, although the theoretical bounds hereafter are independent of this simplification. We also restrict ourselves to convex combinations so that agents do not learn incentives that are directly adverse to other agents. The problem can now be reformulated. Let $\psi_A(f) = A^\top f$ and $\mathcal{D}(f, f') = \sum_i \mathcal{D}_{KL}(\boldsymbol{e}_i \mid\mid A_i)$ where $A \in \mathbb{R}^{n \times n}$ is a right stochastic matrix (rows are non-negative and sum to 1), $\boldsymbol{e}_i \in \mathbb{R}^n$ is a unit vector with a 1 at index i, and $\mathcal{D}_{KL}$ denotes the Kullback-Liebler divergence. Note OpenAI Five (OpenAI et al., 2019) also used a linear mixing approach where the "team spirit" mixture parameter (τ) is manually annealed throughout training from 0.3 to 1.0 (i.e., $A_{ii} = 1 - 0.8\tau, A_{ij} = 0.2\tau, j \neq i$).

2.4 A Local Price of Anarchy

The price of anarchy, $\rho \geq 1$, is defined over the joint strategy space of all players. Computing it is intractable for general games. However, many agents learn via gradient-based training, and so only observe the portion of the strategy space explored by their learning trajectory. Hence, we imbue our agents with the ability to locally estimate the price of anarchy along this trajectory.

[3] The price of anarchy assumes positive losses. This is accounted for in §2.5 to allow for losses in $\mathbb{R}$.

Definition 1 (*Local* Price of Anarchy)**.** Define

$$\rho_{\boldsymbol{x}}(\boldsymbol{f}^A, \Delta t) = \frac{\max_{\mathcal{X}_\tau^*} \sum_i f_i^A(\boldsymbol{x}^*)}{\min_{\tau \in [0, \Delta t]} \sum_i f_i^A(\boldsymbol{x} - \tau F(\boldsymbol{x}))} \geq 1 \tag{3}$$

where $F(\boldsymbol{x}) = [\nabla_{x_1} f_1^A(\boldsymbol{x}); \ldots; \nabla_{x_n} f_n^A(\boldsymbol{x})]$, Δt is a small step size, f_i^A is assumed positive $\forall\, i$, and $\mathcal{X}_\tau^*$ denotes the set of equilibria of the game when constrained to the line.

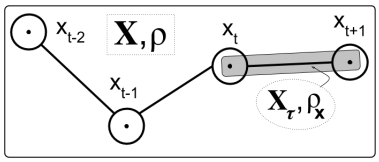


Figure 1: Agents estimate the price of anarchy assuming the joint strategy space, $\mathcal{X}$, of the game is restricted to a local linear region, $\mathcal{X}_\tau$, extending from the currently learned joint strategy, x_t, to the next, x_{t+1}. ρ and ρ_x denote the global and local price of anarchy.

To obtain bounds, we leverage theoretical results on *smooth games*, summarized as a class of games where "the externality imposed on any one player by the others is bounded" (Roughgarden, 2015). We assume a Lipschitz property on all $f_i^A(x)$ (details in Theorem 1), which allows us to appeal to this class of games. The bound in Eqn (4) is tight for some games. Proofs can be found in appendix D.

Theorem 1 (Local *Utilitarian* Price of Anarchy)**.** *Assuming each agent's loss is positive and its loss gradient is Lipschitz, there exists a learning rate $\Delta t > 0$ sufficiently small such that, to $\mathcal{O}(\Delta t^2)$, the local **utilitarian** price of anarchy of the game, $\rho_{\boldsymbol{x}}(\boldsymbol{f}^A, \Delta t)$, is upper bounded by*

$$\max_i \{1 + \Delta t\, \mathtt{ReLU}\Big(\frac{d}{dt} \log(f_i^A(\boldsymbol{x})) + \frac{||\nabla_{x_i} f_i^A(\boldsymbol{x})||^2}{f_i^A(\boldsymbol{x})\bar{\mu}}\Big)\} \tag{4}$$

where i indexes each agent, $\bar{\mu}$ is a user defined positive scalar, $\mathtt{ReLU}(z) \stackrel{\text{def}}{=} \max(z, 0)$, and Lipschitz implies there exists a β_i such that $||\nabla_{x_i} f_i^A(\boldsymbol{x}) - \nabla_{y_i} f_i^A(\boldsymbol{y})|| \leq \beta_i ||\boldsymbol{x} - \boldsymbol{y}||\ \forall \boldsymbol{x}, \boldsymbol{y}, A$.

Recall that this work focuses on price of anarchy defined using total loss as the value of the game. This is a *utilitarian* objective. We also derive an upper bound on the local *egalitarian* price of anarchy where value is defined as the max loss over all agents (replace $\sum_i$ with $\max_i$ in Eqn (3); see §D.2).

2.5 Decentralized Learning of the Loss Mixture Matrix A

Minimizing Eqn (2) w.r.t. A can become intractable if n is large. Moreover, if solving for A at each step is the responsibility of a central authority, the system is vulnerable to this authority failing. A distributed solution is therefore appealing, and the local price of anarchy bound admits a natural decomposition over agents. Equation 2 becomes

$$\min_{A_i \in \Delta^{n-1}} \rho_i + \nu \mathcal{D}_{KL}(\boldsymbol{e}_i \,||\, A_i) \tag{5}$$

where $\rho_i = 1 + \Delta t\, \mathtt{ReLU}\Big(\frac{d}{dt} \log(f_i^A(\boldsymbol{x})) + \frac{||\nabla_{x_i} f_i^A(\boldsymbol{x})||^2}{f_i^A(\boldsymbol{x})\bar{\mu}}\Big)$. This objective is differentiable w.r.t. each A_i with gradient $\nabla_{A_i} \rho_i \propto \nabla_{A_i} \mathtt{ReLU}\Big(\frac{d}{dt} \log(f_i^A(\boldsymbol{x})) + \frac{||\nabla_{x_i} f_i^A(\boldsymbol{x})||^2}{f_i^A(\boldsymbol{x})\bar{\mu}}\Big)$. The $\log$ appears due to price of anarchy being defined as the worst case Nash total loss *divided* by the minimal total loss. We propose the following modified learning rule for a hypothetical price of anarchy which is defined as a *difference* and accepts negative loss: $A_i \leftarrow A_i - \eta_A \tilde{\nabla}_{A_i} \rho_i$ where η_A is a learning rate and

$$\tilde{\nabla}_{A_i} \rho_i = \nabla_{A_i} \mathtt{ReLU}\Big(\frac{d}{dt} f_i^A(\boldsymbol{x}) + \epsilon\Big). \qquad [\epsilon \text{ is a hyperparameter.}] \tag{6}$$

The update direction in (6) is proportional to $\nabla_{A_i} \rho_i$ asymptotically for large f_i^A; see §D.1.1 for further discussion. Each agent i updates x_i and A_i simultaneously using $\nabla_{x_i} f_i^A(\boldsymbol{x})$ and $\tilde{\nabla}_{A_i} \rho_i$.

Improve-Stay, Suffer-Shift—$\nabla_{A_i} \rho_i$ encodes the rule: if the loss is decreasing, maintain the mixing weights, otherwise, change them. This strategy applies Win-Stay, Lose-Shift (WSLS) (Robbins,

1952) to learning (derivatives) rather than outcomes (losses). WSLS was shown to outperform Tit-for-Tat (Rapoport et al., 1965) in an iterated prisoner's dilemma (Nowak and Sigmund, 1993; Imhof et al., 2007).

Note that the trival solution of minimizing average group loss coincides with $A_{ij} = \frac{1}{n}$ for all i, j. If the agent strategies converge to a social optimum, this is a fixed point in the augmented strategy space (x, A). This can be seen by noting that 1) convergence to an optimum implies $\nabla_{x_i} f_i^A(x) = 0$ and 2) convergence alone implies $\frac{df_i}{dt} = 0$ for all agents so $\nabla A_i = 0$ by Eqn (6) assuming $\epsilon = 0$.

2.6 Decentralized Learning & Extending to Reinforcement Learning

The time derivative of each agent's loss, $\frac{d}{dt} f_i^A(\boldsymbol{x})$, in Eqn (6) requires differentiating through potentially all other agent loss functions, which precludes scaling to large populations. In addition, this derivative is not always available as a differentiable function. In order to estimate $\tilde{\nabla}_{A_i} \rho_i$ when only scalar estimates of ρ_i are available as in, e.g., reinforcement learning (RL), each agent perturbs their loss mixture and commits to this perturbation for a random number of training steps. If the loss increases over the trial, the agent updates their mixture in a direction *opposite* the perturbation. Otherwise, no update is performed.

This is formally accomplished with approximate one-shot gradient estimates (Shalev-Shwartz et al., 2012). A one-shot gradient estimate of $\rho_i(A_i)$ is performed by first evaluating $\rho_i(\log(A_i) + \delta \tilde{\boldsymbol{a}}_i)$ where δ is a scalar and $\tilde{\boldsymbol{a}}_i \sim U_{sp}(n)$ is drawn uniformly from the unit sphere in $\mathbb{R}^n$. Then, an unbiased gradient is given by $\frac{n}{\delta} \rho_i(\log(A_i) + \delta \tilde{\boldsymbol{a}}_i) \tilde{\boldsymbol{a}}_i$ where $A_i \in \Delta^{n-1}$. In practice, we cannot evaluate in one shot the $\frac{d}{dt} f_i^A(\boldsymbol{x})$ term that appears in the definition of ρ_i. Instead, Algorithm 1 uses finite differences and we assume the evaluation remains accurate enough across training steps.

Algorithm 1 requires arguments: η_A is a global learning rate for each A_i, δ is a perturbation scalar for the one-shot gradient estimate, $\tau_{\min}$ and τ_{max} specify the lower and upper bounds for the duration of the mixing trial for estimating a finite difference of $\frac{d}{dt} f_i^A(\boldsymbol{x})$, l and h specify lower and upper bounds for clipping A in logit space ($_l\lfloor \cdot \rceil^h$), and $\mathbb{L}_i$ is a learning algorithm that takes A as input (in order to mix rewards) and outputs *discounted return*. $\oslash$ indicates elementwise division.

Algorithm 1 D3C Update for RL Agent i

Input: η_A, δ, ν, $\tau_{\min}$, $\tau_{\max}$, A_i^0, ϵ, l, h, $\mathbb{L}$, iterations T
$A_i \leftarrow A_i^0$ {Initialize Mixing Weights}
{Draw Initial Random Mixing Trial}
$\tilde{A}_i, \tilde{\boldsymbol{a}}, \tau, t_b, G_b = \texttt{trial}(\delta, \tau_{\min}, \tau_{\max}, A_i, 0, G)$
$G = 0$ {Initialize Mean Return of Trial}
for $t = 0 : T$ **do**
  $g = \mathbb{L}_i(\tilde{A}_i \ \forall \ i)$ {Update Policy With Mixed Rewards}
  $\Delta t_b = t - t_b$ {Elapsed Trial Steps}
  $G = (G(\Delta t_b - 1) + g)/\Delta t_b$ {Update Mean Return}
  if $\Delta t_b == \tau$ {Trial Complete} **then**
    $\tilde{\rho}_i = \texttt{ReLU}(\frac{G_b - G}{\tau} + \epsilon)$ {Approximate ρ}
    $\nabla_{A_i} = \tilde{\rho}_i \tilde{\boldsymbol{a}} - \nu \boldsymbol{e}_i \oslash A_i$ {Estimate Gradient —(6)}
    $A_i = \texttt{softmax}\, {}_l\lfloor \log(A_i) - \eta_A \nabla_{A_i} \rceil^h$ {Update}
    {Draw New Random Mixing Trial}
    $\tilde{A}_i, \tilde{\boldsymbol{a}}, \tau, t_b, G_b = \texttt{trial}(\delta, \tau_{\min}, \tau_{\max}, A_i, t, G)$
  end if
end for

Algorithm 2 $\mathbb{L}_i$—example learner

Input: $\tilde{A} = [\tilde{A}_1; \ldots; \tilde{A}_n]$
while episode not terminal **do**
  draw action from agent policy
  play action and observe reward r_i
  broadcast r_i to all agents
  update policy with $\tilde{r}_i = \sum_j \tilde{A}_{ji} r_j$
end while
Output: return over episode g

Algorithm 3 `trial`—helper function

Input: δ, $\tau_{\min}$, $\tau_{\max}$, A_i, t, G
{Sample Perturbation Direction}
$\tilde{\boldsymbol{a}}_i \sim U_{sp}(n)$
{Perturb Mixture}
$\tilde{A}_i = \texttt{softmax}(\log(A_i) + \delta \tilde{\boldsymbol{a}}_i)$
{Draw Random Trial Length}
$\tau \sim \texttt{Uniform}\{\tau_{\min}, \tau_{\max}\}$
Output: $\tilde{A}_i, \tilde{\boldsymbol{a}}, \tau, t, G$

2.7 Assessment

We assess Algorithm 1 with respect to our original design criteria. As described, agents perform gradient descent on a decentralized and local upper bound on the price of anarchy. Recall that a minimal global price of anarchy ($\rho = 1$) implies that even the worst case Nash equilibrium of the

game is socially optimal; similarly, Algorithm 1 searches for a locally socially optimal equilibrium. By design, $A_i \in \Delta^{n-1}$ ensures the approach is budget-balancing. We justify the agents learning weight vectors A_i by initializing them to attend primarily to their own losses as in the original game. If they can minimize their original loss, then they never shift attention according to Eqn (6) because $\frac{df_i}{dt} \leq 0$ for all t. They only shift A_i if their loss increases. We also include a KL term to encourage the weights to return to their initial values. In addition, in our experiments with symmetric games, learning A helps the agents' outcomes in the long run. We also consider experiments in Appendix E.2.2 where only a subset of agents opt into the mechanism. If each agent's original loss is convex with diagonally dominant Hessian and the strategy space is unconstrained, the the unique, globally stable fixed point of the game defined with mixed losses is a Nash (see Appendix H.4). Exact gradients $\nabla_{A_i} \rho_i$ require each agent differentiates through all other agents losses precluding a fully decentralized and scalable algorithm. We circumvent this issue with noisy oneshot gradients. All that is needed in terms of centralization is to share the mixed scalar rewards; this is cheap compared to sharing $x_i \in \mathbb{R}^d$. As mentioned in the introduction, the cost of communicating rewards can be mitigated by learning A_i via sparse optimization or sampling but is outside the scope of this paper.

3 Experiments

Here, we show that agents minimizing local estimates of price of anarchy achieve lower loss on average than selfish, rational agents in three domains. Due to space, we leave two other domains to the appendix. In the first domain, a traffic network (4 players), players optimize using exact gradients (see Eqn (6)). Then in two RL domains, Coins and Cleanup, players optimize with approximate gradients as handled by Algorithm 1. Agents train with deep networks and A2C (Espeholt et al., 2018). We refer to both algorithms as D3C (decentralized, differentiable, dynamic compromise).

For D3C, we initialize $A_{ii} = 0.99$ and $A_{ij} = \frac{0.01}{n-1}$, $j \neq i$. We initialize away from a one-hot because we use entropic mirror descent (Beck and Teboulle, 2003) to update A_i, and this method requires iterates to be initialized to the interior of the simplex. In the RL domains, updates to A_i are clipped in logit-space to be within $l = -5$ and $h = 5$ (see Algorithm 1). We set the $\mathcal{D}_{KL}$ coefficient to 0 except for in Coins, where $\nu = 10^{-5}$. Additional hyperparameters are specified in §G. In experiments, reported price of anarchy refers to the ratio of the sum of losses of the strategy learning converged to over that of the strategy learned by fully cooperative agents ($A_{ij} = \frac{1}{n}$).

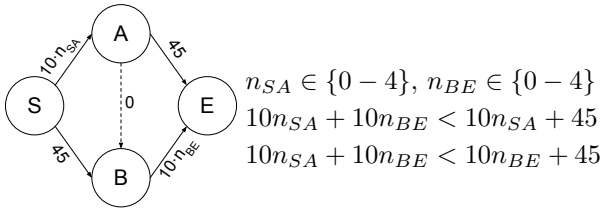
3.1 Traffic Networks and Braess's Paradox

In 2009, New York city's mayor closed Broadway near Times Square to alleviate traffic congestion (Neuman and Barbaro, 2009). This counter-intuitive phenomenon, where restricting commuter choices improves outcomes, is called Braess's paradox (Wardrop, 1952; Beckmann et al., 1956; Braess, 1968), and has been observed in real traffic networks (Youn et al., 2008; Steinberg and Zangwill, 1983). Braess's paradox is also found in physics (Youn et al., 2008), decentralized energy grids (Witthaut and Timme, 2012), and can cause extinction cascades in ecosystems (Sahasrabudhe and Motter, 2011). Knowing when a network may exhibit this paradox is difficult, which means knowing when network dynamics may result in poor outcomes is difficult.

Figure 2a presents a theoretical traffic network. Without edge AB, drivers commute according to the Nash equilibrium, either learned by gradient descent or D3C. Figure 3a shows the price of anarchy approaching 1 for both algorithms. If edge AB is added, the network now exhibits Braess's paradox. Figure 3b shows that while gradient descent converges to Nash ($\rho = \frac{80}{65}$), D3C achieves a price of anarchy near 1. Figure 2b shows that when faced with a randomly drawn network, D3C agents achieve shorter commutes on average than agents without the ability to compromise.

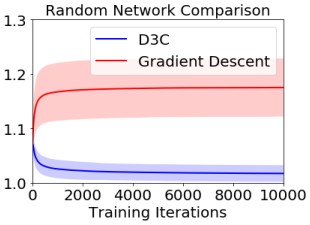
3.2 Coin Dilemma

In the Coins game (Eccles et al., 2019a; Lerer and Peysakhovich, 2017), two agents move on a fully-observed 5×5 gridworld, on which coins of two types corresponding to each agent randomly spawn at each time step with probability 0.005. When an agent moves into a square with a coin of either type, they get a reward of 1. When an agent picks up a coin of the other player's type, the other agent gets a reward of -2. The episode lasts 500 steps. Total reward is maximized when each agent picks up only coins of their own type, but players are tempted to pick up all coins.



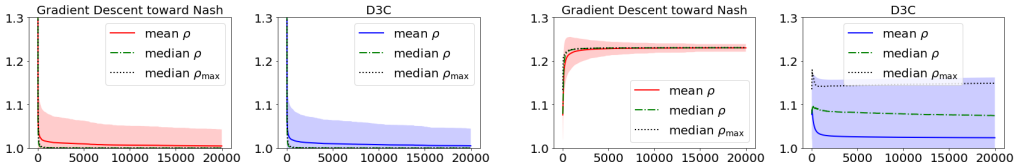
$$n_{SA} \in \{0-4\},\ n_{BE} \in \{0-4\}$$
$$10n_{SA} + 10n_{BE} < 10n_{SA} + 45$$
$$10n_{SA} + 10n_{BE} < 10n_{BE} + 45$$

(a) Traffic Network



(b) Random Network Results

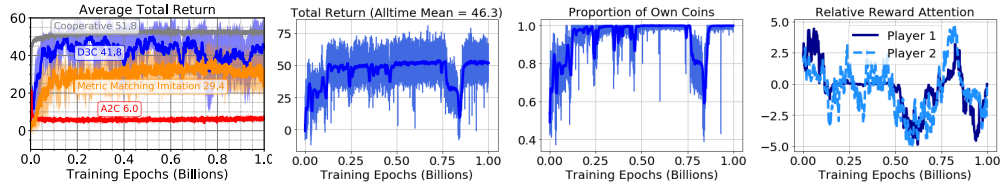
Figure 2: (a) Four drivers aim to minimize commute time from S to E. Commute time on each edge depends on the number of commuters, n_{ij}. Without edge AB, drivers distribute evenly across SAE and SBE for a 65 min commute. After edge AB is added, switching to the shortcut, SABE, always decreases commute time given the other drivers maintain their routes, however, all drivers are incentivized to take the shortcut resulting in an 80 min commute. (b) The price of anarchy throughout training for 1000 randomly generated three road networks exhibiting Braess's paradox. Gradient descent and D3C are compared with shaded regions representing $\pm$ 1 standard devation.



(a) Without Shortcut (Edge AB Removed)

(b) With Shortcut (Edge AB Included)

Figure 3: **Traffic Network** (a) Without edge AB, agents are initialized with random strategies and train with either gradient descent (left) or D3C (right)—similar performance expected. Statistics of 1000 runs are plotted over training. Median $\rho_{\max}$ tracks the median over trials longest-commute among the four drivers. The shaded region captures $\pm$ 1 standard deviation around the mean. (b) After edge AB is added, agents are initialized with random strategies and trained with either gradient descent (left) or D3C (right). Statistics of 1000 runs are plotted over training. Median $\rho_{\max}$ tracks the median over trials of the longest-commute among the four drivers. The shaded region captures $\pm$ 1 standard deviation around the mean.



(a) Trial Statistics

(b) Individual Run

Figure 4: **Coin Dilemma** (a) Mean total return over ten training runs for agents. Mean return over all epochs is next to each method name in the legend. D3C hyperparameters were selected using five independent validation runs. Cooperative agents trained to maximize total return represent the best possible baseline. Shaded region captures $\pm$ 1 standard deviation around the mean. (b) One training run ($A^0_{ii} = 0.9$): sum of agent returns (left); % of coins picked up that were the agent's type (middle); relative reward attention measured as $\ln\left(\frac{A_{ii}}{A_{j\neq i}}\right)$ (right).

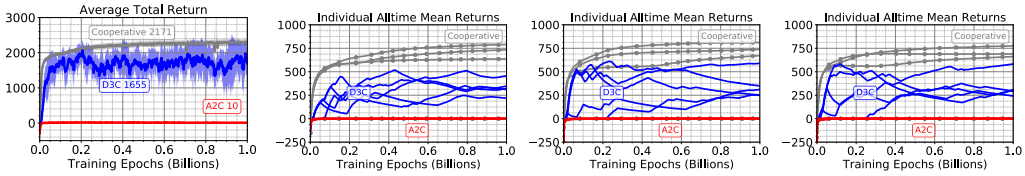
D3C agents approach optimal cooperative returns (see Figure 4a). We compare against Metric Matching Imitation (Eccles et al., 2019b), which was previously tested on Coins and designed to exhibit reciprocal behavior towards co-players.

Figure 4b shows D3C agents learning to cooperate, then temporarily defecting before rediscovering cooperation. Note that the relative reward attention of both players spikes towards selfish during this small defection window; agents collect more of their opponent's coins during this time. Oscillating between cooperation and defection occurred across various hyperparameter settings. Relative reward

attention trajectories between agents appear to be reciprocal, i.e., move in relative synchrony (see §H.2 for analysis).

3.3 CLEANUP

We provide additional results on Cleanup, a five-player gridworld game (Hughes et al., 2018). Agents are rewarded for eating apples, but must keep a river clean to ensure the apples receive sufficient nutrients. The option to be a freeloader and only eat apples presents a social dilemma. D3C is able to increase both welfare and individual reward over A2C (no loss sharing). We also observe that direct welfare maximization (Cooperation) always results in three agents collecting rewards from apples while two agents sacrifice themselves and clean the river. In contrast, D3C avoids this stark division of labor. Agents take turns on each task and all achieve some positive cumulative return over training.



(a) Cleanup (b) Individual Runs

Figure 5: (a) Mean total return over ten training runs for agents. D3C hyperparameters were selected using five independent validation runs. Cooperative agents trained to maximize total return represent the best possible baseline. Shaded region captures $\pm$ 1 standard deviation around the mean. (b) Three randomly selected runs. Each curve shows the mean return up to the current epoch for 1 of 5 agents.

3.4 A ZERO-SUM ELECTION

Consider a hierarchical election in which two parties compete in a zero-sum game—for example, only one candidate becomes president. If, at the primary stage, candidates within one party engage in negative advertising, they hurt their chances of winning the presidential election because these ads are now out in the open. This presents a prisoner's dilemma within each party. The goal then is for each party to solve their respective prisoner's dilemma and come together as one team, but certainly not maximize welfare—the zero-sum game between the two parties should be retained. A simple simulation with two parties consisting of two candidates each initially participating in negative advertising converges to the desired result after running D3C.

The final 4×4 loss mixing matrix, A, after training 1000 steps is an approximate block matrix with 0.46 on the 2×2 block diagonal and 0.04 elsewhere. We make a *duck-typing* argument that when multiple agents are optimizing the same loss, they are functioning as multiple components of a single agent because mathematically, there is no difference between this multi-agent system and a single agent optimization problem. This matrix then indicates that two approximate teams have formed: the first two agents captured by the upper left block and vice versa. Furthermore, the final eigenvalues of the game Jacobian are $(1.84 \pm 0.21i)$ $\times 2$; perfect team formation gives $(2 \pm 0.25i)$ $\times 2$. The existence of imaginary eigenvalues indicates that the zero-sum component of the game is retained. In contrast, minimizing total loss gives 0 imaginary part because Hessians ($\texttt{Jac}(\nabla)$) are symmetric.

4 CONCLUSION

Directly maximizing welfare can solve many social dilemmas, but it fails to draw out the rich behavior we would expect from agents in other interesting scenarios. We formulate learning incentives as a price of anarchy minimization problem and propose a decentralized, gradient-based approach, namely D3C, that incrementally adapts agent incentives to the environment at hand. We demonstrate its effectiveness on achieving near-optimal agent outcomes in socially adversarial environments. Importantly, it also generates reasonable responses where welfare maximization is indifferent.

REFERENCES

Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 1970.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Martin Beckmann, Charles B McGuire, and Christopher B Winsten. Studies in the economics of transportation. Technical report, 1956.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations research*, 59(1):17–31, 2011.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. On the efficiency-fairness trade-off. *Management Science*, 58(12):2234–2250, 2012.

Dietrich Braess. Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung*, 12(1): 258–268, 1968.

Ennio Cavazzuti, Massimo Pappalardo, Mauro Passacantando, et al. Nash equilibria, variational inequalities, and dynamical systems. *Journal of optimization theory and applications*, 114(3): 491–506, 2002.

Edward H Clarke. Multipart pricing of public goods. *Public choice*, pages 17–33, 1971.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. The imitation game: Learned reciprocity in markov games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1934–1936. International Foundation for Autonomous Agents and Multiagent Systems, 2019a.

Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*, 2019b.

Gerald M Edelman and Joseph A Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24):13763–13768, 2001.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

Kurtuluş Gemici, Elias Koutsoupias, Barnabé Monnot, Christos Papadimitriou, and Georgios Piliouras. Wealth inequality and the price of anarchy. *arXiv preprint arXiv:1802.09269*, 2018.

Itzhak Gilboa and Eitan Zemel. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior*, 1(1):80–93, 1989.

Jerry Green and Jean-Jacques Laffont. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica: Journal of the Econometric Society*, pages 427–438, 1977.

Jerry R Green and Jean-Jacques Laffont. Incentives in public decision making. 1979.

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

Jason D Hartline and Tim Roughgarden. Optimal mechanism design and money burning. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 75–84. ACM, 2008.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336, 2018.

Lorens A Imhof, Drew Fudenberg, and Martin A Nowak. Tit-for-tat or win-stay, lose-shift? *Journal of theoretical biology*, 247(3):574–580, 2007.

Raghavendra V Kulkarni, Anna Förster, and Ganesh Kumar Venayagamoorthy. Computational intelligence in wireless sensor networks: A survey. *IEEE communications surveys & tutorials*, 13 (1):68–96, 2010.

Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, and Christoph Koch. Sparse projections onto the simplex. In *International Conference on Machine Learning*, pages 235–243, 2013.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. *arXiv preprint arXiv:1811.08469*, 2018.

Ping Li, Syama Sundar Rangapuram, and Martin Slawski. Methods for sparse and low-rank recovery under simplex constraints. *arXiv preprint arXiv:1605.00507*, 2016.

Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duéñez-Guzmán, Edward Hughes, and Joel Z. Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

Michael P Murray. A drunk and her dog: An illustration of cointegration and error correction. *The American Statistician*, 48(1):37–39, 1994.

Roger B Myerson and Mark A Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2):265–281, 1983.

Anna Nagurney and Ding Zhang. *Projected dynamical systems and variational inequalities with applications*, volume 2. Springer Science & Business Media, 2012.

William Neuman and Michael Barbaro. Mayor plans to close parts of broadway to traffic. *NYTimes.com*, Feb 2009.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*. Cambridge university press, 2007.

Martin Nowak and Karl Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(6432):56, 1993.

Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. 2019. URL https://arxiv.org/abs/1912.06680.

Mert Pilanci, Laurent E Ghaoui, and Venkat Chandrasekaran. Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2012.

Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Michael H Rothkopf. Thirteen reasons why the vickrey-clarke-groves process is not practical. *Operations Research*, 55(2):191–197, 2007.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5): 32, 2015.

Tim Roughgarden and Florian Schoppmann. Local smoothness and the price of anarchy in splittable congestion games. *Journal of Economic Theory*, 156:317–342, 2015.

Sagar Sahasrabudhe and Adilson E Motter. Rescuing ecosystems from extinction cascades through compensatory perturbations. *Nature Communications*, 2:170, 2011.

Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2): 187–217, 1975.

Florian Schäfer and Anima Anandkumar. Competitive gradient descent. In *Advances in Neural Information Processing Systems*, pages 7623–7633, 2019.

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

Richard Steinberg and Willard I Zangwill. The prevalence of braess' paradox. *Transportation Science*, 17(3):301–318, 1983.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993.

John Glen Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers*, 1(3):325–362, 1952.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Daniel J Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.

Dirk Witthaut and Marc Timme. Braess's paradox in oscillator networks, desynchronization and power outage. *New journal of physics*, 14(8):083036, 2012.

Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents. *arXiv preprint arXiv:2006.06051*, 2020.

Hyejin Youn, Michael T Gastner, and Hawoong Jeong. Price of anarchy in transportation networks: efficiency and optimality control. *Physical review letters*, 101(12):128701, 2008.