Two-Player Games for Efficient Non-Convex Constrained Optimization

Andrew Cotter ACOTTER@GOOGLE.COM
Heinrich Jiang HEINRICHJ@GOOGLE.COM
Google AI

Karthik Sridharan SRIDHARAN@CS.CORNELL.EDU
Cornell University

Editors: Aurélien Garivier and Satyen Kale

Abstract

In recent years, constrained optimization has become increasingly relevant to the machine learning community, with applications including Neyman-Pearson classification, robust optimization, and fair machine learning. A natural approach to constrained optimization is to optimize the Lagrangian, but this is not guaranteed to work in the non-convex setting, and, if using a first-order method, cannot cope with non-differentiable constraints (e.g. constraints on rates or proportions).

The Lagrangian can be interpreted as a two-player game played between a player who seeks to optimize over the model parameters, and a player who wishes to maximize over the Lagrange multipliers. We propose a non-zero-sum variant of the Lagrangian formulation that can cope with non-differentiable—even discontinuous—constraints, which we call the "proxy-Lagrangian". The first player minimizes external regret in terms of easy-to-optimize "proxy constraints", while the second player enforces the *original* constraints by minimizing swap regret.

For this new formulation, as for the Lagrangian in the non-convex setting, the result is a stochastic classifier. For both the proxy-Lagrangian and Lagrangian formulations, however, we prove that this classifier, instead of having unbounded size, can be taken to be a distribution over no more than $m + 1$ models (where m is the number of constraints). This is a significant improvement in practical terms.

1. Introduction

We consider the general problem of inequality constrained optimization, in which we wish to find a set of parameters $\theta \in \Theta$ minimizing an objective function subject to m functional constraints:

$$\begin{aligned} &\min_{\theta\in\Theta} g_0\left(\theta\right) \\ &\text{s.t. } \forall i \in [m] \,.\, g_i\left(\theta\right) \le 0 \end{aligned} \tag{1}$$

To highlight some of the challenges that arise in non-convex constrained optimization, consider the specific example of constraining a *fairness* metric. We cast the fairness problem as that of minimizing some empirical loss subject to one or more fairness constraints. One of the simplest examples of such is the following:

$$\begin{aligned}\min_{\theta\in\Theta}\ & \frac{1}{|S|}\sum_{x,y\in S}\ell\left(f\left(x;\theta\right),y\right) \\ \text{s.t. } & \frac{1}{|S|}\sum_{x\in S_{\min}}\mathbf{1}_{f(x;\theta)>0}\geq\frac{0.8}{|S|}\sum_{x\in S}\mathbf{1}_{f(x;\theta)>0}\end{aligned} \tag{2}$$

Here, $f\left(\cdot;\theta\right)$ is a classification function with parameters θ, S is the training dataset, and $S_{\min}\subseteq S$ represents a minority population. The constraint represents a version of the so-called "80% rule" (e.g. Biddle, 2005; Vuolo and Levy, 2013), and forces the resulting classifier to make at least 80% of its positive predictions on the minority population. Unfortunately, several serious challenges arise when we attempt to optimize this problem:

1. The constraint is data-dependent, and could therefore be very expensive to check.
2. The classification function f may be a badly-behaving function of θ (e.g. a deep neural network), resulting in non-convex objective and constraint functions.
3. Worse, the constraint is a linear combination of *indicators*, hence is not even subdifferentiable w.r.t. θ.

Perhaps the most "familiar" technique for constrained optimization is to formulate the Lagrangian:

Definition 1 *The Lagrangian $\mathcal{L}:\Theta\times\Lambda\rightarrow\mathbb{R}$ of Equation 1 is:*

$$\mathcal{L}\left(\theta,\lambda\right):=g_0\left(\theta\right)+\sum_{i=1}^{m}\lambda_i g_i\left(\theta\right)$$

where $\Lambda\subseteq\mathbb{R}_+^m$.

and jointly minimize over $\theta\in\Theta$ and maximize over $\lambda\in\Lambda\subseteq\mathbb{R}_+^m$. By itself, using this formulation doesn't address the challenges we identified above, but we will see that, compared to the alternatives (Section 2.1), it's a good starting point for an approach that does.

1.1. Dealing with non-Convexity

Optimizing the Lagrangian can be interpreted as playing a two-player zero-sum game: the first player chooses θ to minimize $\mathcal{L}\left(\theta,\lambda\right)$, and the second player chooses λ to maximize it. The essential difficulty is that, without strong duality—equivalently, unless the minimax theorem holds, giving that $\min_{\theta\in\Theta}\max_{\lambda\in\Lambda}\mathcal{L}\left(\theta,\lambda\right)=\max_{\lambda\in\Lambda}\min_{\theta\in\Theta}\mathcal{L}\left(\theta,\lambda\right)$—then the θ-player, who is working on the primal (minimax) problem, and the λ-player, who is working on the dual (maximin) problem, might fail to converge to a solution satisfying both players simultaneously (i.e. a pure Nash equilibrium).

If Equation 1 is a convex optimization problem and the action spaces Θ and Λ are compact and convex, then the minimax theorem holds (von Neumann, 1928), and optimizing the Lagrangian will work. Otherwise it might not, and in fact it's quite easy to construct a counterexample: Figure 1 shows a case in which a pure Nash equilibrium of the Lagrangian game *does not exist*. For this reason, the standard approach for handling non-convex machine learning problems, i.e. pretending that the problem is convex and using a stochastic first order algorithm anyway, should not be expected to reliably converge to a pure Nash equilibrium—even on a problem as trivial as that in Figure 1—since there may be none for it to converge *to*.

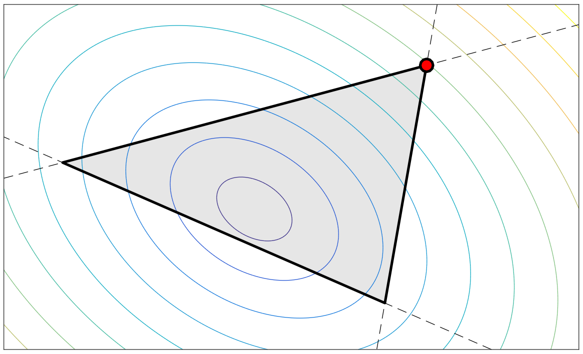

Figure 1: The plotted rectangular region is the domain Θ, the contours are those of the *strictly concave* minimization objective function g_0, and the shaded triangle is the feasible region determined by the three linear inequality constraints $g_1, \dots, g_3$. The red dot is the optimal feasible point. The Lagrangian $\mathcal{L}(\theta, \lambda)$ is strictly concave in θ for any choice of λ, so the optimal choice(s) for the θ-player will always lie on the boundary of the plotted rectangle. However, these points are infeasible, and therefore suboptimal for the λ-player.

Under general conditions, however, even when there is no *pure* Nash equilibrium, a *mixed* equilibrium (i.e. a pair of distributions over θ and λ) does exist. Such an equilibrium defines a stochastic classifier: upon receiving an example x to classify, one would sample θ from its equilibrium distribution, and then evaluate the classification function $f(x; \theta)$. Furthermore, and this is our first main contribution, this equilibrium can be taken to consist of a discrete distribution over at most $m+1$ distinct θs (m being the number of constraints), and a single non-random λ. This is a crucial improvement in practical terms, since a machine learning model consisting of e.g. a distribution over thousands (or more) of deep neural networks—or worse, a continuous distribution—would likely be so unwieldy as to be unusable.

1.2. Introducing Proxy Constraints

Most real-world machine learning implementations perform optimization using a first-order method (even on non-convex problems, e.g. DNNs). To use such a method, however, one must have gradients, and gradients are unavailable for non-differentiable constraints like that in the fairness example of Equation 2, or in the myriad of other situations in which one wishes to constrain *counts* or *proportions* instead of smooth losses (e.g. recall, coverage or churn as in Goh et al. (2016)). In all of these cases, the constraint functions are piecewise-constant, so their gradients are zero almost everywhere, and a gradient-based method cannot be expected to succeed.

The obvious solution is to use a surrogate. For example, one could replace the indicators of Equation 2 with sigmoids, and then optimize the Lagrangian. This solves the differentiability problem, but introduces a new one: a (mixed) Nash equilibrium would correspond to a solution satisfying the sigmoid-relaxed constraint, instead of the *actual* constraint. Interestingly, it turns out that we can seek to satisfy the original un-relaxed constraint, even while using a surrogate. Our proposal is motivated by the observation that, while differentiating the Lagrangian (Definition 1) w.r.t. θ requires differentiating the constraint functions $g_i(\theta)$, to differentiate it w.r.t. λ we only need to *evaluate* them. Hence, a surrogate is only necessary for the θ-player; the λ-player can continue to use the original constraint functions.

We refer to a surrogate that is used by only one of the two players as a "proxy", and introduce the notion of "proxy constraints" by taking $\tilde{g}_i(\theta)$ to be a sufficiently-smooth upper bound on $g_i(\theta)$ for $i \in [m]$, and formulating two functions that we call "proxy-Lagrangians":

Definition 2 *Given proxy constraint functions $\tilde{g}_i(\theta) \geq g_i(\theta)$ for $i \in [m]$, the proxy-Lagrangians $\mathcal{L}_\theta, \mathcal{L}_\lambda : \Theta \times \Lambda \rightarrow \mathbb{R}$ of Equation 1 are:*

$$\begin{aligned}\mathcal{L}_\theta(\theta, \lambda) :=& \lambda_1 g_0(\theta) + \sum_{i=1}^{m} \lambda_{i+1} \tilde{g}_i(\theta) \\ \mathcal{L}_\lambda(\theta, \lambda) :=& \sum_{i=1}^{m} \lambda_{i+1} g_i(\theta)\end{aligned}$$

where $\Lambda := \Delta^{m+1} \ni \lambda$ is the $(m+1)$-dimensional simplex.

As one might expect, the θ-player wishes to minimize $\mathcal{L}_\theta(\theta, \lambda)$, while the λ-player wishes to maximize $\mathcal{L}_\lambda(\theta, \lambda)$. Notice that the $\tilde{g}_i$s are *only* used by the θ-player. Intuitively, the λ-player chooses how much to weigh the proxy constraint functions, but—and this is the key to our proposal—does so in such a way as to satisfy the *original* constraints.

Unfortunately, because the two players are optimizing different functions, this is a non-zero-sum game, and finding a (mixed) Nash equilibrium of such a game is known to be PPAD-complete even in the finite setting (Chen and Deng, 2006). We prove, however, that a *weaker* type of equilibrium (a Φ-correlated equilibrium (Rakhlin et al., 2011), i.e. a joint distribution over θ and λ w.r.t. which neither player can improve)—one that we *can* find efficiently—suffices to guarantee a nearly-optimal and nearly-feasible solution to Equation 1 in expectation.

1.3. Contributions

We first focus on the standard Lagrangian formulation, in the non-convex setting. In Section 3, we provide an algorithm that, given access to an approximate Bayesian optimization oracle, finds a stochastic classifier that, in expectation, is provably approximately feasible and optimal. Many previous authors have approached constrained optimization using similar techniques (see Section 2)—our main contribution is to show how such a classifier can be efficiently "shrunk" to one that is *at least as good*, but is supported on only $m+1$ solutions.

Our next major contribution is the introduction of the proxy-Lagrangian formulation, which allows us to optimize constrained problems with extremely general (even non-differentiable) constraints. In Section 4, we prove that a particular type of Φ-correlated equilibrium results in a stochastic classifier that is feasible and optimal, and go on to provide a novel algorithm that converges to such an equilibrium. Interestingly, to get the "right" sort of equilibrium, the θ-player needs only minimize the usual external regret, but the λ-player must minimize the *swap regret*. While the resulting distribution is supported on a large number of (θ, λ) pairs, applying the same "shrinking" procedure as before yields a distribution over only $m+1$ θs that is at least as good as the original.

Finally, in Section 5, we tie everything together by describing an end-to-end recipe for provably solving a non-convex constrained optimization problem with potentially non-differentiable constraints, yielding a stochastic model that is a supported on at most $m+1$ solutions. In practice, one would use SGD instead of an oracle, which results in an efficient procedure that can be easily plugged-in to existing workflows, as is experimentally verified in Section 6.

2. Related Work

The interpretation of constrained optimization as a two-player game has a long history: Arora et al. (2012) surveys some such work, and there are several more recent examples (e.g. Kearns et al., 2017; Narasimhan, 2018; Agarwal et al., 2018). In particular, Agarwal et al. (2018) propose an algorithm for fair classification that is very similar to the Lagrangian-based approach that we outline in Section 3—the main differences are our introduction of "shrinking", and that our setting (Equation 1) is more general. The recent work of Chen et al. (2017) addresses non-convex robust optimization, i.e. problems of the form:

$$\min_{\theta \in \Theta} \max_{i \in [m]} g_i\left(\theta\right)$$

Like both us and Agarwal et al. (2018), they: (i) model such a problem as a two-player game where one player chooses a mixture of objective functions, and the other player minimizes the loss of the mixture, and (ii) they find a *distribution* over solutions rather than a pure equilibrium. These similarities are unsurprising in light of the fact that robust optimization can be reformulated as constrained optimization via the introduction of a slack variable:

$$\begin{aligned} &\min_{\theta \in \Theta, \xi \in \Xi} \xi \\ &\text{s.t. } \forall i \in [m] \,. \xi \geq g_i\left(\theta\right) \end{aligned} \tag{3}$$

Correspondingly, one can transform a robust problem to a constrained one at the cost of an extra bisection search (e.g. Christiano et al., 2011; Rakhlin and Sridharan, 2013). As this relationship suggests, our main contributions can be adapted to the robust optimization setting. In particular: (i) our proposed shrinking procedure can be applied to Equation 3 to yield a distribution over only $m+1$ solutions, and (ii) one could perform robust optimization over non-differentiable (even discontinuous) losses using "proxy objectives", just as we use proxy constraints.

2.1. Alternative Approaches

Given the difficulties involved in using a Lagrangian-like formulation for non-convex problems, it's natural to wonder whether one should instead favor a procedure based on entirely different principles. Unfortunately, the alternatives each present their own challenges.

The potential complexity of the constraints all but rules out approaches based on projections (e.g. projected SGD) or optimization of constrained subproblems (e.g. Frank-Wolfe, as in Hazan and Kale (2012); Jaggi (2013); Garber and Hazan (2013)). Similarly, attempting to penalize violations (e.g. Arora et al., 2012; Rakhlin and Sridharan, 2013; Mahdavi et al., 2012; Cotter et al., 2016), for example by adding $\gamma \max_{i \in [m]} \max\left\{0, g_i\left(\theta\right)\right\}$ to the objective, where $\gamma \in \mathbb{R}_+$ is a hyperparameter, and optimizing the resulting problem using a first order method, fails if the constraint functions are non-differentiable. Even if they are, they may still be data-dependent, so evaluating g_i, or even determining whether it is positive (as is necessary for such techniques, due to the max with 0), requires enumerating over the entire dataset. Hence, unlike the Lagrangian and proxy-Lagrangian formulations, such "penalized" formulations are incompatible with the use of a computationally-cheap stochastic optimizer.

In response to the idea of proxy constraints, it's natural to ask "why not just relax the constraints for *both* players, instead of just the θ-player?". This is indeed a popular approach, having been proposed e.g. for Neyman-Pearson classification (Davenport et al., 2010; Gasso et al., 2011), more general rate metrics (Goh et al., 2016), and AUC (Eban et al., 2017). The answer is that in many cases, particularly when constraints are data dependent, they represent real-world restrictions on how the learned model is permitted to behave. For example, the "80% rule" of Equation 2 can be found in the HOPA Act of 1995 (Wikipedia, 2018), and it requires an 80% threshold in terms of the *number of positive predictions*—not a relaxation—which is precisely the target that the proxy-Lagrangian approach will attempt to hit.

This point, in turn, raises the question of *generalization*: satisfying the correct un-relaxed constraints on training data does not necessarily mean that they will be satisfied at evaluation time. This issue is outside the scope of this paper, but is vital. For certain specific applications, the post-training correction approach of Woodworth et al. (2017) can improve generalization performance, and Cotter et al. (2018)'s more recent proposal (which is based on our proxy-Lagrangian formulation) can be applied more generally, but there is still room for future work.

3. Starting Point: Lagrangian Optimization

Our ultimate interest is in constrained optimization, so before we present our proposed algorithm for optimizing the Lagrangian (Definition 1) in the non-convex setting, we will characterize the relationship between an approximate Nash equilibrium of the Lagrangian game, and a nearly-optimal nearly-feasible solution to the original constrained problem (Equation 1):

Theorem 3 *Define* $\Lambda := \left\{ \lambda \in \mathbb{R}_+^m : \left\Vert \lambda \right\Vert_1 \le R \right\}$*, and let* $\theta^{(1)}, \dots, \theta^{(T)} \in \Theta$ *and* $\lambda^{(1)}, \dots, \lambda^{(T)} \in \Lambda$ *be sequences of parameter vectors and Lagrange multipliers that comprise an approximate mixed Nash equilibrium, i.e.:*

$$\max_{\lambda^* \in \Lambda} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(\theta^{(t)}, \lambda^*\right) - \inf_{\theta^* \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(\theta^*, \lambda^{(t)}\right) \le \epsilon$$

Define $\bar{\theta}$ *as a random variable for which* $\bar{\theta} = \theta^{(t)}$ *with probability* $1/T$*, and let* $\bar{\lambda} := \left(\sum_{t=1}^{T} \lambda^{(t)} \right) / T$*. Then* $\bar{\theta}$ *is nearly-optimal in expectation:*

$$\mathbb{E}_{\bar{\theta}}\left[g_0\left(\bar{\theta}\right) \right] \le \inf_{\theta^* \in \Theta : \forall i . g_i(\theta^*) \le 0} g_0\left(\theta^*\right) + \epsilon$$

and nearly-feasible:

$$\max_{i \in [m]} \mathbb{E}_{\bar{\theta}}\left[g_i\left(\bar{\theta}\right) \right] \le \frac{\epsilon}{R - \left\Vert \bar{\lambda} \right\Vert_1} \tag{4}$$

Additionally, if there exists a $\theta' \in \Theta$ *that satisfies all of the constraints with margin* γ *(i.e.* $g_i\left(\theta'\right) \le -\gamma$ *for all* $i \in [m]$*), then:*

$$\left\Vert \bar{\lambda} \right\Vert_1 \le \frac{\epsilon + B_{g_0}}{\gamma}$$

where $B_{g_0} \ge \sup_{\theta \in \Theta} g_0\left(\theta\right) - \inf_{\theta \in \Theta} g_0\left(\theta\right)$ *is a bound on the range of the objective function* g_0*.*

Proof This is a special case of Theorem 11 and Lemma 12 in Appendix A. ■

This theorem has a few differences from the more typically-encountered equivalence between Nash equilibria and optimal feasible solutions in the convex setting. First, it characterizes *mixed*

Algorithm 1 Optimizes the Lagrangian formulation (Definition 1) in the non-convex setting via the use of an approximate Bayesian optimization oracle $\mathcal{O}_\rho$ (Definition 4) for the θ-player. The parameter R is the radius of the Lagrange multiplier space $\Lambda := \left\{\lambda \in \mathbb{R}_+^m : \|\lambda\|_1 \leq R\right\}$, and the function Π_Λ projects its argument onto Λ w.r.t. the Euclidean norm.

```
OracleLagrangian ($R \in \mathbb{R}_+, \mathcal{L} : \Theta \times \Lambda \to \mathbb{R}, \mathcal{O}_\rho : (\Theta \to \mathbb{R}) \to \Theta, T \in \mathbb{N}, \eta_\lambda \in \mathbb{R}_+$):
    Initialize $\lambda^{(1)} = 0$
    For $t \in [T]$:
        Let $\theta^{(t)} = \mathcal{O}_\rho\left(\mathcal{L}\left(\cdot, \lambda^{(t)}\right)\right)$                // Oracle optimization
        Let $\Delta_\lambda^{(t)}$ be a gradient of $\mathcal{L}\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. $\lambda$
        Update $\lambda^{(t+1)} = \Pi_\Lambda\left(\lambda^{(t)} + \eta_\lambda \Delta_\lambda^{(t)}\right)$          // Projected gradient update
    Return $\theta^{(1)}, \dots, \theta^{(T)}$ and $\lambda^{(1)}, \dots, \lambda^{(T)}$
```

equilibria, in that uniformly sampling from the sequences $\theta^{(t)}$ and $\lambda^{(t)}$ can be interpreted as defining distributions over Θ and Λ. A convexity assumption would enable us to eliminate this added complexity by appealing to Jensen's inequality to replace these sequences with their averages. Second, for the technical reason that we require compact domains in order to prove convergence rates (below), Λ is taken to consist only of sets of Lagrange multipliers with bounded 1-norm[1].

As a consequence of this second point, the feasibility guarantee of Equation 4 only holds if the Lagrange multipliers are, on average, smaller than the maximum 1-norm radius R. Thankfully, as is shown by the final result of Theorem 3, if there exists a point satisfying the constraints with some margin γ, then there will exist Rs that are large enough to guarantee feasibility to within $O(\epsilon)$.

Our proposed algorithm (Algorithm 1) requires an *oracle* that performs approximate non-convex minimization, similarly to Chen et al. (2017)'s algorithm for robust optimization and Agarwal et al. (2018)'s for fair classification (the latter reference uses the terminology "best response"):

Definition 4 *A ρ-approximate Bayesian optimization oracle is a function $\mathcal{O}_\rho : (\Theta \to \mathbb{R}) \to \Theta$ for which:*

$$f\left(\mathcal{O}_\rho(f)\right) \leq \inf_{\theta^* \in \Theta} f\left(\theta^*\right) + \rho$$

for any $f : \Theta \to \mathbb{R}$ that can be written as a nonnegative linear combination of the objective and constraint functions $g_0, g_1, \dots, g_m$.

The θ-player uses this oracle, and the λ-player uses projected gradient ascent. Notice that, unlike the oracle of Chen et al. (2017), which provides a multiplicative approximation, $\mathcal{O}_\rho$ provides an *additive* approximation. Algorithm 1's convergence rate is:

Lemma 5 *Suppose that Λ and R are as in Theorem 3, and define the upper bound $B_\Delta \geq \max_{t \in [T]} \left\|\Delta_\lambda^{(t)}\right\|_2$.*

If we run Algorithm 1 with the step size $\eta_\lambda := R / B_\Delta \sqrt{2T}$, then the result satisfies the conditions of Theorem 3 for:

$$\epsilon = \rho + R B_\Delta \sqrt{\frac{2}{T}}$$

[1] In Appendix A, this is generalized to p-norms.

where ρ is the error associated with the oracle $\mathcal{O}_\rho$.

Proof In Appendix C.3. ■

Combined with Theorem 3, we therefore have that if R is sufficiently large, then Algorithm 1 will converge to a distribution over Θ that is, in expectation, $O(\rho)$-far from being optimal and feasible at a $O(1/\sqrt{T})$ rate, where ρ is as in Definition 4.

3.1. Shrinking

Aside from the unrealistic oracle assumption (which will be partially addressed in Section 4), the main disadvantage of Algorithm 1 is that it results in a mixture of T models, which presumably would be far too many to use in practice. However, a classical result (e.g. Bohnenblust et al., 1950; Parthasarathy and Raghavan, 1975) gives that much smaller Nash equilibria exist:

Lemma 6 *If Θ is a compact Hausdorff space, Λ is compact, and the objective and constraint functions $g_0, g_1, \dots, g_m$ are continuous, then the Lagrangian game (Definition 1) has a mixed Nash equilibrium pair (θ, λ) where θ is a random variable supported on at most $m+1$ elements of Θ, and λ is non-random.*

Proof Follows from Theorem 15 in Appendix B. ■

Of course, the mere existence of such an equilibrium is insufficient—we need to be able to *find* it, and Algorithm 1 manifestly does not. Thankfully, we can re-formulate the problem of finding the optimal ϵ-feasible mixture of the $\theta^{(t)}$s as a linear program (LP) that can be solved to "shrink" the support set. We must first evaluate the objective and constraint functions for every $\theta^{(t)}$, yielding a T-dimensional vector of objective function values, and m such vectors of constraint function evaluations, which are then used to specify the LP:

Lemma 7 *Let $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)} \in \Theta$ be a sequence of T "candidate solutions" of Equation 1. Define $\boldsymbol{g_0}, \boldsymbol{g_i} \in \mathbb{R}^T$ such that $(\boldsymbol{g_0})_t = g_0\left(\theta^{(t)}\right)$ and $(\boldsymbol{g_i})_t = g_i\left(\theta^{(t)}\right)$ for $i \in [m]$, and consider the linear program:*

$$\begin{aligned} \min_{p \in \Delta^T} \; & \langle p, \boldsymbol{g_0} \rangle \\ \text{s.t. } & \forall i \in [m] \, . \, \langle p, \boldsymbol{g_i} \rangle \leq \epsilon \end{aligned}$$

where Δ^T is the T-dimensional simplex. Then every vertex p^* of the feasible region—in particular an optimal one—has at most $m^* + 1 \leq m + 1$ nonzero elements, where m^* is the number of active $\langle p^*, \boldsymbol{g_i} \rangle \leq \epsilon$ constraints.

Proof In Appendix B. ■

This result suggests a two-phase approach to optimization. In the first phase, we apply Algorithm 1, yielding a sequence of iterates for which the uniform distribution over the $\theta^{(t)}$s is approximately feasible and optimal. We then apply the procedure of Lemma 7 to find the *best* distribution over these iterates, which in particular is guaranteed to be no worse than the uniform distribution, and is supported on at most $m+1$ iterates. We'll expand upon this further in Section 5.

4. Proxy-Lagrangian Optimization

While the Lagrangian formulation can be used to solve constrained problems in the form of Equation 1, Algorithm 1 isn't actually implementable, due to its reliance on an oracle. If one wished to apply it in practice, one would need to replace the oracle with something else, and for large-scale machine learning problems, "something else" is overwhelmingly likely to be SGD (Robbins and Monro, 1951; Zinkevich, 2003) or another first-order stochastic algorithm (e.g. AdaGrad (Duchi et al., 2011) or ADAM (Kingma and Ba, 2014)).

This leads to the issue we raised in Section 1.2: for non-differentiable constraints like those in the fairness example of Equation 2, we cannot compute gradients, and therefore cannot use a first-order algorithm. "Fixing" this issue by replacing the constraints with differentiable surrogates introduces a new difficulty: solutions to the resulting problem will satisfy the *surrogate* constraints, rather than the *actual* constraints.

The proxy-Lagrangian formulation of Definition 2 sidesteps this issue by using a non-zero-sum two-player game. The λ-player chooses how much the θ-player should penalize the (differentiable) proxy constraints, but does so in such a way as to satisfy the *original* constraints. Unfortunately, since the proxy-Lagrangian game is non-zero-sum, we cannot expect to find a Nash equilibrium, at least not efficiently. However, the analogous result to Theorem 3 requires a *weaker* type of equilibrium: a joint distribution over Θ and Λ w.r.t. which the θ-player can only make a negligible improvement compared to the best constant strategy, and the λ-player compared to the best action-swapping strategy (this is a particular type of Φ-correlated equilibrium (Rakhlin et al., 2011)):

Theorem 8 *Define $\mathcal{M}$ as the set of all left-stochastic $(m+1) \times (m+1)$ matrices, $\Lambda := \Delta^{m+1}$ as the $(m+1)$-dimensional simplex, and assume that each $\tilde{g}_i$ upper bounds the corresponding g_i. Let $\theta^{(1)}, \dots, \theta^{(T)} \in \Theta$ and $\lambda^{(1)}, \dots, \lambda^{(T)} \in \Lambda$ be sequences satisfying:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right) - \inf_{\theta^* \in \Theta} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\theta\left(\theta^{*}, \lambda^{(t)}\right) \le \epsilon_\theta$$

$$\max_{M^* \in \mathcal{M}} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\lambda\left(\theta^{(t)}, M^*\lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right) \le \epsilon_\lambda$$

Define $\bar{\theta}$ as a random variable for which $\bar{\theta} = \theta^{(t)}$ with probability $\lambda_1^{(t)} / \sum_{s=1}^{T} \lambda_1^{(s)}$, and let $\bar{\lambda} := \left(\sum_{t=1}^{T} \lambda^{(t)}\right)/T$. Then $\bar{\theta}$ is nearly-optimal in expectation:

$$\mathbb{E}_{\bar{\theta}}\left[g_0\left(\bar{\theta}\right)\right] \le \inf_{\theta^* \in \Theta : \forall i . \tilde{g}_i(\theta^*) \le 0} g_0\left(\theta^*\right) + \frac{\epsilon_\theta + \epsilon_\lambda}{\bar{\lambda}_1} \tag{5}$$

and nearly-feasible:

$$\max_{i \in [m]} \mathbb{E}_{\bar{\theta}}\left[g_i\left(\bar{\theta}\right)\right] \le \frac{\epsilon_\lambda}{\bar{\lambda}_1} \tag{6}$$

Additionally, if there exists a $\theta' \in \Theta$ that satisfies all of the proxy constraints with margin γ (i.e. $\tilde{g}_i\left(\theta'\right) \le -\gamma$ for all $i \in [m]$), then:

$$\bar{\lambda}_1 \ge \frac{\gamma - \epsilon_\theta - \epsilon_\lambda}{\gamma + B_{g_0}}$$

where $B_{g_0} \ge \sup_{\theta \in \Theta} g_0\left(\theta\right) - \inf_{\theta \in \Theta} g_0\left(\theta\right)$ is a bound on the range of the objective function g_0.

Algorithm 2 Optimizes the proxy-Lagrangian formulation (Definition 2) in the convex setting, with the θ-player minimizing external regret, and the λ-player minimizing swap regret. The $\operatorname{fix} M$ operation on line 3 results in a stationary distribution of M (i.e. a $\lambda \in \Lambda$ such that $M\lambda = \lambda$, which can be derived from the top eigenvector). The function Π_Θ projects its argument onto Θ w.r.t. the Euclidean norm.

```
  StochasticProxyLagrangian $\left(\mathcal{L}_\theta, \mathcal{L}_\lambda : \Theta \times \Delta^{m+1} \rightarrow \mathbb{R}, T \in \mathbb{N}, \eta_\theta, \eta_\lambda \in \mathbb{R}_+\right)$:
1     Initialize $\theta^{(1)} = 0$, and $M^{(1)} \in \mathbb{R}^{(m+1)\times(m+1)}$ with $M_{i,j} = 1/(m+1)$   // Assumes $0 \in \Theta$
2     For $t \in [T]$:
3         Let $\lambda^{(t)} = \operatorname{fix} M^{(t)}$   // Stationary distribution of $M^{(t)}$
4         Let $\check{\Delta}_\theta^{(t)}$ be a stochastic subgradient of $\mathcal{L}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. $\theta$
5         Let $\Delta_\lambda^{(t)}$ be a stochastic gradient of $\mathcal{L}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. $\lambda$
6         Update $\theta^{(t+1)} = \Pi_\Theta\left(\theta^{(t)} - \eta_\theta \check{\Delta}_\theta^{(t)}\right)$   // Projected SGD update
7         Update $\tilde{M}^{(t+1)} = M^{(t)} \odot . \exp\left(\eta_\lambda \Delta_\lambda^{(t)} \left(\lambda^{(t)}\right)^T\right)$   // $\odot$ and $.\exp$ are element-wise
8         Project $M_{:,i}^{(t+1)} = \tilde{M}_{:,i}^{(t+1)} / \left\|\tilde{M}_{:,i}^{(t+1)}\right\|_1$ for $i \in [m+1]$   // Column-wise projection
9     Return $\theta^{(1)}, \ldots, \theta^{(T)}$ and $\lambda^{(1)}, \ldots, \lambda^{(T)}$
```

Proof This is a special case of Theorem 13 and Lemma 14 in Appendix A. ■

Notice that while Equation 6 guarantees feasibility w.r.t. the original constraints, the comparator in Equation 5 is feasible w.r.t. the *proxy* constraints. Hence, the overall guarantee is no better than what we would achieve if we took $g_i := \tilde{g}_i$ for all $i \in [m]$, and optimized the Lagrangian as in Section 3. However, as will be demonstrated experimentally in Section 6.2, because the feasible region w.r.t. the original constraints is larger (perhaps significantly so) than that w.r.t. the proxy constraints, the proxy-Lagrangian approach—which finds a solution that is feasible w.r.t. this larger region—has more "room" to find a better solution in practice.

One key difference between this result and Theorem 3 is that the R parameter is absent. Instead, its role, and that of $\left\|\bar{\lambda}\right\|_1$, is played by the first coordinate of $\bar{\lambda}$. Inspection of Definition 2 reveals that, if one or more of the constraints are violated, then the λ-player would prefer λ_1 to be zero, whereas if they are satisfied (with some margin), then it would prefer λ_1 to be one. In other words, the first coordinate of $\lambda^{(t)}$ encodes the λ-player's belief about the feasibility of $\theta^{(t)}$, for which reason $\theta^{(t)}$ is weighted by $\lambda_1^{(t)}$ in the density defining $\bar{\theta}$.

Algorithm 2 is motivated by the observation that, while Theorem 8 only requires that the $\theta^{(t)}$ sequence suffer low external regret w.r.t. $\mathcal{L}_\theta\left(\cdot, \lambda^{(t)}\right)$, the condition on the $\lambda^{(t)}$ sequence is stronger, requiring it to suffer low *swap regret* (Blum and Mansour, 2007). Hence, the θ-player uses SGD to minimize external regret, while the λ-player uses a swap-regret minimization algorithm of the type proposed by Gordon et al. (2008), yielding the convergence guarantee:

Lemma 9 *Suppose that Θ is a compact convex set, $\mathcal{M}$ and Λ are as in Theorem 8, and that the objective and proxy constraint functions $g_0, \tilde{g}_1, \ldots, \tilde{g}_m$ are convex (but not $g_1, \ldots, g_m$). Define the three upper bounds $B_\Theta \geq \max_{\theta \in \Theta} \|\theta\|_2$, $B_{\check{\Delta}} \geq \max_{t \in [T]} \left\|\check{\Delta}_\theta^{(t)}\right\|_2$, and $B_\Delta \geq \max_{t \in [T]} \left\|\Delta_\lambda^{(t)}\right\|_\infty$.*

If we run Algorithm 2 with the step sizes $\eta_\theta := B_\Theta / B_{\check{\Delta}} \sqrt{2T}$ *and* $\eta_\lambda := \sqrt{(m+1)\ln(m+1)/TB_{\Delta}^2}$, *then the result satisfies the conditions of Theorem 8 for:*

$$\begin{aligned}\epsilon_\theta =& 2B_\Theta B_{\check{\Delta}} \sqrt{\frac{1 + 16\ln\frac{2}{\delta}}{T}} \\ \epsilon_\lambda =& 2B_\Delta \sqrt{\frac{2\left(m+1\right)\ln\left(m+1\right)\left(1+16\ln\frac{2}{\delta}\right)}{T}}\end{aligned}$$

with probability $1-\delta$ *over the draws of the stochastic (sub)gradients.*

Proof In Appendix C.3. ■

Algorithm 2 is designed for the convex setting (except for the g_is), for which reason it uses SGD for the θ-updates. However, this convexity requirement is not innate to our approach: it's straightforward to design an oracle-based algorithm that, like Algorithm 1, doesn't require convexity[2]. Our reason for presenting the SGD-based algorithm, instead of the oracle-based one, is that the purpose of proxy constraints is to substitute optimizable constraints for unoptimizable ones, and there is no need to do so if you have an oracle.

4.1. Shrinking

It turns out that the same existence result that we provided for the Lagrangian game (Lemma 6)—of a *Nash* equilibrium—holds for the proxy-Lagrangian:

Lemma 10 *If* Θ *is a compact Hausdorff space and the objective, constraint and proxy constraint functions* $g_0, g_1, \dots, g_m, \tilde{g}_1, \dots, \tilde{g}_m$ *are continuous, then the proxy-Lagrangian game (Definition 2) has a mixed Nash equilibrium pair* (θ, λ) *where* θ *is a random variable supported on at most* $m+1$ *elements of* Θ*, and* λ *is non-random.*

Proof In Appendix B. ■

Furthermore, the exact same linear programming procedure of Lemma 7 can be applied (with the $\boldsymbol{g_i}$s being defined in terms of the *original*—not proxy—constraints) to yield a solution with support size $m+1$, and works equally well. This is easy to verify: since $\bar{\theta}$, as defined in Theorem 8, is a distribution over the $\theta^{(t)}$s, and is therefore feasible for the LP, the *best* distribution over the iterates will be at least as good.

5. Overall Procedure

The pieces are now in place to propose a complete two-phase optimization procedure, for both convex and non-convex problems, with or without proxy constraints. In the first phase, we apply the appropriate algorithm to yield a distribution over the T "candidates" $\theta^{(1)}, \dots, \theta^{(T)}$ that is approximately feasible and optimal, according to either Theorems 3 or 8. Then, in the second phase, we construct $\boldsymbol{g_0}, \boldsymbol{g_1}, \dots, \boldsymbol{g_m} \in \mathbb{R}^T$ by evaluating the objective and constraint functions for each $\theta^{(t)}$,

[2] This is Algorithm 4, with Lemma 25 being its convergence guarantee, both in Appendix C.3.

	Testing				Training			
	Set 1	Set 2	Set 3	Set 4	Set 1	Set 2	Set 3	Set 4
Baseline ($\bar{\theta}$)	2.58	2.66	2.01	2.52	1.06	1.45	0.43	1.10
Baseline ($\theta^{(T)}$)	1.77	1.92	1.77	1.75	0.04	0.40	0.01	0.08
Lagrangian ($\bar{\theta}$)	2.04	2.15	1.96	2.04	0.42	0.70	0.30	0.49
Lagrangian (LP)	1.66	1.67	1.63	1.62	0.00	0.01	0.00	0.00

Table 1: Error rates on the experiments of Section 6.1. The columns correspond to the four corrupted datasets of Chen et al. (2017).

and then optimize the LP of Lemma 7 to find the *best* distribution over $\theta^{(1)}, \dots, \theta^{(T)}$ (which will have support size $\leq m+1$). If we take the ϵ parameter to this LP to be either the RHS of Equation 4 in Theorem 3 (for the Lagrangian case), or of Equation 6 in Theorem 8 (for the proxy-Lagrangian case), then the resulting size-$(m+1)$ distribution will have the same guarantees as the original.

Practical Procedure: The approach outlined above provably works, but is still somewhat idealized. In practice, we'll dispense with the oracle $\mathcal{O}_\rho$—even on non-convex problems—in favor of the "typical" approach: pretending that the problem is convex, and using SGD (or another cheap stochastic algorithm) for the θ-updates[3]. On a non-convex problem, this has no guarantees, but one would still hope that it would result in a "candidate set" of $\theta^{(t)}$s that contains enough good solutions to pass on to the LP of Lemma 7. If necessary, this candidate set can first be subsampled to make it a reasonable size. To choose the ϵ parameter of the LP, one can use a bisection search to find the smallest $\epsilon \geq 0$ for which there exists a feasible solution.

Evaluation: The ultimate result of either of these procedures is a distribution over at most $m+1$ distinct θs. If the underlying problem is one of classification, with $f(\cdot;\theta)$ being the scoring function, then this distribution defines a stochastic classifier: at evaluation time, upon receiving an example x, we would sample θ, and then return $f(x;\theta)$. If a stochastic classifier is not acceptable (as is often the case in real-world applications), then one could heuristically convert it into a deterministic one, e.g. by weighted averaging or voting, which is made significantly easier by its small size.

6. Experiments

We present two experiments: the first, on the robust MNIST problem of Chen et al. (2017), tests the performance of the "practical procedure" of Section 5 using the Lagrangian formulation (with the norms of the Lagrange multipliers being unbounded, i.e. $R = \infty$), while the second, a fairness problem on the UCI Adult dataset (Dheeru and Karra Taniskidou, 2017), uses the proxy-Lagrangian formulation. Both were implemented in TensorFlow[4].

In both cases, the θ and λ-updates both used AdaGrad with the same initial learning rates. In the proxy-Lagrangian case, however, the λ-update (line 7 of Algorithm 2) was performed in the log domain so that it would be multiplicative. To choose the initial AdaGrad learning rate, we performed

[3] In the Lagrangian case, this is Algorithm 3, with Lemma 24 being its convergence guarantee in the convex setting, both in Appendix C.3. In the proxy-Lagrangian case, this is Algorithm 2.

[4] `https://github.com/google-research/tensorflow_constrained_optimization`.

Algorithm	Support	Dataset	Error	Female	Male	Black	White
Baseline ($\theta^{(T)}$)	1	Train		**88.9%**	102%	**82.8%**	101%
		Test	14.2%	**89.5%**	102%	**81.6%**	101%
Lagrangian ($\bar{\theta}$)	100	Train		113%	97.8%	121%	99.7%
		Test	16.3%	114%	97.5%	126%	99.8%
Lagrangian (LP)	3	Train		104%	99.4%	105%	101%
		Test	15.5%	106%	99.0%	111%	101%
Proxy ($\bar{\theta}$)	100	Train		**94.7%**	101%	**94.5%**	100%
		Test	14.4%	**94.1%**	101%	**94.9%**	100%
Proxy (LP)	3	Train		95.0%	101%	95.0%	100%
		Test	14.2%	**94.4%**	101%	**94.9%**	100%

Table 2: Support sizes, test error rates, and "equal opportunity" values for the experiments of Section 6.2. For the constraints, each reported number is the ratio of the positive prediction rate on positively-labeled members of the protected class, to the positive prediction rate on the set of all positively-labeled data. The constraints attempt to force this ratio to be at least 95%—quantities lower than this threshold violate the constraint, and are marked in **bold**.

a grid search over powers-of-two, and chose the best model on a validation set. In all experiments, the optimum was in the interior of the grid.

Our constrained optimization algorithms result in stochastic classifiers, and we report results for *both* the $\bar{\theta}$ of Theorems 3 or 8 (in the Lagrangian or proxy-Lagrangian cases, respectively), *and* the optimal distribution found by the LP of Lemma 7, optimized on the training dataset.

6.1. Robust Optimization

In robust optimization, there are multiple objective functions $g_1, ..., g_m : \Theta \to \mathbb{R}$, and the goal is to find a $\theta \in \Theta$ minimizing $\max_{i \in [m]} g_i(\theta)$. As was discussed in Section 2, this can be re-written as a constrained problem by introducing a slack variable, as in Equation 3.

The task is the modified MNIST problem created by Chen et al. (2017), which is based on four datasets, each of which is a version of MNIST that has been corrupted in different ways. One would therefore hope that choosing g_i to be an empirical loss on the ith such dataset, and optimizing the corresponding robust problem, will result in a classifier that is "robust" to all four types of corruption.

We used a neural network with one 1024-neuron hidden layer, and ReLu activations. The four objective functions were the cross-entropy losses on the corrupted datasets. All models were trained for $50\,000$ iterations using a minibatch size of 100, and a $\theta^{(t)}$ was extracted every 500 iterations, yielding a sequence of length $T = 100$.

Baselines: For our baselines, we trained the neural network over the union of the four datasets. We report two variants: (i) the "Uniform Distribution Baseline" of Chen et al. (2017) is a stochastic classifier, uniformly sampled over the $\theta^{(t)}$s (like our $\bar{\theta}$ classifier), and (ii) a non-stochastic classifier taking its parameters from the last iterate $\theta^{(T)}$.

Results: Table 6 lists, for each of the corrupted datasets, the error rates of the compared models on both the training and testing datasets. Interestingly, although our proposed shrinking procedure is only guaranteed to give a distribution over $m + 1$ solutions, in these experiments it chose only one. Hence, the "Lagrangian (LP)" model of Table 6 is, like "Baseline ($\theta^{(T)}$)", non-stochastic.

While we did not quite match the raw performance reported by Chen et al. (2017)'s algorithm, our results, and theirs, tell similar stories. In particular, we can see that both of our algorithms outperformed their natural baseline equivalents. In particular, the use of shrinking not only greatly simplified the model, but also significantly improved performance.

6.2. Equal Opportunity

These experiments were performed on the UCI Adult dataset, which consists of census data including 14 features such as age, gender, race, occupation, and education. The goal was to predict whether income exceeds 50k/year. The dataset contains $32\,561$ training examples, from which we split off 20% to form a validation set, and $16\,281$ testing examples.

We dropped the "fnlwgt" weighting feature, and processed the remaining features as in Platt (1998), yielding 120 binary features, on which we trained linear models. The objective was to minimize the average hinge loss, subject to one 95% equal opportunity (Hardt et al., 2016) constraint in the style of Goh et al. (2016) for each "protected class": g_i was defined such that $g_i(\theta) \leq 0$ iff the positive prediction rate on the set of positively-labeled examples for the associated class was at least 95% of the positive prediction rate on the set of all positively-labeled examples.

When using proxy constraints, $\tilde{g}_i$ was taken to be a version of g_i with the indicator functions defining the positive prediction rates replaced with hinge upper bounds. When not using proxy constraints, the indicator-based constraints were dropped entirely, with these upper bounds being used throughout.

All models were trained for $5\,000$ iterations with a minibatch size of 100, with a $\theta^{(t)}$ being extracted every 50 iterations, yielding a sequence of length $T = 100$.

Baseline: The baseline classifier was optimized to simply minimize training hinge loss. Since this problem is unconstrained, we took the last iterate $\theta^{(T)}$.

"Best-model" Heuristic: For hyperparameter tuning using a grid search, we needed to choose the "best" model on the validation set. Due to the presence of constraints, however, the "best" model was not necessarily that with the lowest validation error. Instead, we used the following heuristic: the models were each ranked in terms of their objective function value, as well as the magnitude of the ith constraint violation (i.e. $\max\{0, g_i(\theta)\}$). The "score" of each model was then taken to be the maximal such rank, and the model with the lowest score was chosen, with the objective function serving as a tiebreaker.

Results: Table 2 lists the test error rates, (indicator-based) constraint function values on both the training and testing datasets, and support sizes of the stochastic classifiers, for each of the compared algorithms. The "LP" versions of our models, which were found using the shrinking procedure of Lemma 7, uniformly outperformed their $\bar{\theta}$-analogues. We can see, however, that the generalization issue discussed in Section 2.1 caused the proxy-Lagrangian LP model to slightly violate the constraints on the testing dataset, despite satisfying them on the training dataset. The non-proxy algorithms satisfied all constraints, on both the training and testing datasets, because there was sufficient “room” between the hinge upper bound that they actually constrained, and the true constraint, to absorb the generalization error. Inspection of the error rates, however, reveals that the relaxed constraints were so overly-conservative that satisfying them significantly damaged classification performance. In contrast, our proxy-Lagrangian approach matched the classification performance of the unconstrained baseline.

Acknowledgments

We thank Seungil You for initially posing the question of whether constraint functions could be relaxed for only the θ-player, as well as Maya Gupta, Taman Narayan and Serena Wang for helping to develop the heuristic used to choose the "best" model on the validation dataset in Section 6.2.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In *ICML'18*, pages 60–69, 2018.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, May 2003.

Dan Biddle. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower, 2005.

Avrim Blum and Yishay Mansour. From external to internal regret. *JMLR*, 8:1307–1324, 2007.

H. F. Bohnenblust, Samuel Karlin, and L. S. Shapley. Games with continuous, convex pay-off. *Contributions to the Theory of Games*, 1(24):181–192, 1950.

Robert S. Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In *Nips'17*, 2017.

Xi Chen and Xiaotie Deng. Settling the complexity of two-player nash equilibrium. In *FOCS'06*, pages 261–272. IEEE, 2006.

Paul Christiano, Jonathan A. Kelner, Aleksander Madry, Daniel A. Spielman, and Shang-Hua Teng. Electrical flows, Laplacian systems, and faster approximation of maximum flow in undirected graphs. In *STOC '11*, pages 273–282, 2011.

Andrew Cotter, Maya Gupta, and Jan Pfeifer. A Light Touch for heavily constrained SGD. In *29th Annual Conference on Learning Theory*, pages 729–771, 2016.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints, 2018. URL `https://arxiv.org/abs/1807.00028`.

Mark Davenport, Richard G. Baraniuk, and Clayton D. Scott. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12(Jul):2121–2159, 2011.

Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. Scalable learning of non-decomposable objectives. *AIStats'17*, 2017.

Dan Garber and Elad Hazan. Playing non-linear games with linear oracles. In *FOCS*, pages 420–428. IEEE Computer Society, 2013.

Gilles Gasso, Aristidis Pappaionannou, Marina Spivak, and Léon Bottou. Batch and online learning algorithms for nonconvex Neyman-Pearson classification. *ACM Transactions on Intelligent Systems and Technology*, 2011.

I. L. Glicksberg. A further generalization of the Kakutani fixed point theorem with application to Nash equilibrium points. *Amer. Math. Soc.*, 3:170–174, 1952.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *NIPS*, pages 2415–2423. 2016.

Geoffrey J. Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *ICML'08*, pages 360–367, 2008.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

Elad Hazan and Satyen Kale. Projection-free online learning. In *ICML'12*, 2012.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML'13*, volume 28, pages 427–435, 2013.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, 2017. URL `https://arxiv.org/abs/1711.05144`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR'14*, 2014.

Mehrdad Mahdavi, Tianbao Yang, Rong Jin, Shenghuo Zhu, and Jinfeng Yi. Stochastic gradient descent with only one projection. In *NIPS'12*, pages 494–502. 2012.

Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *AIStats*, 2018.

Arkadi Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons Ltd, 1983.

T. Parthasarathy and T. E. S. Raghavan. Equilibria of continuous two-person games. *Pacific Journal of Mathematics*, 57(1):265–270, 1975.

John C. Platt. Fast training of support vector machines using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS'13*, pages 3066–3074, 2013.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Beyond regret. In *COLT'11*, pages 559–594, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *NIPS'11*, 2011.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Matthew S. Vuolo and Norma B. Levy. Disparate impact doctrine in fair housing. *New York Law Journal*, 2013.

Wikipedia. Housing for older persons act — wikipedia, the free encyclopedia, 2018. URL `https://en.wikipedia.org/w/index.php?title=Housing_for_Older_Persons_Act&oldid=809132145`. [Online; accessed 6-February-2018].

Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *COLT'17*, pages 1920–1953, 2017.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML'03*, 2003.

Appendix A. Proofs of Sub{optimality,feasibility} Guarantees

Theorem 11 ***(Lagrangian Sub{optimality,feasibility})*** *Define $\Lambda = \left\{ \lambda \in \mathbb{R}_+^m : \left\| \lambda \right\|_p \leq R \right\}$, and consider the Lagrangian of Equation 1 (Definition 1). Suppose that $\theta \in \Theta$ and $\lambda \in \Lambda$ are random variables such that:*

$$\max_{\lambda^* \in \Lambda} \mathbb{E}_\theta \left[\mathcal{L}\left(\theta, \lambda^*\right) \right] - \inf_{\theta^* \in \Theta} \mathbb{E}_\lambda \left[\mathcal{L}\left(\theta^*, \lambda\right) \right] \leq \epsilon \tag{7}$$

i.e. θ, λ is an ϵ-approximate Nash equilibrium. Then θ is ϵ-suboptimal:

$$\mathbb{E}_\theta \left[g_0\left(\theta\right) \right] \leq \inf_{\theta^* \in \Theta : \forall i \in [m] . g_i(\theta^*) \leq 0} g_0\left(\theta^*\right) + \epsilon$$

Furthermore, if λ is in the interior of Λ, in the sense that $\left\| \bar{\lambda} \right\|_p < R$ where $\bar{\lambda} := \mathbb{E}_\lambda \left[\lambda \right]$, then θ is $\epsilon / \left(R - \left\| \bar{\lambda} \right\|_p \right)$-feasible:

$$\left\| \left(\mathbb{E}_\theta \left[g_:\left(\theta\right) \right] \right)_+ \right\|_q \leq \frac{\epsilon}{R - \left\| \bar{\lambda} \right\|_p}$$

where $g_:\left(\theta\right)$ is the m-dimensional vector of constraint evaluations, and $\left(\cdot\right)_+$ takes the positive part of its argument, so that $\left\| \left(\mathbb{E}_\theta \left[g_:\left(\theta\right) \right] \right)_+ \right\|_q$ is the q-norm of the vector of expected constraint violations.

Proof First notice that $\mathcal{L}$ is linear in λ, so:

$$\max_{\lambda^* \in \Lambda} \mathbb{E}_\theta \left[\mathcal{L}\left(\theta, \lambda^*\right) \right] - \inf_{\theta^* \in \Theta} \mathcal{L}\left(\theta^*, \bar{\lambda}\right) \leq \epsilon \tag{8}$$

Optimality: Choose θ^* to be the optimal *feasible* solution in Equation 8, so that $g_i\left(\theta^*\right) \leq 0$ for all $i \in [m]$, and also choose $\lambda^* = 0$, which combined with the definition of $\mathcal{L}$ (Definition 1) gives that:

$$\mathbb{E}_\theta \left[g_0\left(\theta\right) \right] - g_0\left(\theta^*\right) \leq \epsilon$$

which is the optimality claim.

Feasibility: Choose $\theta^* = \theta$ in Equation 8. By the definition of $\mathcal{L}$ (Definition 1):

$$\max_{\lambda^* \in \Lambda} \sum_{i=1}^m \lambda_i^* \mathbb{E}_\theta \left[g_i\left(\theta\right) \right] - \sum_{i=1}^m \bar{\lambda}_i \mathbb{E}_\theta \left[g_i\left(\theta\right) \right] \leq \epsilon$$

Then by the definition of a dual norm, Hölder's inequality, and the assumption that $\left\| \bar{\lambda} \right\|_p < R$:

$$R \left\| \left(\mathbb{E}_\theta \left[g_:\left(\theta\right) \right] \right)_+ \right\|_q - \left\| \bar{\lambda} \right\|_p \left\| \left(\mathbb{E}_\theta \left[g_:\left(\theta\right) \right] \right)_+ \right\|_q \leq \epsilon$$

Rearranging terms gives the feasibility claim. ■

Lemma 12 *In the context of Theorem 11, suppose that there exists a $\theta' \in \Theta$ that satisfies all of the constraints, and does so with q-norm margin γ, i.e. $g_i\left(\theta'\right) \leq 0$ for all $i \in [m]$ and $\left\| g_:\left(\theta'\right) \right\|_q \geq \gamma$. Then:*

$$\left\| \bar{\lambda} \right\|_p \leq \frac{\epsilon + B_{g_0}}{\gamma}$$

where $B_{g_0} \geq \sup_{\theta \in \Theta} g_0\left(\theta\right) - \inf_{\theta \in \Theta} g_0\left(\theta\right)$ is a bound on the range of the objective function g_0.

Proof Starting from Equation 7 (in Theorem 11), and choosing $\theta^* = \theta'$ and $\lambda^* = 0$:

$$\begin{aligned}
\epsilon \geq & \mathbb{E}_{\theta}\left[g_0\left(\theta\right)\right] - \mathbb{E}_{\lambda}\left[g_0\left(\theta'\right) + \sum_{i=1}^{m} \lambda_i g_i\left(\theta'\right)\right] \\
\epsilon \geq & \mathbb{E}_{\theta}\left[g_0\left(\theta\right) - \inf_{\theta' \in \Theta} g_0\left(\theta'\right)\right] - \left(g_0\left(\theta'\right) - \inf_{\theta' \in \Theta} g_0\left(\theta'\right)\right) + \gamma \left\|\bar{\lambda}\right\|_p \\
\epsilon \geq & - B_{g_0} + \gamma \left\|\bar{\lambda}\right\|_p
\end{aligned}$$

Solving for $\left\|\bar{\lambda}\right\|_p$ yields the claim. ∎

Theorem 13 ***(Proxy-Lagrangian Sub{optimality,feasibility})*** *Let $\mathcal{M}$ be the set of all left-stochastic $(m+1) \times (m+1)$ matrices (i.e. $\mathcal{M} := \left\{M \in \mathbb{R}^{(m+1)\times(m+1)} : \forall i \in [m+1] . M_{:,i} \in \Delta^{m+1}\right\}$), and consider the "proxy-Lagrangians" of Equation 1 (Definition 2). Suppose that $\theta \in \Theta$ and $\lambda \in \Lambda$ are jointly distributed random variables such that:*

$$\begin{aligned}
\mathbb{E}_{\theta,\lambda}\left[\mathcal{L}_{\theta}\left(\theta, \lambda\right)\right] - \inf_{\theta^* \in \Theta} \mathbb{E}_{\lambda}\left[\mathcal{L}_{\theta}\left(\theta^*, \lambda\right)\right] \leq & \epsilon_{\theta} \\
\max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta,\lambda}\left[\mathcal{L}_{\lambda}\left(\theta, M^*\lambda\right)\right] - \mathbb{E}_{\theta,\lambda}\left[\mathcal{L}_{\lambda}\left(\theta, \lambda\right)\right] \leq & \epsilon_{\lambda}
\end{aligned} \tag{9}$$

Define $\bar{\lambda} := \mathbb{E}_{\lambda}\left[\lambda\right]$, let $(\Omega, \mathcal{F}, P)$ be the probability space, and define a random variable $\bar{\theta}$ such that:

$$\Pr\left\{\bar{\theta} \in S\right\} = \frac{\int_{\theta^{-1}(S)} \lambda_1\left(x\right) dP\left(x\right)}{\int_{\Omega} \lambda_1\left(x\right) dP\left(x\right)}$$

In words, $\bar{\theta}$ is a version of θ that has been resampled with λ_1 being treated as an importance weight. In particular $\mathbb{E}_{\bar{\theta}}\left[f\left(\bar{\theta}\right)\right] = \mathbb{E}_{\theta,\lambda}\left[\lambda_1 f\left(\theta\right)\right] / \bar{\lambda}_1$ for any $f : \Theta \to \mathbb{R}$. Then $\bar{\theta}$ is nearly-optimal:

$$\mathbb{E}_{\bar{\theta}}\left[g_0\left(\bar{\theta}\right)\right] \leq \inf_{\theta^* \in \Theta : \forall i \in [m] . \tilde{g}_i(\theta^*) \leq 0} g_0\left(\theta^*\right) + \frac{\epsilon_{\theta} + \epsilon_{\lambda}}{\bar{\lambda}_1}$$

and nearly-feasible:

$$\left\|\left(\mathbb{E}_{\bar{\theta}}\left[g_{:}\left(\bar{\theta}\right)\right]\right)_+\right\|_{\infty} \leq \frac{\epsilon_{\lambda}}{\bar{\lambda}_1}$$

Notice the optimality inequality is weaker than it may appear, since the comparator in this equation is not *the optimal solution w.r.t. the constraints g_i, but rather w.r.t. the* proxy *constraints $\tilde{g}_i$.*

Proof **Optimality:** If we choose M^* to be the matrix with its first row being all-one, and all other rows being all-zero, then $\mathcal{L}_{\lambda}\left(\theta, M^*\lambda\right) = 0$, which shows that the first term in the LHS of the second line of Equation 9 is nonnegative. Hence, $-\mathbb{E}_{\theta,\lambda}\left[\mathcal{L}_{\lambda}\left(\theta, \lambda\right)\right] \leq \epsilon_{\lambda}$, so by the definition of $\mathcal{L}_{\lambda}$ (Definition 2), and the fact that $\tilde{g}_i \geq g_i$:

$$\mathbb{E}_{\theta,\lambda}\left[\sum_{i=1}^{m} \lambda_{i+1} \tilde{g}_i\left(\theta\right)\right] \geq -\epsilon_{\lambda}$$

Notice that $\mathcal{L}_{\theta}$ is linear in λ, so the first line of Equation 9, combined with the above result and the definition of $\mathcal{L}_{\theta}$ (Definition 2) becomes:

$$\mathbb{E}_{\theta,\lambda}\left[\lambda_1 g_0\left(\theta\right)\right] - \inf_{\theta^* \in \Theta}\left(\bar{\lambda}_1 g_0\left(\theta^*\right) + \sum_{i=1}^{m} \bar{\lambda}_{i+1} \tilde{g}_i\left(\theta^*\right)\right) \leq \epsilon_{\theta} + \epsilon_{\lambda} \tag{10}$$

Choose θ^* to be the optimal solution that satisfies the *proxy* constraints $\tilde{g}$, so that $\tilde{g}_i\left(\theta^*\right) \leq 0$ for all $i \in [m]$. Then:

$$\mathbb{E}_{\theta,\lambda}\left[\lambda_1 g_0\left(\theta\right)\right] - \bar{\lambda}_1 g_0\left(\theta^*\right) \leq \epsilon_\theta + \epsilon_\lambda$$

which is the optimality claim.

Feasibility: We'll simplify our notation by defining $\ell_1\left(\theta\right) := 0$ and $\ell_{i+1}\left(\theta\right) := g_i\left(\theta\right)$ for $i \in [m]$, so that $\mathcal{L}_\lambda\left(\theta, \lambda\right) = \left\langle \lambda, \ell_:\left(\theta\right)\right\rangle$. Consider the first term in the LHS of the second line of Equation 9:

$$\begin{aligned}
\max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta,\lambda}\left[\mathcal{L}_\lambda\left(\theta, M^*\lambda\right)\right] =& \max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta,\lambda}\left[\left\langle M^*\lambda, \ell_:\left(\theta\right)\right\rangle\right] \\
=& \max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta,\lambda}\left[\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} M^*_{j,i}\lambda_i \ell_j\left(\theta\right)\right] \\
=& \sum_{i=1}^{m+1} \max_{M^*_{:,i} \in \Delta^{m+1}} \sum_{j=1}^{m+1} \mathbb{E}_{\theta,\lambda}\left[M^*_{j,i}\lambda_i \ell_j\left(\theta\right)\right] \\
=& \sum_{i=1}^{m+1} \max_{j \in [m+1]} \mathbb{E}_{\theta,\lambda}\left[\lambda_i \ell_j\left(\theta\right)\right]
\end{aligned}$$

where we used the fact that, since M^* is left-stochastic, each of its columns is a $(m+1)$-dimensional multinoulli distribution. For the second term in the LHS of the second line of Equation 9, we can use the fact that $\ell_1\left(\theta\right) = 0$:

$$\mathbb{E}_{\theta,\lambda}\left[\sum_{i=2}^{m+1} \lambda_i \ell_i\left(\theta\right)\right] \leq \sum_{i=2}^{m+1} \max_{j \in [m+1]} \mathbb{E}_{\theta,\lambda}\left[\lambda_i \ell_j\left(\theta\right)\right]$$

Plugging these two results into the second line of Equation 9, the two sums collapse, leaving:

$$\max_{i \in [m+1]} \mathbb{E}_{\theta,\lambda}\left[\lambda_1 \ell_i\left(\theta\right)\right] \leq \epsilon_\lambda$$

The definition of ℓ_i then yields the feasibility claim. ■

Lemma 14 *In the context of Theorem 13, suppose that there exists a $\theta' \in \Theta$ that satisfies all of the* proxy *constraints with margin γ, i.e. $\tilde{g}_i\left(\theta'\right) \leq -\gamma$ for all $i \in [m]$. Then:*

$$\bar{\lambda}_1 \geq \frac{\gamma - \epsilon_\theta - \epsilon_\lambda}{\gamma + B_{g_0}}$$

where $B_{g_0} \geq \sup_{\theta \in \Theta} g_0\left(\theta\right) - \inf_{\theta \in \Theta} g_0\left(\theta\right)$ is a bound on the range of the objective function g_0.

Proof Starting from Equation 10 (in the proof of Theorem 13), and choosing $\theta^* = \theta'$:

$$\mathbb{E}_{\theta,\lambda}\left[\lambda_1 g_0\left(\theta\right)\right] - \left(\bar{\lambda}_1 g_0\left(\theta'\right) + \sum_{i=1}^{m} \bar{\lambda}_{i+1}\tilde{g}_i\left(\theta'\right)\right) \leq \epsilon_\theta + \epsilon_\lambda$$

Since $\tilde{g}_i\left(\theta'\right) \leq -\gamma$ for all $i \in [m]$:

$$\begin{aligned}
\epsilon_\theta + \epsilon_\lambda \geq & \mathbb{E}_{\theta,\lambda}\left[\lambda_1 g_0\left(\theta\right)\right] - \bar{\lambda}_1 g_0\left(\theta'\right) + \left(1 - \bar{\lambda}_1\right)\gamma \\
\geq & \mathbb{E}_{\theta,\lambda}\left[\lambda_1\left(g_0\left(\theta\right) - \inf_{\theta' \in \Theta} g_0\left(\theta'\right)\right)\right] - \bar{\lambda}_1\left(g_0\left(\theta'\right) - \inf_{\theta' \in \Theta} g_0\left(\theta'\right)\right) + \left(1 - \bar{\lambda}_1\right)\gamma \\
\geq & -\bar{\lambda}_1 B_{g_0} + \left(1 - \bar{\lambda}_1\right)\gamma
\end{aligned}$$

Solving for $\bar{\lambda}_1$ yields the claim. ■

Appendix B. Proofs of Existence of Sparse Equilibria

Theorem 15 *Consider a two player game, played on the compact Hausdorff spaces Θ and $\Lambda \subseteq \mathbb{R}^m$. Imagine that the θ-player wishes to minimize $\mathcal{L}_\theta : \Theta \times \Lambda \to \mathbb{R}$, and the λ-player wishes to maximize $\mathcal{L}_\lambda : \Theta \times \Lambda \to \mathbb{R}$, with both of these functions being continuous in θ and linear in λ. Then there exists a Nash equilibrium θ, λ:*

$$\begin{aligned}
\mathbb{E}_\theta\left[\mathcal{L}_\theta\left(\theta, \lambda\right)\right] &= \min_{\theta^* \in \Theta} \mathcal{L}_\theta\left(\theta^*, \lambda\right) \\
\mathbb{E}_\theta\left[\mathcal{L}_\lambda\left(\theta, \lambda\right)\right] &= \max_{\lambda^* \in \Lambda} \mathbb{E}_\theta\left[\mathcal{L}_\lambda\left(\theta, \lambda^*\right)\right]
\end{aligned}$$

where θ is a random variable placing nonzero probability mass on at most $m+1$ *elements of Θ, and $\lambda \in \Lambda$ is non-random.*

Proof There are some extremely similar (and in some ways more general) results than this in the game theory literature (e.g. Bohnenblust et al., 1950; Parthasarathy and Raghavan, 1975), but for our particular (Lagrangian and proxy-Lagrangian) setting it's possible to provide a fairly straightforward proof.

To begin with, Glicksberg (1952) gives that there exists a mixed strategy in the form of two random variables $\tilde{\theta}$ and $\tilde{\lambda}$:

$$\begin{aligned}
\mathbb{E}_{\tilde{\theta},\tilde{\lambda}}\left[\mathcal{L}_\theta\left(\tilde{\theta}, \tilde{\lambda}\right)\right] &= \min_{\theta^* \in \Theta} \mathbb{E}_{\tilde{\lambda}}\left[\mathcal{L}_\theta\left(\theta^*, \tilde{\lambda}\right)\right] \\
\mathbb{E}_{\tilde{\theta},\tilde{\lambda}}\left[\mathcal{L}_\lambda\left(\tilde{\theta}, \tilde{\lambda}\right)\right] &= \max_{\lambda^* \in \Lambda} \mathbb{E}_{\tilde{\theta}}\left[\mathcal{L}_\lambda\left(\tilde{\theta}, \lambda^*\right)\right]
\end{aligned}$$

Since both functions are linear in $\tilde{\lambda}$, we can define $\lambda := \mathbb{E}_{\tilde{\lambda}}\left[\tilde{\lambda}\right]$, and these conditions become:

$$\begin{aligned}
\mathbb{E}_{\tilde{\theta}}\left[\mathcal{L}_\theta\left(\tilde{\theta}, \lambda\right)\right] &= \min_{\theta^* \in \Theta} \mathcal{L}_\theta\left(\theta^*, \lambda\right) := \ell_{\min} \\
\mathbb{E}_{\tilde{\theta}}\left[\mathcal{L}_\lambda\left(\tilde{\theta}, \lambda\right)\right] &= \max_{\lambda^* \in \Lambda} \mathbb{E}_{\tilde{\theta}}\left[\mathcal{L}_\lambda\left(\tilde{\theta}, \lambda^*\right)\right]
\end{aligned}$$

Let's focus on the first condition. Let $p_\epsilon := \Pr\left\{\mathcal{L}_\theta\left(\tilde{\theta}, \lambda\right) \geq \ell_{\min} + \epsilon\right\}$, and notice that $p_{1/n}$ must equal zero for any $n \in \{1, 2, \dots\}$ (otherwise we would contradict the above), implying by the countable additivity of measures that $\Pr\left\{\mathcal{L}_\theta\left(\tilde{\theta}, \lambda\right) = \ell_{\min}\right\} = 1$. We therefore assume henceforth, without loss of generality, that the support of $\tilde{\theta}$ consists entirely of minimizers of $\mathcal{L}_\theta\left(\cdot, \lambda\right)$. Let $S \subseteq \Theta$ be this support set.

Define $G := \left\{\nabla_{\tilde{\lambda}} \mathcal{L}_\lambda\left(\theta', \lambda\right) : \theta' \in S\right\}$, and take $\bar{G}$ to be the closure of the convex hull of G. Since $\mathbb{E}_{\tilde{\theta}}\left[\nabla_{\tilde{\lambda}} \mathcal{L}_\lambda\left(\tilde{\theta}, \lambda\right)\right] \in \bar{G} \subseteq \mathbb{R}^m$, we can write it as a convex combination of at most $m+1$ extreme points of $\bar{G}$, or equivalently of $m+1$ elements of G. Hence, we can take θ to be a discrete random variable that places nonzero mass on at most $m+1$ elements of S, and:

$$\mathbb{E}_\theta\left[\nabla_{\tilde{\lambda}} \mathcal{L}_\lambda\left(\theta, \lambda\right)\right] = \mathbb{E}_{\tilde{\theta}}\left[\nabla_{\tilde{\lambda}} \mathcal{L}_\lambda\left(\tilde{\theta}, \lambda\right)\right]$$

Linearity in λ then implies that $\mathbb{E}_\theta\left[\mathcal{L}_\lambda\left(\theta, \cdot\right)\right]$ and $\mathbb{E}_{\tilde{\theta}}\left[\mathcal{L}_\lambda\left(\tilde{\theta}, \cdot\right)\right]$ are the *same function* (up to a constant), and therefore have the same maximizer(s). Correspondingly, θ is supported on S, which contains only minimizers of $\mathcal{L}_\theta\left(\cdot, \lambda\right)$ by construction. ■

Lemma 10 *If Θ is a compact Hausdorff space and the objective, constraint and proxy constraint functions $g_0, g_1, \dots, g_m, \tilde{g}_1, \dots, \tilde{g}_m$ are continuous, then the proxy-Lagrangian game (Definition 2) has a mixed Nash equilibrium pair (θ, λ) where θ is a random variable supported on at most $m+1$ elements of Θ, and λ is non-random.*

Proof Applying Theorem 15 directly would result in a support size of $m+2$, rather than the desired $m+1$, since Λ is $(m+1)$-dimensional. Instead, we define $\tilde{\Lambda} = \left\{\tilde{\lambda} \in \mathbb{R}_+^m : \left\|\tilde{\lambda}\right\|_1 \le 1\right\}$ as the space containing the last m coordinates of Λ. Then we can rewrite the proxy-Lagrangian functions $\tilde{\mathcal{L}}_\theta, \tilde{\mathcal{L}}_\lambda : \Theta \times \tilde{\Lambda} \to \mathbb{R}$ as:

$$\begin{aligned}
\tilde{\mathcal{L}}_\theta\left(\theta, \tilde{\lambda}\right) &= \left(1 - \left\|\tilde{\lambda}\right\|_1\right) g_0\left(\theta\right) + \sum_{i=1}^m \tilde{\lambda}_i \tilde{g}_i\left(\theta\right) \\
\tilde{\mathcal{L}}_\lambda\left(\theta, \tilde{\lambda}\right) &= \sum_{i=1}^m \tilde{\lambda}_i g_i\left(\theta\right)
\end{aligned}$$

These functions are linear in $\tilde{\lambda}$, which is a m-dimensional space, so the conditions of Theorem 15 apply, yielding the claimed result. ■

Lemma 7 *Let $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)} \in \Theta$ be a sequence of T "candidate solutions" of Equation 1. Define $\boldsymbol{g_0}, \boldsymbol{g_i} \in \mathbb{R}^T$ such that $\left(\boldsymbol{g_0}\right)_t = g_0\left(\theta^{(t)}\right)$ and $\left(\boldsymbol{g_i}\right)_t = g_i\left(\theta^{(t)}\right)$ for $i \in [m]$, and consider the linear program:*

$$\begin{aligned}
\min_{p \in \Delta^T} \; & \left\langle p, \boldsymbol{g_0} \right\rangle \\
\text{s.t. } & \forall i \in [m] \,.\, \left\langle p, \boldsymbol{g_i} \right\rangle \le \epsilon
\end{aligned}$$

where Δ^T is the T-dimensional simplex. Then every vertex p^* of the feasible region—in particular an optimal one—has at most $m^* + 1 \le m+1$ nonzero elements, where m^* is the number of active $\left\langle p^*, \boldsymbol{g_i} \right\rangle \le \epsilon$ constraints.

Proof The linear program contains not only the m explicit linearized functional constraints, but also, since $p \in \Delta^T$, the T nonnegativity constraints $p_t \geq 0$, and the sum-to-one constraint $\sum_{t=1}^{T} p_t = 1$.

Since p is T-dimensional, every vertex p^* of the feasible region must include T active constraints. Letting $m^* \leq m$ be the number of active linearized functional constraints, and accounting for the sum-to-one constraint, it follows that at least $T - m^* - 1$ nonnegativity constraints are active, implying that p^* contains at most $m^* + 1$ nonzero elements. ■

Appendix C. Proofs of Convergence Rates

C.1. Non-Stochastic One-Player Convergence Rates

Theorem 16 ***(Mirror Descent)*** *Let $f_1, f_2, \dots : \Theta \to \mathbb{R}$ be a sequence of convex functions that we wish to minimize on a compact convex set Θ. Suppose that the "distance generating function" $\Psi : \Theta \to \mathbb{R}_+$ is nonnegative and 1-strongly convex w.r.t. a norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$.*

Define the step size $\eta = \sqrt{B_\Psi / T B_{\check{\nabla}}^2}$, where $B_\Psi \geq \max_{\theta \in \Theta} \Psi(\theta)$ is a uniform upper bound on Ψ, and $B_{\check{\nabla}} \geq \left\| \check{\nabla} f_t \left(\theta^{(t)} \right) \right\|_*$ is a uniform upper bound on the norms of the subgradients. Suppose that we perform T iterations of the following update, starting from $\theta^{(1)} = \operatorname{argmin}_{\theta \in \Theta} \Psi(\theta)$:

$$\begin{aligned} \tilde{\theta}^{(t+1)} =& \nabla \Psi^* \left(\nabla \Psi \left(\theta^{(t)} \right) - \eta \check{\nabla} f_t \left(\theta^{(t)} \right) \right) \\ \theta^{(t+1)} =& \operatorname*{argmin}_{\theta \in \Theta} D_\Psi \left(\theta \mid \tilde{\theta}^{(t+1)} \right) \end{aligned}$$

where $\check{\nabla} f_t(\theta) \in \partial f_t(\theta^{(t)})$ is a subgradient of f_t at θ, and $D_\Psi(\theta \mid \theta') := \Psi(\theta) - \Psi(\theta') - \langle \nabla \Psi(\theta'), \theta - \theta' \rangle$ is the Bregman divergence associated with Ψ. Then:

$$\frac{1}{T} \sum_{t=1}^{T} f_t \left(\theta^{(t)} \right) - \frac{1}{T} \sum_{t=1}^{T} f_t(\theta^*) \leq 2 B_{\check{\nabla}} \sqrt{\frac{B_\Psi}{T}}$$

where $\theta^* \in \Theta$ is an arbitrary reference vector.

Proof Mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003) dates back to 1983, but this particular statement is taken from Lemma 2 of Srebro et al. (2011). ■

Corollary 17 ***(Gradient Descent)*** *Let $f_1, f_2, \dots : \Theta \to \mathbb{R}$ be a sequence of convex functions that we wish to minimize on a compact convex set Θ.*

Define the step size $\eta = B_\Theta / B_{\check{\nabla}} \sqrt{2T}$, where $B_\Theta \geq \max_{\theta \in \Theta} \|\theta\|_2$, and $B_{\check{\nabla}} \geq \left\| \check{\nabla} f_t \left(\theta^{(t)} \right) \right\|_2$ is a uniform upper bound on the norms of the subgradients. Suppose that we perform T iterations of the following update, starting from $\theta^{(1)} = \operatorname{argmin}_{\theta \in \Theta} \|\theta\|_2$:

$$\theta^{(t+1)} = \Pi_\Theta \left(\theta^{(t)} - \eta \check{\nabla} f_t \left(\theta^{(t)} \right) \right)$$

where $\check{\nabla} f_t(\theta) \in \partial f_t(\theta^{(t)})$ is a subgradient of f_t at θ, and Π_Θ projects its argument onto Θ w.r.t. the Euclidean norm. Then:

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(\theta^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} f_t\left(\theta^*\right) \le B_\Theta B_{\check{\nabla}}\sqrt{\frac{2}{T}}$$

where $\theta^* \in \Theta$ is an arbitrary reference vector.

Proof Follows from taking $\Psi(\theta) = \|\theta\|_2^2/2$ in Theorem 16. ∎

Corollary 18 *Let $\mathcal{M} := \left\{M \in \mathbb{R}^{\tilde{m}\times\tilde{m}} : \forall i \in [\tilde{m}] . M_{:,i} \in \Delta^{\tilde{m}}\right\}$ be the set of all left-stochastic $\tilde{m}\times\tilde{m}$ matrices, and let $f_1, f_2, \ldots : \mathcal{M} \to \mathbb{R}$ be a sequence of concave functions that we wish to maximize.*

Define the step size $\eta = \sqrt{\tilde{m}\ln\tilde{m}/TB_{\hat{\nabla}}^2}$, where $B_{\hat{\nabla}} \ge \left\|\hat{\nabla} f_t\left(M^{(t)}\right)\right\|_{\infty,2}$ is a uniform upper bound on the norms of the supergradients, and $\|\cdot\|_{\infty,2} := \sqrt{\sum_{i=1}^{\tilde{m}} \|M_{:,i}\|_\infty^2}$ is the $L_{\infty,2}$ matrix norm. Suppose that we perform T iterations of the following update starting from the matrix $M^{(1)}$ with all elements equal to $1/\tilde{m}$:

$$\begin{aligned}\tilde{M}^{(t+1)} =& M^{(t)} \odot . \exp\left(\eta\hat{\nabla} f_t\left(M^{(t)}\right)\right)\\ M_{:,i}^{(t+1)} =& \tilde{M}_{:,i}^{(t+1)} / \left\|\tilde{M}_{:,i}^{(t+1)}\right\|_1\end{aligned}$$

where $-\hat{\nabla} f_t\left(M^{(t)}\right) \in \partial\left(-f_t(M^{(t)})\right)$, i.e. $\hat{\nabla} f_t\left(M^{(t)}\right)$ is a supergradient of f_t at $M^{(t)}$, and the multiplication and exponentiation in the first step are performed element-wise. Then:

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(M^*\right) - \frac{1}{T}\sum_{t=1}^{T} f_t\left(M^{(t)}\right) \le 2B_{\hat{\nabla}}\sqrt{\frac{\tilde{m}\ln\tilde{m}}{T}}$$

where $M^* \in \mathcal{M}$ is an arbitrary reference matrix.

Proof Define $\Psi : \mathcal{M} \to \mathbb{R} := \tilde{m}\ln\tilde{m} + \sum_{i,j\in[\tilde{m}]} M_{i,j}\ln M_{i,j}$ as $\tilde{m}\ln\tilde{m}$ plus the negative Shannon entropy, applied to its (matrix) argument element-wise ($\tilde{m}\ln\tilde{m}$ is added to make Ψ nonnegative on $\mathcal{M}$). As in the vector setting, the resulting mirror descent update will be (element-wise) multiplicative.

The Bregman divergence satisfies:

$$\begin{aligned}D_\Psi\left(M|M'\right) =& \Psi(M) - \Psi\left(M'\right) - \left\langle\nabla\Psi\left(M'\right), M - M'\right\rangle\\ =& \left\|M'\right\|_{1,1} - \|M\|_{1,1} + \sum_{i=1}^{\tilde{m}} D_{KL}\left(M_{:,i}\|M'_{:,i}\right)\end{aligned} \tag{11}$$

where $\|M\|_{1,1} = \sum_{i=1}^{\tilde{m}}\|M_{:,i}\|_1$ is the $L_{1,1}$ matrix norm. This incidentally shows that one projects onto $\mathcal{M}$ w.r.t. D_Ψ by projecting each column w.r.t. the KL divergence, i.e. by normalizing the columns.

By Pinsker's inequality (applied to each column of an $M \in \mathcal{M}$):

$$\left\|M - M'\right\|_{1,2}^{2} \leq 2 \sum_{i=1}^{\tilde{m}} D_{KL}\left(M_{:,i} \| M'_{:,i}\right)$$

where $\left\|M\right\|_{1,2} = \sqrt{\sum_{i=1}^{\tilde{m}} \left\|M_{:,i}\right\|_{1}^{2}}$ is the $L_{1,2}$ matrix norm. Substituting this into Equation 11, and using the fact that $\left\|M\right\|_{1,1} = \tilde{m}$ for all $M \in \mathcal{M}$, we have that for all $M, M' \in \mathcal{M}$:

$$D_{\Psi}\left(M | M'\right) \geq \frac{1}{2}\left\|M - M'\right\|_{1,2}^{2}$$

which shows that Ψ is 1-strongly convex w.r.t. the $L_{1,2}$ matrix norm. The dual norm of the $L_{1,2}$ matrix norm is the $L_{\infty,2}$ norm, which is the last piece needed to apply Theorem 16, yielding the claimed result. ■

Lemma 19 *Let $\Lambda := \Delta^{\tilde{m}}$ be the $\tilde{m}$-dimensional simplex, define $\mathcal{M}$ as the set of all left-stochastic $\tilde{m} \times \tilde{m}$ matrices (i.e. $\mathcal{M} := \left\{M \in \mathbb{R}^{\tilde{m} \times \tilde{m}} : \forall i \in [\tilde{m}] . M_{:,i} \in \Delta^{\tilde{m}}\right\}$), and take $f_1, f_2, \ldots : \Lambda \to \mathbb{R}$ to be a sequence of concave functions that we wish to maximize.*

Define the step size $\eta = \sqrt{\tilde{m} \ln \tilde{m} / T B_{\hat{\nabla}}^{2}}$, where $B_{\hat{\nabla}} \geq \left\|\hat{\nabla} f_t\left(\lambda^{(t)}\right)\right\|_{\infty}$ is a uniform upper bound on the ∞-norms of the supergradients. Suppose that we perform T iterations of the following update, starting from the matrix $M^{(1)}$ with all elements equal to $1/\tilde{m}$:

$$\begin{aligned}
\lambda^{(t)} &= \operatorname{fix} M^{(t)} \\
A^{(t)} &= \left(\hat{\nabla} f_t\left(\lambda^{(t)}\right)\right)\left(\lambda^{(t)}\right)^{T} \\
\tilde{M}^{(t+1)} &= M^{(t)} \odot . \exp\left(\eta A^{(t)}\right) \\
M_{:,i}^{(t+1)} &= \tilde{M}_{:,i}^{(t+1)} / \left\|\tilde{M}_{:,i}^{(t+1)}\right\|_{1}
\end{aligned}$$

where $\operatorname{fix} M$ is a stationary distribution of M (i.e. a $\lambda \in \Lambda$ such that $M\lambda = \lambda$—such always exists, since M is left-stochastic), $-\hat{\nabla} f_t\left(\lambda^{(t)}\right) \in \partial\left(-f_t(\lambda^{(t)})\right)$, i.e. $\hat{\nabla} f_t\left(\lambda^{(t)}\right)$ is a supergradient of f_t at $\lambda^{(t)}$, and the multiplication and exponentiation of the third step are performed element-wise. Then:

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(M^{*}\lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} f_t\left(\lambda^{(t)}\right) \leq 2 B_{\hat{\nabla}} \sqrt{\frac{\tilde{m} \ln \tilde{m}}{T}}$$

where $M^{*} \in \mathcal{M}$ is an arbitrary left-stochastic reference matrix.

Proof This algorithm is an instance of that contained in Figure 1 of Gordon et al. (2008).

Define $\tilde{f}_t\left(M\right) := f_t\left(M^{(t)}\lambda^{(t)}\right)$. Observe that since $\hat{\nabla} f_t\left(\lambda^{(t)}\right)$ is a supergradient of f_t at $\lambda^{(t)}$, and $M^{(t)}\lambda^{(t)} = \lambda^{(t)}$:

$$\begin{aligned}
f_t\left(\tilde{M}\lambda^{(t)}\right) \leq & f_t\left(M^{(t)}\lambda^{(t)}\right) + \left\langle \hat{\nabla} f_t\left(\lambda^{(t)}\right), \tilde{M}\lambda^{(t)} - M^{(t)}\lambda^{(t)} \right\rangle \\
\leq & f_t\left(M^{(t)}\lambda^{(t)}\right) + A^{(t)} \cdot \left(\tilde{M} - M^{(t)}\right)
\end{aligned}$$

where the matrix product on the last line is performed element-wise. This shows that $A^{(t)}$ is a supergradient of $\tilde{f}_t$ at $M^{(t)}$, from which we conclude that the final two steps of the update are performing the algorithm of Corollary 18, so:

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{f}_t\left(M^{*}\right)-\frac{1}{T}\sum_{t=1}^{T}\tilde{f}_t\left(M^{(t)}\right)\leq 2B_{\hat{\nabla}}\sqrt{\frac{\tilde{m}\ln\tilde{m}}{T}}$$

where the $B_{\hat{\nabla}}$ of Corollary 18 is a uniform upper bound on the $L_{\infty,2}$ matrix norms of the $A^{(t)}$s. However, by the definition of $A^{(t)}$ and the fact that $\lambda^{(t)}\in\Delta^{\tilde{m}}$, we can instead take $B_{\hat{\nabla}}$ to be a uniform upper bound on $\left\|\hat{\nabla}^{(t)}\right\|_{\infty}$. Substituting the definition of $\tilde{f}_t$ and again using the fact that $M^{(t)}\lambda^{(t)}=\lambda^{(t)}$ then yields the claimed result. ■

C.2. Stochastic One-Player Convergence Rates

Theorem 20 ***(Stochastic Mirror Descent)*** *Let Ψ, $\|\cdot\|$, D_{Ψ} and B_{Ψ} be as in Theorem 16, and let $f_1,f_2,\ldots:\Theta\rightarrow\mathbb{R}$ be a sequence of convex functions that we wish to minimize on a compact convex set Θ.*

Define the step size $\eta=\sqrt{B_{\Psi}/TB_{\check{\Delta}}^{2}}$, where $B_{\check{\Delta}}\geq\left\|\check{\Delta}^{(t)}\right\|_{*}$ is a uniform upper bound on the norms of the stochastic subgradients. Suppose that we perform T iterations of the following stochastic *update, starting from $\theta^{(1)}=\operatorname{argmin}_{\theta\in\Theta}\Psi\left(\theta\right)$:*

$$\begin{aligned}\tilde{\theta}^{(t+1)}&=\nabla\Psi^{*}\left(\nabla\Psi\left(\theta^{(t)}\right)-\eta\check{\Delta}^{(t)}\right)\\ \theta^{(t+1)}&=\operatorname*{argmin}_{\theta\in\Theta}D_{\Psi}\left(\theta|\tilde{\theta}^{(t+1)}\right)\end{aligned}$$

where $\mathbb{E}\left[\check{\Delta}^{(t)}\mid\theta^{(t)}\right]\in\partial f_t(\theta^{(t)})$, i.e. $\check{\Delta}^{(t)}$ is a stochastic subgradient of f_t at $\theta^{(t)}$. Then, with probability $1-\delta$ over the draws of the stochastic subgradients:

$$\frac{1}{T}\sum_{t=1}^{T}f_t\left(\theta^{(t)}\right)-\frac{1}{T}\sum_{t=1}^{T}f_t\left(\theta^{*}\right)\leq 2B_{\check{\nabla}}\sqrt{\frac{2B_{\Psi}\left(1+16\ln\frac{1}{\delta}\right)}{T}}$$

where $\theta^{*}\in\Theta$ is an arbitrary reference vector.

Proof This is nothing more than the usual transformation of a uniform regret guarantee into a stochastic one via the Hoeffding-Azuma inequality—we include a proof for completeness.

Define the sequence:

$$\tilde{f}_t\left(\theta\right)=f_t\left(\theta^{(t)}\right)+\left\langle\check{\Delta}^{(t)},\theta-\theta^{(t)}\right\rangle$$

Then applying non-stochastic mirror descent to the sequence $\tilde{f}_t$ will result in exactly the same sequence of iterates $\theta^{(t)}$ as applying stochastic mirror descent (above) to f_t. Hence, by Theorem 16 and the definition of $\tilde{f}_t$ (notice that we can take $B_{\check{\nabla}} = B_{\check{\Delta}}$):

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\tilde{f}_t\left(\theta^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T}\tilde{f}_t\left(\theta^{*}\right) \leq & 2B_{\check{\nabla}}\sqrt{\frac{B_\Psi}{T}} \\
\frac{1}{T}\sum_{t=1}^{T}f_t\left(\theta^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T}f_t\left(\theta^{*}\right) \leq & 2B_{\check{\nabla}}\sqrt{\frac{B_\Psi}{T}} + \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{f}_t\left(\theta^{*}\right) - f_t\left(\theta^{*}\right)\right) \\
\leq & 2B_{\check{\nabla}}\sqrt{\frac{B_\Psi}{T}} + \frac{1}{T}\sum_{t=1}^{T}\left\langle \check{\Delta}^{(t)} - \check{\nabla} f_t\left(\theta^{(t)}\right), \theta^{*} - \theta^{(t)}\right\rangle \qquad (12)
\end{aligned}$$

where the last step follows from the convexity of the f_ts. Consider the second term on the RHS. Observe that, since the $\check{\Delta}^{(t)}$s are stochastic subgradients, each of the terms in the sum is zero in expectation (conditioned on the past), and the partial sums therefore form a martingale. Furthermore, by Hölder's inequality:

$$\left\langle \check{\Delta}^{(t)} - \check{\nabla} f_t\left(\theta^{(t)}\right), \theta^{*} - \theta^{(t)}\right\rangle \leq \left\|\check{\Delta}^{(t)} - \check{\nabla} f_t\left(\theta^{(t)}\right)\right\|_{*}\left\|\theta^{*} - \theta^{(t)}\right\| \leq 4B_{\check{\Delta}}\sqrt{2B_\Psi}$$

the last step because $\left\|\theta^{*} - \theta^{(t)}\right\| \leq \left\|\theta^{*} - \theta^{(1)}\right\| + \left\|\theta^{(t)} - \theta^{(1)}\right\| \leq 2\sup_{\theta\in\Theta}\sqrt{2D_\Psi\left(\theta \mid \theta^{(1)}\right)} \leq 2\sqrt{2B_\Psi}$, using the fact that D_Ψ is 1-strongly convex w.r.t. $\left\|\cdot\right\|$, and the definition of $\theta^{(1)}$. Hence, by the Hoeffding-Azuma inequality:

$$\Pr\left\{\frac{1}{T}\sum_{t=1}^{T}\left\langle \check{\Delta}^{(t)} - \check{\nabla} f_t\left(\theta^{(t)}\right), \theta^{*} - \theta^{(t)}\right\rangle \geq \epsilon\right\} \leq \exp\left(-\frac{T\epsilon^2}{64B_\Psi B_{\check{\Delta}}^2}\right)$$

equivalently:

$$\Pr\left\{\frac{1}{T}\sum_{t=1}^{T}\left\langle \check{\Delta}^{(t)} - \check{\nabla} f_t\left(\theta^{(t)}\right), \theta^{*} - \theta^{(t)}\right\rangle \geq 8B_{\check{\Delta}}\sqrt{\frac{B_\Psi \ln\frac{1}{\delta}}{T}}\right\} \leq \delta$$

substituting this into Equation 12, and applying the inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2a + 2b}$, yields the claimed result. ■

Corollary 21 ***(Stochastic Gradient Descent)*** *Let $f_1, f_2, \ldots : \Theta \to \mathbb{R}$ be a sequence of convex functions that we wish to minimize on a compact convex set Θ.*

Define the step size $\eta = B_\Theta / B_{\check{\Delta}}\sqrt{2T}$, where $B_\Theta \geq \max_{\theta\in\Theta}\left\|\theta\right\|_2$, and $B_{\check{\Delta}} \geq \left\|\check{\Delta}^{(t)}\right\|_2$ is a uniform upper bound on the norms of the stochastic subgradients. Suppose that we perform T iterations of the following stochastic *update, starting from $\theta^{(1)} = \operatorname{argmin}_{\theta\in\Theta}\left\|\theta\right\|_2$:*

$$\theta^{(t+1)} = \Pi_\Theta\left(\theta^{(t)} - \eta\check{\Delta}^{(t)}\right)$$

where $\mathbb{E}\left[\check{\Delta}^{(t)} \mid \theta^{(t)}\right] \in \partial f_t(\theta^{(t)})$, i.e. $\check{\Delta}^{(t)}$ is a stochastic subgradient of f_t at $\theta^{(t)}$, and Π_Θ projects its argument onto Θ w.r.t. the Euclidean norm. Then, with probability $1 - \delta$ over the draws of the *stochastic subgradients:*

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(\theta^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} f_t\left(\theta^{*}\right) \leq 2B_{\Theta}B_{\check{\nabla}}\sqrt{\frac{1+16\ln\frac{1}{\delta}}{T}}$$

where $\theta^{*} \in \Theta$ is an arbitrary reference vector.

Proof Follows from taking $\Psi\left(\theta\right) = \left\lVert\theta\right\rVert_2^2/2$ in Theorem 20. ■

Corollary 22 *Let $\mathcal{M} := \left\{M \in \mathbb{R}^{\tilde{m}\times\tilde{m}} : \forall i \in [\tilde{m}]\,.M_{:,i} \in \Delta^{\tilde{m}}\right\}$ be the set of all left-stochastic $\tilde{m}\times\tilde{m}$ matrices, and let $f_1, f_2, \ldots : \mathcal{M} \to \mathbb{R}$ be a sequence of concave functions that we wish to maximize.*

Define the step size $\eta = \sqrt{\tilde{m}\ln\tilde{m}/TB_{\hat{\Delta}}^2}$, where $B_{\hat{\Delta}} \geq \left\lVert\hat{\Delta}^{(t)}\right\rVert_{\infty,2}$ is a uniform upper bound on the norms of the stochastic supergradients, and $\left\lVert\cdot\right\rVert_{\infty,2} := \sqrt{\sum_{i=1}^{\tilde{m}}\left\lVert M_{:,i}\right\rVert_{\infty}^2}$ is the $L_{\infty,2}$ matrix norm. Suppose that we perform T iterations of the following stochastic update starting from the matrix $M^{(1)}$ with all elements equal to $1/\tilde{m}$:

$$\begin{aligned}\tilde{M}^{(t+1)} =& M^{(t)} \odot .\exp\left(\eta\hat{\Delta}^{(t)}\right)\\ M_{:,i}^{(t+1)} =& \tilde{M}_{:,i}^{(t+1)}/\left\lVert\tilde{M}_{:,i}^{(t+1)}\right\rVert_1\end{aligned}$$

where $\mathbb{E}\left[-\hat{\Delta}^{(t)} \mid M^{(t)}\right] \in \partial\left(-f_t(M^{(t)})\right)$, i.e. $\hat{\Delta}^{(t)}$ is a stochastic supergradient of f_t at $M^{(t)}$, and the multiplication and exponentiation in the first step are performed element-wise. Then with probability $1-\delta$ over the draws of the stochastic supergradients:

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(M^{*}\right) - \frac{1}{T}\sum_{t=1}^{T} f_t\left(M^{(t)}\right) \leq 2B_{\hat{\Delta}}\sqrt{\frac{2\left(\tilde{m}\ln\tilde{m}\right)\left(1+16\ln\frac{1}{\delta}\right)}{T}}$$

where $M^{*} \in \mathcal{M}$ is an arbitrary reference matrix.

Proof The same reasoning as was used to prove Corollary 18 from Theorem 16 applies here (but starting from Theorem 20). ■

Lemma 23 *Let $\Lambda := \Delta^{\tilde{m}}$ be the $\tilde{m}$-dimensional simplex, define $\mathcal{M}$ as the set of all left-stochastic $\tilde{m}\times\tilde{m}$ matrices (i.e. $\mathcal{M} := \left\{M \in \mathbb{R}^{\tilde{m}\times\tilde{m}} : \forall i \in [\tilde{m}]\,.M_{:,i} \in \Delta^{\tilde{m}}\right\}$), and take $f_1, f_2, \ldots : \Lambda \to \mathbb{R}$ to be a sequence of concave functions that we wish to maximize.*

Define the step size $\eta = \sqrt{\tilde{m}\ln\tilde{m}/TB_{\hat{\Delta}}^2}$, where $B_{\hat{\Delta}} \geq \left\lVert\hat{\Delta}^{(t)}\right\rVert_{\infty}$ is a uniform upper bound on the ∞-norms of the stochastic supergradients. Suppose that we perform T iterations of the following *update, starting from the matrix $M^{(1)}$ with all elements equal to $1/\tilde{m}$:*

$$\begin{aligned}\lambda^{(t)} =& \operatorname{fix} M^{(t)} \\ A^{(t)} =& \hat{\Delta}^{(t)} \left(\lambda^{(t)}\right)^T \\ \tilde{M}^{(t+1)} =& M^{(t)} \odot . \exp\left(\eta A^{(t)}\right) \\ M^{(t+1)}_{:,i} =& \tilde{M}^{(t+1)}_{:,i} / \left\| \tilde{M}^{(t+1)}_{:,i} \right\|_1\end{aligned}$$

where fix M *is a stationary distribution of M (i.e. a $\lambda \in \Lambda$ such that $M\lambda = \lambda$—such always exists, since M is left-stochastic), $\mathbb{E}\left[-\hat{\Delta}^{(t)} \mid \lambda^{(t)}\right] \in \partial\left(-f_t(\lambda^{(t)})\right)$, i.e. $\hat{\Delta}^{(t)}$ is a stochastic supergradient of f_t at $\lambda^{(t)}$, and the multiplication and exponentiation of the third step are performed element-wise. Then with probability $1-\delta$ over the draws of the stochastic supergradients:*

$$\frac{1}{T}\sum_{t=1}^{T} f_t\left(M^*\lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} f_t\left(\lambda^{(t)}\right) \le 2B_{\hat{\Delta}}\sqrt{\frac{2\left(\tilde{m}\ln\tilde{m}\right)\left(1+16\ln\frac{1}{\delta}\right)}{T}}$$

where $M^* \in \mathcal{M}$ is an arbitrary left-stochastic reference matrix.

Proof The same reasoning as was used to prove Lemma 19 from Corollary 18 applies here (but starting from Corollary 22). ■

C.3. Two-Player Convergence Rates

Lemma 5 ***(Algorithm 1)*** *Suppose that Λ and R are as in Theorem 3, and define the upper bound $B_\Delta \ge \max_{t\in[T]}\left\|\Delta_\lambda^{(t)}\right\|_2$.*

If we run Algorithm 1 with the step size $\eta_\lambda := R/B_\Delta\sqrt{2T}$, then the result satisfies the conditions of Theorem 3 for:

$$\epsilon = \rho + RB_\Delta\sqrt{\frac{2}{T}}$$

where ρ is the error associated with the oracle $\mathcal{O}_\rho$.

Proof Applying Corollary 17 to the optimization over λ gives:

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\theta^{(t)},\lambda^*\right) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\theta^{(t)},\lambda^{(t)}\right) \le B_\Lambda B_\Delta\sqrt{\frac{2}{T}}$$

By the definition of $\mathcal{O}_\rho$ (Definition 4):

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\theta^{(t)},\lambda^*\right) - \inf_{\theta^*\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\theta^*,\lambda^{(t)}\right) \le \rho + B_\Lambda B_\Delta\sqrt{\frac{2}{T}}$$

Using the linearity of $\mathcal{L}$ in λ, the fact that $B_\Lambda = R$, and the definitions of $\bar{\theta}$ and $\bar{\lambda}$, yields the claimed result. ■

Algorithm 3 Optimizes the Lagrangian formulation (Definition 1) in the convex setting. The parameter R is the radius of the Lagrange multiplier space $\Lambda := \left\{\lambda \in \mathbb{R}_+^m : \|\lambda\|_1 \leq R\right\}$, and the functions Π_Θ and Π_Λ project their arguments onto Θ and Λ (respectively) w.r.t. the Euclidean norm.

```
StochasticLagrangian ($R \in \mathbb{R}_+, \mathcal{L} : \Theta \times \Lambda \to \mathbb{R}, T \in \mathbb{N}, \eta_\theta, \eta_\lambda \in \mathbb{R}_+$):
    Initialize $\theta^{(1)} = 0$, $\lambda^{(1)} = 0$                                    // Assumes $0 \in \Theta$
    For $t \in [T]$:
        Let $\check{\Delta}_\theta^{(t)}$ be a stochastic subgradient of $\mathcal{L}\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. $\theta$
        Let $\Delta_\lambda^{(t)}$ be a stochastic gradient of $\mathcal{L}\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. $\lambda$
        Update $\theta^{(t+1)} = \Pi_\Theta\left(\theta^{(t)} - \eta_\theta \check{\Delta}_\theta^{(t)}\right)$              // Projected SGD updates ...
        Update $\lambda^{(t+1)} = \Pi_\Lambda\left(\lambda^{(t)} + \eta_\lambda \Delta_\lambda^{(t)}\right)$              //     ...
    Return $\theta^{(1)}, \dots, \theta^{(T)}$ and $\lambda^{(1)}, \dots, \lambda^{(T)}$
```

Lemma 24 ***(Algorithm 3)*** *Suppose that Θ is a compact convex set, Λ and R are as in Theorem 3, and that the objective and constraint functions $g_0, g_1, \dots, g_m$ are convex. Define the three upper bounds $B_\Theta \geq \max_{\theta \in \Theta} \|\theta\|_2$, $B_{\check{\Delta}} \geq \max_{t \in [T]} \left\|\check{\Delta}_\theta^{(t)}\right\|_2$, and $B_\Delta \geq \max_{t \in [T]} \left\|\Delta_\lambda^{(t)}\right\|_2$.*

If we run Algorithm 3 with the step sizes $\eta_\theta := B_\Theta / B_{\check{\Delta}} \sqrt{2T}$ and $\eta_\lambda := R / B_\Delta \sqrt{2T}$, then the result satisfies the conditions of Theorem 3 for:

$$\epsilon = 2\left(B_\Theta B_{\check{\Delta}} + R B_\Delta\right) \sqrt{\frac{1 + 16 \ln \frac{2}{\delta}}{T}}$$

with probability $1 - \delta$ over the draws of the stochastic (sub)gradients.

Proof Applying Corollary 21 to the two optimizations (over θ and λ) gives that with probability $1 - 2\delta'$ over the draws of the stochastic (sub)gradients:

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(\theta^{(t)}, \lambda^{(t)}\right) - \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(\theta^{*}, \lambda^{(t)}\right) \leq & 2 B_\Theta B_{\check{\Delta}} \sqrt{\frac{1 + 16 \ln \frac{1}{\delta'}}{T}} \\
\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(\theta^{(t)}, \lambda^{*}\right) - \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(\theta^{(t)}, \lambda^{(t)}\right) \leq & 2 B_\Lambda B_\Delta \sqrt{\frac{1 + 16 \ln \frac{1}{\delta'}}{T}}
\end{aligned}$$

Adding these inequalities, taking $\delta = 2\delta'$, using the linearity of $\mathcal{L}$ in λ, the fact that $B_\Lambda = R$, and the definitions of $\bar{\theta}$ and $\bar{\lambda}$, yields the claimed result. ■

Lemma 25 ***(Algorithm 4)*** *Suppose that $\mathcal{M}$ and Λ are as in Theorem 8, and define the upper bound $B_\Delta \geq \max_{t \in [T]} \left\|\Delta_\lambda^{(t)}\right\|_\infty$.*

Algorithm 4 Optimizes the proxy-Lagrangian formulation (Definition 2) in the non-convex setting via the use of an approximate Bayesian optimization oracle $\mathcal{O}_\rho$ (Definition 4, but with $\tilde{g}_i$s instead of g_is in the linear combination defining f) for the θ-player, with the λ-player minimizing swap regret. The fix M operation on line 3 results in a stationary distribution of M (i.e. a $\lambda \in \Lambda$ such that $M\lambda = \lambda$, which can be derived from the top eigenvector).

```
  OracleProxyLagrangian $\left(\mathcal{L}_\theta, \mathcal{L}_\lambda : \Theta \times \Delta^{m+1} \to \mathbb{R}, \mathcal{O}_\rho : (\Theta \to \mathbb{R}) \to \Theta, T \in \mathbb{N}, \eta_\lambda \in \mathbb{R}_+\right)$:
1     Initialize $M^{(1)} \in \mathbb{R}^{(m+1)\times(m+1)}$ with $M_{i,j} = 1/(m+1)$
2     For $t \in [T]$:
3         Let $\lambda^{(t)} = \operatorname{fix} M^{(t)}$                              // Stationary distribution of $M^{(t)}$
4         Let $\theta^{(t)} = \mathcal{O}_\rho\left(\mathcal{L}_\theta\left(\cdot, \lambda^{(t)}\right)\right)$                      // Oracle optimization
5         Let $\Delta_\lambda^{(t)}$ be a gradient of $\mathcal{L}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. $\lambda$
6         Update $\tilde{M}^{(t+1)} = M^{(t)} \odot .\exp\left(\eta_\lambda \Delta_\lambda^{(t)} \left(\lambda^{(t)}\right)^T\right)$      // $\odot$ and $.\exp$ are element-wise
7         Project $M_{:,i}^{(t+1)} = \tilde{M}_{:,i}^{(t+1)} / \left\|\tilde{M}_{:,i}^{(t+1)}\right\|_1$ for $i \in [m+1]$     // Column-wise projection
8     Return $\theta^{(1)}, \dots, \theta^{(T)}$ and $\lambda^{(1)}, \dots, \lambda^{(T)}$
```

If we run Algorithm 4 with the step size $\eta_\lambda := \sqrt{(m+1)\ln(m+1)/TB_\Delta^2}$*, then the result satisfies satisfies the conditions of Theorem 8 for:*

$$\begin{aligned}\epsilon_\theta =& \rho \\ \epsilon_\lambda =& 2B_\Delta \sqrt{\frac{(m+1)\ln(m+1)}{T}}\end{aligned}$$

where ρ *is the error associated with the oracle* $\mathcal{O}_\rho$*.*

Proof Applying Lemma 19 to the optimization over λ (with $\tilde{m} := m+1$) gives:

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\lambda\left(\theta^{(t)}, M^* \lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right) \le 2B_\Delta \sqrt{\frac{(m+1)\ln(m+1)}{T}}$$

By the definition of $\mathcal{O}_\rho$ (Definition 4):

$$\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right) - \inf_{\theta^* \in \Theta} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\theta\left(\theta^*, \lambda^{(t)}\right) \le \rho$$

Using the definitions of $\bar{\theta}$ and $\bar{\lambda}$ yields the claimed result. ■

Lemma 9 ***(Algorithm 2)*** *Suppose that* Θ *is a compact convex set,* $\mathcal{M}$ *and* Λ *are as in Theorem 8, and that the objective and proxy constraint functions* $g_0, \tilde{g}_1, \dots, \tilde{g}_m$ *are convex (but not* $g_1, \dots, g_m$*). Define the three upper bounds* $B_\Theta \ge \max_{\theta \in \Theta} \|\theta\|_2$*,* $B_{\check{\Delta}} \ge \max_{t \in [T]} \left\|\check{\Delta}_\theta^{(t)}\right\|_2$*, and* $B_\Delta \ge \max_{t \in [T]} \left\|\Delta_\lambda^{(t)}\right\|_\infty$.

If we run Algorithm 2 with the step sizes $\eta_\theta := B_\Theta / B_{\check{\Delta}} \sqrt{2T}$ *and* $\eta_\lambda := \sqrt{(m+1)\ln(m+1)/TB_{\check{\Delta}}^2}$, *then the result satisfies the conditions of Theorem 8 for:*

$$\begin{aligned}\epsilon_\theta =& 2B_\Theta B_{\check{\Delta}} \sqrt{\frac{1 + 16 \ln \frac{2}{\delta}}{T}} \\ \epsilon_\lambda =& 2B_\Delta \sqrt{\frac{2(m+1)\ln(m+1)\left(1 + 16\ln\frac{2}{\delta}\right)}{T}}\end{aligned}$$

with probability $1 - \delta$ *over the draws of the stochastic (sub)gradients.*

Proof Applying Corollary 21 to the optimization over θ, and Lemma 23 to that over λ (with $\tilde{m} := m + 1$), gives that with probability $1 - 2\delta'$ over the draws of the stochastic (sub)gradients:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\theta\left(\theta^{*}, \lambda^{(t)}\right) \le& 2B_\Theta B_{\check{\Delta}} \sqrt{\frac{1 + 16\ln\frac{1}{\delta'}}{T}} \\ \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\lambda\left(\theta^{(t)}, M^{*}\lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right) \le& 2B_\Delta \sqrt{\frac{2(m+1)\ln(m+1)\left(1 + 16\ln\frac{1}{\delta'}\right)}{T}}\end{aligned}$$

Taking $\delta = 2\delta'$, and using the definitions of $\bar{\theta}$ and $\bar{\lambda}$, yields the claimed result. ■