Table 5. Key notation.

Symbol	Description
m	Number of constraints in Equation 1
ℓ_0	Objective loss function in Equation 1
ℓ_i	ith constraint loss function in Equation 1
$\tilde{\ell}_i$	ith proxy-constraint loss function (corresponding to ℓ_i) in Definition 1
x	Feature vector for an example
$\mathcal{X}$	Space of feature vectors $x \in \mathcal{X}$
$\mathcal{D}$	Data distribution over feature vectors $x \in \mathcal{X}$
$S^{(\mathrm{trn})}$	An *i.i.d.* training sample from $\mathcal{D}$
$S^{(\mathrm{val})}$	An *i.i.d.* "validation" sample from $\mathcal{D}$
θ	Parameter vector defining a model
Θ	Space of parameter vectors ($\theta \in \Theta$)
$\hat{\Theta}$	Subset of Θ consisting of the sequence of algorithm outputs $\theta^{(1)}, \ldots, \theta^{(T)}$
$\bar{\theta}$	Random variable over $\hat{\Theta}$ defining a stochastic model (Theorems 1 and 2)
Δ^m	The m-probability-simplex, i.e. the set of all $p \in \mathbb{R}_+^{m+1}$ for which $\sum_{i=1}^{m+1} p_i = 1$
λ	Vector of Lagrange-multiplier-like "hyperparameters" (Definition 1)
Λ	Space of Lagrange-multiplier-like "hyperparameter" vectors ($\lambda \in \Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$)
$\bar{\lambda}$	Average "hyperparameters" $\bar{\lambda} := (\sum_{t=1}^{T} \lambda^{(t)})/T$ (Theorems 1 and 2)
$\bar{\lambda}_1$	First coordinate of $\bar{\lambda}$, measuring our "belief" that $\bar{\theta}$ is feasible
$\mathcal{L}_\theta$	In-expectation proxy-Lagrangian function minimized by θ-player (Definition 4)
$\mathcal{L}_\lambda$	In-expectation proxy-Lagrangian function maximized by λ-player (Definition 4)
$\hat{\mathcal{L}}_\theta$	Empirical proxy-Lagrangian function minimized by θ-player (Definition 1)
$\hat{\mathcal{L}}_\lambda$	Empirical proxy-Lagrangian function maximized by λ-player (Definition 1)
$\tilde{G}^{(\mathrm{trn})}(\Theta)$	Generalization bound for $\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m$ on $S^{(\mathrm{trn})}$ for $\theta \in \Theta$ (Definition 2)
$G^{(\mathrm{val})}(\hat{\Theta})$	Generalization bound for $\ell_1, \ldots, \ell_m$ on $S^{(\mathrm{val})}$ for $\theta \in \hat{\Theta}$ (Definition 2)
$\mathcal{O}_\rho$	Bayesian oracle of Definition 3
ρ	Additive approximation of the Bayesian oracle of Definition 3
C_r	A radius-r external covering of $\Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$ w.r.t. the 1-norm
r	Radius of the covering C_r
Θ_{C_r}	Oracle evaluations at the covering centers ($\Theta_{C_r} := \{\mathcal{O}_\rho(\hat{\mathcal{L}}_\theta(\cdot, \tilde{\lambda})) : \tilde{\lambda} \in C_r\}$)
T	Number of iterations performed in Algorithms 1, 3 and 4
T_λ	Number of iterations performed in the outer loop of Algorithm 2
T_θ	Number of iterations performed in the inner loop of Algorithm 2
η_λ	Step size associated with λ-player in Algorithms 1, 2, 3 and 4
η_θ	Step size associated with θ-player in Algorithms 3 and 4
M	A left-stochastic $(m+1) \times (m+1)$ matrix
$\mathcal{M}$	Space of all left-stochastic $(m+1) \times (m+1)$ matrices ($M \in \mathcal{M}$)
γ	Maximum margin by which the proxy-constraints can be satisfied in Theorems 1 and 2
μ	Strong convexity parameter of $\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m$ in Theorem 2
L	Lipschitz constant of $\ell_1, \ldots, \ell_m$ in Theorem 2
B_{ℓ_0}	Upper bound on $b_0 - a_0$, where $\mathrm{range}(\ell_0) = [a_0, b_0]$
B_ℓ	Upper bound on $b_i - a_i$ for all $i \in [m]$, where $\mathrm{range}(\ell_i) = [a_i, b_i]$
$B_{\tilde{\ell}}$	Upper bound on $\lvert\ell(x;\theta)\rvert$ for all $\ell \in \{\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m\}$
$B_{\hat{\Delta}}$	Upper bound on the 2-norms of subgradients of $\hat{\mathcal{L}}_\theta$ w.r.t. θ
B_Δ	Upper bound on the ∞-norms of gradients of $\hat{\mathcal{L}}_\lambda$ w.r.t. λ

Algorithm 3 "Practical" algorithm for optimizing the empirical proxy-Lagrangian game (Definition 1). This is essentially Algorithm 2 of Cotter et al. (2019), differing only in that it is applied to the two-dataset formulation of Definition 1.

PracticalTwoDataset $\left(\hat{\mathcal{L}}_\theta, \hat{\mathcal{L}}_\lambda : \Theta \times \Delta^m \to \mathbb{R}, T \in \mathbb{N}, \eta_\theta, \eta_\lambda \in \mathbb{R}_+\right)$:

Initialize $\theta^{(1)} = 0$ *// Assumes* $0 \in \Theta$

Initialize $M^{(1)} \in \mathbb{R}^{(m+1)\times(m+1)}$ with $M_{i,j} = 1/(m+1)$

For $t \in [T]$:

  Let $\lambda^{(t)} = \operatorname{fix} M^{(t)}$ *// fixed point of* $M^{(t)}$*, i.e. a stationary distribution*

  Let $\check{\Delta}_\theta^{(t)}$ be a stochastic subgradient of $\hat{\mathcal{L}}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. θ

  Let $\Delta_\lambda^{(t)}$ be a stochastic gradient of $\hat{\mathcal{L}}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. λ

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

  Update $\tilde{M}^{(t+1)} = M^{(t)} \odot \cdot\exp\left(\eta_\lambda \Delta_\lambda^{(t)}\left(\lambda^{(t)}\right)^T\right)$ *// $\Delta\lambda^T$ is an outer product; $\odot$ and $\cdot\exp$ are element-wise*

  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 w.r.t. KL divergence*

Return $\theta^{(1)}, \ldots, \theta^{(T)}$ and $\lambda^{(1)}, \ldots, \lambda^{(T)}$

Algorithm 4 "Practical" algorithm for optimizing a variant of the standard Lagrangian game, modified to support proxy constraints and two datasets. Here, instead of using the proxy-Lagrangian formulation of Definition 1, we take $\hat{\mathcal{L}}_\theta := \mathbb{E}_{x\sim S^{(\text{trn})}}\left[\ell_0(x;\theta) + \sum_{i=1}^m \lambda_i \tilde{\ell}_i(x;\theta)\right]$ and $\hat{\mathcal{L}}_\lambda := \mathbb{E}_{x\sim S^{(\text{val})}}\left[\ell_0(x;\theta) + \sum_{i=1}^m \lambda_i \ell_i(x;\theta)\right]$, with $\lambda \in \Lambda := \mathbb{R}_+^m$. Compared to Algorithm 3, this algorithm is further from those for which we can prove theoretical results (Algorithms 1 and 2), but is much closer to the Lagrangian-based approach of Agarwal et al. (2018). We include it to demonstrate that our two-dataset proposal works *as a heuristic*.

LagrangianTwoDataset $\left(\hat{\mathcal{L}}_\theta, \hat{\mathcal{L}}_\lambda : \Theta \times \mathbb{R}_+^m \to \mathbb{R}, T \in \mathbb{N}, \eta_\theta, \eta_\lambda \in \mathbb{R}_+\right)$:

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 $\hat{\mathcal{L}}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. θ

  Let $\Delta_\lambda^{(t)}$ be a stochastic gradient of $\hat{\mathcal{L}}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right)$ w.r.t. λ

  Update $\theta^{(t+1)} = \Pi_\Theta\left(\theta^{(t)} - \eta_\theta \check{\Delta}_\theta^{(t)}\right)$ *// Projected SGD updates (w.r.t. the Euclidean norm) ...*

  Update $\lambda^{(t+1)} = \Pi_\Lambda\left(\lambda^{(t)} + \eta_\lambda \Delta_\lambda^{(t)}\right)$ *// ...*

Return $\theta^{(1)}, \ldots, \theta^{(T)}$ and $\lambda^{(1)}, \ldots, \lambda^{(T)}$

A. Glossary

(External) Regret: Suppose we have a sequence of losses $\ell_t : \Theta \to \mathbb{R}$ that we wish to minimize, and an algorithm that produces a sequence $\theta^{(1)}, \ldots, \theta^{(T)} \in \Theta$. We can measure the performance of this algorithm via its *external regret*:

$$R = \sum_{t=1}^T \ell_t\left(\theta^{(t)}\right) - \min_{\theta^* \in \Theta} \sum_{t=1}^T \ell_t\left(\theta^*\right)$$

External regret is the difference between the total loss of the sequence generated by the algorithm, and that of the best element of Θ, chosen with the benefit of hindsight.

Swap Regret: To define swap regret, we must adjust the above setting slightly. In particular, we will assume that at each step we may select from a *finite* number of choices—we'll call this number m—and define each $\ell_t \in \mathbb{R}^m$ as a vector for which its ith element is the loss associated with the ith choice. The algorithm we are evaluating will be assumed to produce a sequence of probability vectors $p^{(1)}, \ldots, p^{(T)} \in \Delta^{m-1} \subseteq \mathbb{R}_+^m$, for which the external regret would be:

$$R = \sum_{t=1}^T \left\langle p^{(t)}, \ell_t \right\rangle - \min_{p^* \in \Delta^{m-1}} \sum_{t=1}^T \left\langle p^*, \ell_t \right\rangle$$

Unlike external regret, which compares to the best constant strategy, *swap regret* compares to a strategy that varies in a particular way: we can retroactively decide that the probability mass that the $p^{(t)}$ sequence assigned to the ith element should instead be assigned to the i'th, and the jth to the j'th, and so on. Mathematically, swap regret can be defined as:

$$R = \sum_{t=1}^{T} \left\langle p^{(t)}, \ell_t \right\rangle - \min_{M \in \mathcal{M}} \sum_{t=1}^{T} \left\langle M p^{(t)}, \ell_t \right\rangle$$

here, $\mathcal{M} \subseteq \mathbb{R}^{m \times m}$ is the space of left-stochastic matrices.

Pure Equilibrium: Consider a two-player game with payoff functions $\mathcal{L}_\theta : \Theta \times \Lambda \to \mathbb{R}$ and $\mathcal{L}_\lambda : \Theta \times \Lambda \to \mathbb{R}$. We'll adopt the convention that the first player wishes to choose θ to minimize $\mathcal{L}_\theta$, while the second wishes to choose λ to maximize $\mathcal{L}_\lambda$. If $\mathcal{L}_\theta = \mathcal{L}_\lambda$, then the game is said to be *zero-sum*. Otherwise, it is non-zero-sum. In either case, a *pure equilibrium* is a pair $\theta^*, \lambda^* \in \Theta \times \Lambda$ satisfying some properties (depending on the type of equilibrium). For example, for a *pure Nash equilibrium*, both players suffer no external regret:

$$\begin{aligned} \mathcal{L}_\theta\left(\theta^*, \lambda^*\right) \leq & \mathcal{L}_\theta\left(\theta, \lambda^*\right) \\ \mathcal{L}_\lambda\left(\theta^*, \lambda^*\right) \geq & \mathcal{L}_\lambda\left(\theta^*, \lambda\right) \end{aligned}$$

the above holding for all $\theta \in \Theta$ and $\lambda \in \Lambda$. In words, the θ-player cannot improve upon θ^*, and the λ-player cannot improve upon λ^*.

Mixed Equilibrium: A *mixed equilibrium* differs from a pure equilibrium in that, instead of being a pair of elements taken from $\Theta \times \Lambda$, it is defined either as a joint distribution P supported on $\Theta \times \Lambda$, or—when possible—as a pair of marginal distributions P_θ and P_λ supported on Θ and Λ, respectively (with $P := P_\theta \times P_\lambda$). For example, a *mixed Nash equilibrium* is a pair of marginal distributions for which both players suffer no external regret:

$$\begin{aligned} \mathbb{E}_{\theta^* \sim P_\theta, \lambda^* \sim P_\lambda}\left[\mathcal{L}_\theta\left(\theta^*, \lambda^*\right)\right] \leq & \mathbb{E}_{\lambda^* \sim P_\lambda}\left[\mathcal{L}_\theta\left(\theta, \lambda^*\right)\right] \\ \mathbb{E}_{\theta^* \sim P_\theta, \lambda^* \sim P_\lambda}\left[\mathcal{L}_\theta\left(\theta^*, \lambda^*\right)\right] \geq & \mathbb{E}_{\theta^* \sim P_\theta}\left[\mathcal{L}_\theta\left(\theta^*, \lambda\right)\right] \end{aligned}$$

the above holding, as before, for all $\theta \in \Theta$ and $\lambda \in \Lambda$.

Coarse Correlated Equilibrium: A *coarse correlated equilibrium* is a joint distribution P for which both players suffer no external regret:

$$\begin{aligned} \mathbb{E}_{\theta^*, \lambda^* \sim P}\left[\mathcal{L}_\theta\left(\theta^*, \lambda^*\right)\right] \leq & \mathbb{E}_{\theta^*, \lambda^* \sim P}\left[\mathcal{L}_\theta\left(\theta, \lambda^*\right)\right] \\ \mathbb{E}_{\theta^*, \lambda^* \sim P}\left[\mathcal{L}_\theta\left(\theta^*, \lambda^*\right)\right] \geq & \mathbb{E}_{\theta^*, \lambda^* \sim P}\left[\mathcal{L}_\theta\left(\theta^*, \lambda\right)\right] \end{aligned}$$

the above holding, once again, for all $\theta \in \Theta$ and $\lambda \in \Lambda$. A coarse correlated equilibrium is *weaker* than a mixed Nash equilibrium: the former is a joint distribution P supported on $\Theta \times \Lambda$, while the latter can be decomposed as the product of two marginal distributions ($P := P_\theta \times P_\lambda$).

In Theorems 1 and 2, the relevant type of mixed equilibrium—which does not have an established name [¶]—is slightly stronger than a coarse correlated equilibrium, but is still easier to find than a mixed Nash equilibrium: it is a joint distribution P for which the θ-player suffers no external regret, while the λ-player suffers no *swap* regret. For further details, see Theorem 3 in Appendix E, or Theorem 8 of Cotter et al. (2019).

B. Examples of Constraints

In this appendix, we'll provide some examples of constrained optimization problems in the form of Equation 1.

B.1. Neyman-Pearson

The first example we'll consider is Neyman-Pearson classification (Davenport et al., 2010; Gasso et al., 2011). Imagine that we wish to learn a classification function $f : \mathcal{X} \times \Theta \to \mathbb{R}$ parameterized by θ, with the goal being to minimize the false

[¶] We sometimes call it a *semi-coarse correlated equilibrium*, since a *correlated equilibrium* is a joint distribution for which all players suffer no swap regret.

positive rate, subject to the constraint that the false negative rate be at most 10%:

$$\begin{aligned}\underset{\theta\in\Theta}{\text{minimize}}\ &\mathbb{E}_{x,y|y=-1}\left[\mathbf{1}\left\{f\left(x;\theta\right)\geq 0\right\}\right]\\ \text{s.t. }&\mathbb{E}_{x,y|y=1}\left[\mathbf{1}\left\{f\left(x;\theta\right)\leq 0\right\}\right]\leq 0.1\end{aligned}$$

One way to convert this problem into the form of Equation 1 is to define $\mathcal{D}_+$ and $\mathcal{D}_-$ as the marginal distributions over xs for which $y=+1$ and $y=-1$ (respectively), and take $\mathcal{D}:=\mathcal{D}_+\times\mathcal{D}_-$ so that $\mathcal{D}$ is a distribution over *pairs* of feature vectors, the first having a positive label, and the second a negative label. Defining $\ell_0\left(x_+,x_-;\theta\right):=\mathbf{1}\left\{f\left(x_-\right)\geq 0\right\}$ and $\ell_1\left(x_+,x_-;\theta\right):=\mathbf{1}\left\{f\left(x_+\right)\leq 0\right\}-0.1$ puts the original Neyman-Pearson problem in the form of Equation 1.

In practice, the fact that ℓ_0 and ℓ_1 are defined in terms of indicator functions, and are therefore discontinuous, will be problematic. To fix this, using the formulation of Definition 1, one could instead define $\ell_0\left(x_+,x_-;\theta\right):=\max\left\{0,1+f\left(x_-\right)\right\}$ as a hinge upper bound on the false positive rate, and leave ℓ_1 as-is while defining the corresponding proxy-constraint to be $\tilde{\ell}_1\left(x_+,x_-;\theta\right):=\max\left\{0,1-f\left(x_+\right)\right\}$.

B.2. Equal Opportunity

The second example we'll consider is a fairness-constrained problem. As before, we'll take $f:\mathcal{X}\times\Theta\to\mathbb{R}$ to be a classification function, but we'll imagine that each $x\in\mathcal{X}$ contains a feature $x_k\in\{1,2,3\}$ indicating to which of three protected classes the corresponding example belongs. We will seek to minimize the overall error rate, subject to the constraint that, for each of the three protected classes, the false negative rate is at most 110% of the false negative rate across all three classes (this is essentially an equal opportunity constraint (Hardt et al., 2016)):

$$\begin{aligned}\underset{\theta\in\Theta}{\text{minimize}}\ &\mathbb{E}_{x,y}\left[\mathbf{1}\left\{yf\left(x;\theta\right)\leq 0\right\}\right]\\ \underset{\forall i\in\{1,2,3\}}{\text{s.t.}}\ &\mathbb{E}_{x,y|y=1\wedge x_k=i}\left[\mathbf{1}\left\{f\left(x;\theta\right)\leq 0\right\}\right]\leq 1.1\cdot\mathbb{E}_{x,y|y=1}\left[\mathbf{1}\left\{f\left(x;\theta\right)\leq 0\right\}\right]\end{aligned}$$

While we could use the same approach as in the Neyman-Pearson example, i.e. taking marginals and crossing them to define a data distribution over tuples of examples, we'll instead take $\mathcal{D}$ to be the data distribution over $\mathcal{X}\times\{\pm 1\}$ pairs, and use the indicator feature x_k to define:

$$\begin{aligned}\ell_0\left(x,y;\theta\right):=&\mathbf{1}\left\{yf\left(x;\theta\right)\leq 0\right\}\\ \ell_i\left(x,y;\theta\right):=&\frac{\mathbf{1}\{y=1\wedge x_k=i\}\mathbf{1}\left\{f\left(x;\theta\right)\leq 0\right\}}{\Pr\{y=1\wedge x_k=i\mid x,y\sim\mathcal{D}\}}-1.1\cdot\frac{\mathbf{1}\{y=1\}\mathbf{1}\left\{f\left(x;\theta\right)\leq 0\right\}}{\Pr\{y=1\mid x,y\sim\mathcal{D}\}}\end{aligned}$$

for all $i\in\{1,2,3\}$, where we assume that the probabilities in the denominators of the ratios defining ℓ_i are constants known a priori.

As in the Neyman-Pearson example, in practice the indicators in the objective function could be replaced with differentiable upper bounds, and a differentiable proxy-constraint $\tilde{\ell}_i$ could be introduced for each ℓ_i.

C. Shrinking

Cotter et al. (2019) introduced a procedure for "shrinking" the support size of a $\bar{\theta}$. When adapted to our setting, the first step is to evaluate the objective and constraints for every iterate (in practice, this is overkill; one should subsample the iterates):

$$\begin{aligned}\vec{\ell_0}^{(t)}=&\frac{1}{\left|S^{(\text{trn})}\right|}\sum_{x\in S^{(\text{trn})}}\ell_0\left(x;\theta^{(t)}\right)\\ \vec{\ell_i}^{(t)}=&\frac{1}{\left|S^{(\text{val})}\right|}\sum_{x\in S^{(\text{val})}}\ell_i\left(x;\theta^{(t)}\right)\end{aligned}$$

Next, we optimize a linear program (LP) that seeks a distribution p over $\hat{\Theta}$ that minimizes the objective while violating no constraint by more than ϵ:

$$\min_{p\in\Delta^{T-1}\subseteq\mathbb{R}_+^T}\left\langle p,\vec{\ell_0}\right\rangle\quad\underset{\forall i\in[m]}{\text{s.t.}}\ \left\langle p,\vec{\ell_i}\right\rangle\leq\epsilon$$

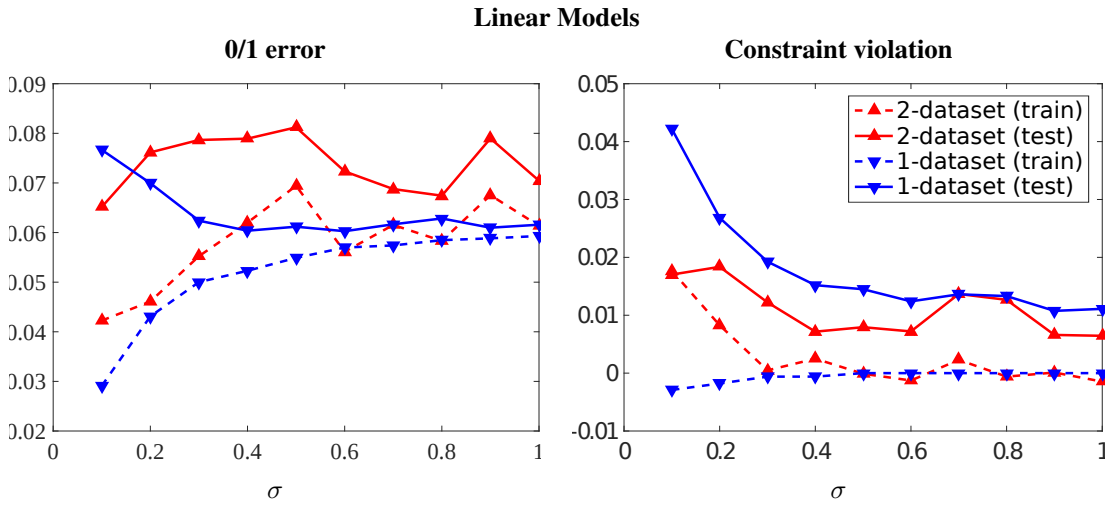

Figure 1. Results of the experiment of Appendix D, using Algorithm 3 on a linear model. The left-hand plot demonstrates both that overfitting increases as σ decreases, as expected, and that our model performs worse than the baseline, in terms of accuracy. In the right-hand plot, however, we can see that our approach generalizes better on the constraints (the two red curves are closer together than the two blue ones), and also does a better job of satisfying the constraints on the testing set (the solid red curve is below the solid blue one).

Notice that the $p \in \Delta^{T-1}$ condition adds several implicit simplex constraints to this LP. The key to this procedure is that, as Cotter et al. (2019) show, every vertex p of this linear program has at most $m+1$ nonzero elements, where m is the number of constraints.

In particular, if ϵ is chosen to be the maximum validation constraint violation of the "original" stochastic classifier $\bar{\theta}$, then an *optimal* vertex p^* will have a training objective function value and maximum validation constraint violation that are no larger than those of $\bar{\theta}$, and p^* will be supported on only $m+1$ $\theta^{(t)}$s. Furthermore, since the resulting stochastic classifier is still supported on a subset of $\hat{\Theta}$, the generalization definitions of Section 3.1 apply to it just as well as they did to $\bar{\theta}$.

While it would be possible to provide optimality and feasibility guarantees for the result of this "shrinking" procedure, we will only use it in our experiments (Section 5). There, instead of taking the LP's ϵ parameter to be the maximum validation constraint violation of $\bar{\theta}$, we follow Cotter et al. (2019)'s suggestion to use an outer bisection search to find the smallest $\epsilon \geq 0$ for which the LP is feasible.

D. Simulated-data Experiments

These experiments are performed on a simulated binary classification problem designed to be especially prone to overfitting, and is intended to explore the relationship between generalization and model complexity in the context of our proposed approach.

To generate the dataset, we first draw $n = 1000$ points $z_1, \ldots, z_n$ from two overlapping Gaussians in $\mathbb{R}^2$, and another n points $w_1, \ldots, w_n$ from the same distribution. For each i, we let the classification label y_i indicate which of the two Gaussians z_i was drawn from, and generate a feature vector $x_i \in \mathcal{X} := \mathbb{R}^n$ such that the jth feature satisfies $x_{i,j} := \exp(-\|z_i - w_j\|^2 / 2\sigma^2)$. Our results are averaged over ten runs, with different random splits of the data into equally-sized training, validation and testing datasets.

The classification task is learn a classifier on $\mathcal{X}$ that determines which of the two Gaussian distributions generated the example, with the model's recall constrained to be at least 97%. The σ parameter partly controls the amount of overfitting: as $\sigma \to 0$, a linear classifier on $\mathcal{X}$ approaches a 1-nearest neighbor classifier over $w_1, \ldots, w_n$, which one would expect to overfit badly.

We trained four sets of models using Algorithm 3: linear, and one-hidden-layer neural networks with 5, 10 and 100 hidden ReLU units. We also varied σ between 0 and 1. Figures 1 and 2 show that our approach consistently comes closer to satisfying the constraints on the testing set—and the training and testing constraint violations are much closer *to each other* for our approach than for the baseline—but that, as one would expect, this comes at a slight cost in testing accuracy. In light

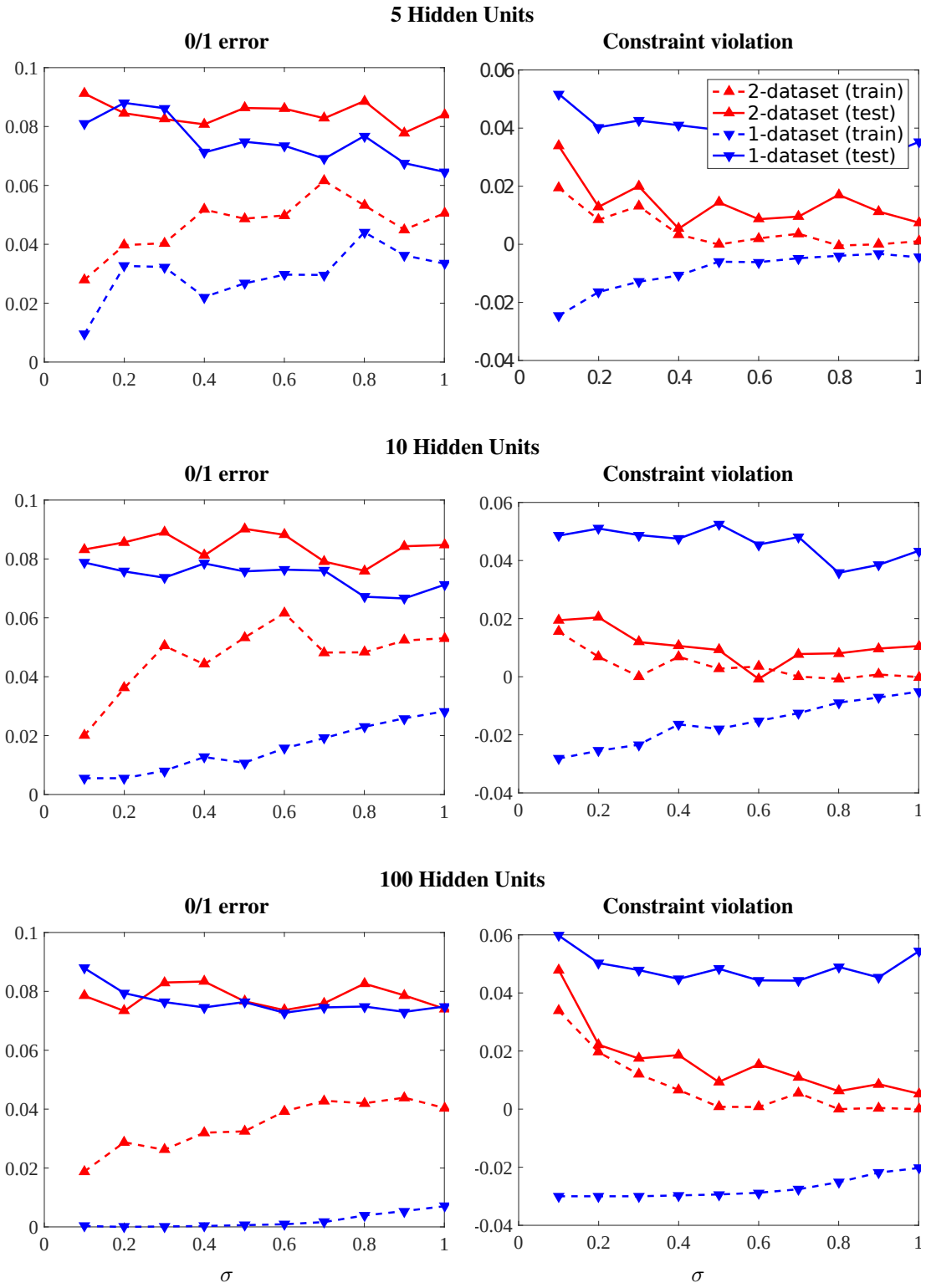

Figure 2. Same as Figure 1, but for one-hidden-layer neural networks with 5, 10 and 100 hidden units. The results follow the same general trend as those of Figure 1, but, just as one would expect, the benefits of our proposed two-dataset approach become more pronounced as the model complexity increases.

of the fact that our theoretical results (Section 4) bound constraint generalization independently of model complexity, it's unsurprising that our approach is most advantageous for the most complex models (100-hidden unit), and less so for the simplest (linear).

E. Proofs

We'll begin by reproducing a definition from Cotter et al. (2019), which introduced the idea of the in-expectation proxy-Lagrangian game:

Definition 4. ***(Definition 2 of Cotter et al. (2019))*** *Given proxy loss functions $\tilde{\ell}_i(x;\theta) \geq \ell_i(x;\theta)$ for all $x \in \mathcal{X}$ and $i \in [m]$, the proxy-Lagrangians $\mathcal{L}_\theta, \mathcal{L}_\lambda : \Theta \times \Lambda \to \mathbb{R}$ of Equation 1 are:*

$$\mathcal{L}_\theta(\theta,\lambda) := \mathbb{E}_{x\sim\mathcal{D}}\left[\lambda_1 \ell_0(x;\theta) + \sum_{i=1}^{m} \lambda_{i+1}\tilde{\ell}_i(x;\theta)\right]$$
$$\mathcal{L}_\lambda(\theta,\lambda) := \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{i=1}^{m} \lambda_{i+1}\ell_i(x;\theta)\right]$$

where $\Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$ is the m-probability-simplex.

This definition differs from Definition 1 (in Section 3) in that the former is the in-expectation version of the latter, which additionally is written in terms of separate *i.i.d.* training and validation sets.

Theorem 2 of Cotter et al. (2019) characterizes the optimality and feasibility properties of a particular type of Φ-correlated equilibrium of Definition 4. We adapt it to our setting, giving the analogous result for such an equilibrium of Definition 1:

Theorem 3. *Define $\mathcal{M}$ as the set of all left-stochastic $(m+1)\times(m+1)$ matrices, $\Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$ as the m-probability-simplex, and assume that each $\tilde{\ell}_i$ upper bounds the corresponding ℓ_i. Let $\theta^{(1)},\ldots,\theta^{(T)} \in \Theta$ and $\lambda^{(1)},\ldots,\lambda^{(T)} \in \Lambda$ be sequences satisfying:*

$$\frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{L}}_\theta\left(\theta^{(t)},\lambda^{(t)}\right) - \inf_{\theta^*\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{L}}_\theta\left(\theta^*,\lambda^{(t)}\right) \leq \epsilon_\theta \tag{8}$$
$$\max_{M^*\in\mathcal{M}} \frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{L}}_\lambda\left(\theta^{(t)},M^*\lambda^{(t)}\right) - \frac{1}{T}\sum_{t=1}^{T} \hat{\mathcal{L}}_\lambda\left(\theta^{(t)},\lambda^{(t)}\right) \leq \epsilon_\lambda$$

Define $\hat{\Theta} := \left\{\theta^{(1)},\ldots,\theta^{(T)}\right\}$. Let $\bar{\theta}$ be a random variable taking values from $\hat{\Theta}$, defined such that $\bar{\theta} = \theta^{(t)}$ 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:

$$\mathbb{E}_{\bar{\theta},x\sim\mathcal{D}}\left[\ell_0\left(x;\bar{\theta}\right)\right] \leq \inf_{\theta^*\in\Theta:\forall i.\mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{\ell}_i(x;\theta^*)\right]\leq 0} \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0(x;\theta^*)\right] + \frac{1}{\bar{\lambda}_1}\left(\epsilon_\theta + \epsilon_\lambda + 2\tilde{G}^{(\mathrm{trn})}(\Theta) + G^{(\mathrm{val})}(\hat{\Theta})\right) \tag{9}$$

and nearly-feasible:

$$\max_{i\in[m]} \mathbb{E}_{\bar{\theta},x\sim\mathcal{D}}\left[\ell_i\left(x;\bar{\theta}\right)\right] \leq \frac{\epsilon_\lambda}{\bar{\lambda}_1} + G^{(\mathrm{val})}(\hat{\Theta}) \tag{10}$$

Additionally, if there exists a $\theta' \in \Theta$ that satisfies all of the constraints with margin γ (i.e. $\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i(x;\theta')\right] \leq -\gamma$ for all $i \in [m]$), then:

$$\bar{\lambda}_1 \geq \frac{\gamma - \epsilon_\theta - \epsilon_\lambda - 2\tilde{G}^{(\mathrm{trn})}(\Theta) - G^{(\mathrm{val})}(\hat{\Theta})}{\gamma + B_{\ell_0}} \tag{11}$$

where $B_{\ell_0} \geq \sup_{\theta\in\Theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0(x;\theta)\right] - \inf_{\theta\in\Theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0(x;\theta)\right]$ is a bound on the range of the objective loss.

Proof. This proof closely follows those of Theorem 4 and Lemma 7 in Cotter et al. (2019)—the only addition is that, in this proof, we use the empirical proxy-Lagrangian formulation with separate training and validation datasets (Definition 1), instead of the in-expectation proxy-Lagrangian (Definition 4), and therefore need to account for generalization.

Optimality: If we choose M^* to be the matrix with its first row being all-one, and all other rows being all-zero, then $\hat{\mathcal{L}}_\lambda\left(\theta, M^*\lambda\right) = 0$, which shows that the first term in the LHS of the second line of Equation 8 is nonnegative. Hence:

$$-\mathbb{E}_{t\sim[T]}\left[\hat{\mathcal{L}}_\lambda\left(\theta^{(t)}, \lambda^{(t)}\right)\right] \leq \epsilon_\lambda$$

so by the definitions of $\hat{\mathcal{L}}_\lambda$ (Definition 1) and $G^{(\mathrm{val})}(\hat{\Theta})$ (Definition 2), and the facts that $\tilde{\ell}_i \geq \ell_i$ and $\lambda^{(t)} \in \Delta^m$:

$$\mathbb{E}_{t\sim[T],x\sim\mathcal{D}}\left[\sum_{i=1}^{m}\lambda_{i+1}^{(t)}\tilde{\ell}_i\left(x;\theta^{(t)}\right)\right] \geq -\epsilon_\lambda - G^{(\mathrm{val})}(\hat{\Theta})$$

Notice that $\hat{\mathcal{L}}_\theta$ is linear in λ, so the first line of Equation 8, combined with the definition of $\tilde{G}^{(\mathrm{trn})}(\Theta)$, becomes:

$$\begin{aligned}\mathbb{E}_{t\sim[T],x\sim\mathcal{D}}&\left[\lambda_1^{(t)}\ell_0\left(x;\theta^{(t)}\right) + \sum_{i=1}^{m}\lambda_{i+1}^{(t)}\tilde{\ell}_i\left(x;\theta^{(t)}\right)\right] - \inf_{\theta^*\in\Theta}\mathbb{E}_{x\sim\mathcal{D}}\left[\bar{\lambda}_1\ell_0\left(x;\theta^*\right) + \sum_{i=1}^{m}\bar{\lambda}_{i+1}\tilde{\ell}_i\left(x;\theta^*\right)\right]\\ \leq&\epsilon_\theta + 2\tilde{G}^{(\mathrm{trn})}(\Theta)\end{aligned}$$

Combining the above two results:

$$\begin{aligned}\mathbb{E}_{t\sim[T],x\sim\mathcal{D}}&\left[\lambda_1^{(t)}\ell_0\left(x;\theta^{(t)}\right)\right] - \inf_{\theta^*\in\Theta}\mathbb{E}_{x\sim\mathcal{D}}\left[\bar{\lambda}_1\ell_0\left(x;\theta^*\right) + \sum_{i=1}^{m}\bar{\lambda}_{i+1}\tilde{\ell}_i\left(x;\theta^*\right)\right]\\ \leq&\epsilon_\theta + \epsilon_\lambda + 2\tilde{G}^{(\mathrm{trn})}(\Theta) + G^{(\mathrm{val})}(\hat{\Theta})\end{aligned} \tag{12}$$

Choose θ^* to be the optimal solution that satisfies the *proxy* constraints, so that $\mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{\ell}_i\left(x;\theta^*\right)\right] \leq 0$ for all $i \in [m]$. Then:

$$\mathbb{E}_{t\sim[T],x\sim\mathcal{D}}\left[\lambda_1^{(t)}\ell_0\left(x;\theta^{(t)}\right)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[\bar{\lambda}_1\ell_0\left(x;\theta^*\right)\right] \leq \epsilon_\theta + \epsilon_\lambda + 2\tilde{G}^{(\mathrm{trn})}(\Theta) + G^{(\mathrm{val})}(\hat{\Theta})$$

which is the optimality claim.

Feasibility: We'll begin by simplifying our notation: define $g_1\left(\theta\right) := 0$ and $g_{i+1}\left(\theta\right) := \mathbb{E}_{x\sim S^{(\mathrm{val})}}\left[\ell_i\left(x;\theta\right)\right]$ for $i \in [m]$, so that $\hat{\mathcal{L}}_\lambda\left(\theta,\lambda\right) = \langle\lambda, g_:\left(\theta\right)\rangle$. Consider the first term in the LHS of the second line of Equation 8:

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

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 8, we can use the fact that $g_1\left(\theta\right) = 0$:

$$\mathbb{E}_{t\sim[T]}\left[\sum_{i=2}^{m+1}\lambda_i^{(t)}g_i\left(\theta^{(t)}\right)\right] \leq \sum_{i=2}^{m+1}\max_{j\in[m+1]}\mathbb{E}_{t\sim[T]}\left[\lambda_i^{(t)}g_j\left(\theta^{(t)}\right)\right]$$

Plugging these two results into the second line of Equation 8, the two sums collapse, leaving:

$$\max_{i\in[m+1]}\mathbb{E}_{t\sim[T]}\left[\lambda_1^{(t)}g_i\left(\theta^{(t)}\right)\right] \leq \epsilon_\lambda$$

Substituting the definitions of g_i and $G^{(\text{val})}(\hat{\Theta})$ (since the g_is are defined on $S^{(\text{val})}$, but we want our result to hold on $\mathcal{D}$) then yields the feasibility claim.

Bound on $\bar{\lambda}_1$: Choosing $\theta^* = \theta'$ in Equation 12 (recall that θ' satisfies all of the proxy constraints with margin γ) and substituting the definition of B_{ℓ_0}:

$$\begin{aligned}\epsilon_\theta + \epsilon_\lambda + 2\tilde{G}^{(\text{trn})}(\Theta) + G^{(\text{val})}(\hat{\Theta}) \geq & \mathbb{E}_{t\sim[T],x\sim\mathcal{D}}\left[\lambda_1^{(t)}\ell_0\left(x;\theta^{(t)}\right) - \lambda_1^{(t)}\ell_0\left(x;\theta'\right)\right] + \left(1-\bar{\lambda}_1\right)\gamma \\ \geq & -\bar{\lambda}_1 B_{\ell_0} + \left(1-\bar{\lambda}_1\right)\gamma\end{aligned}$$

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

Before moving on to the convergence and generalization properties of our actual algorithms, we need to state some (fairly standard) elementary results:

Definition 5. *We say that $C_r \subseteq \mathbb{R}^{m+1}$ is a radius-r external covering of $\Lambda := \Delta^m \subseteq \mathbb{R}^{m+1}_+$ w.r.t. the 1-norm if for every $\lambda \in \Lambda$ there exists a $\tilde{\lambda} \in C_r$ for which $\left\|\lambda - \tilde{\lambda}\right\|_1 \leq r$. Notice that we do not require C_r to be a subset of Λ—this is why it's an* external *covering.*

Lemma 1. *Assuming that $r \leq 1$, there exists a radius-r external covering of $\Lambda := \Delta^m \subseteq \mathbb{R}^{m+1}_+$ w.r.t. the 1-norm of size no larger than $(5/r)^m$.*

Proof. Consider the m-dimensional unit ball $\tilde{B} := \left\{\tilde{\lambda} \in \mathbb{R}^m : \left\|\tilde{\lambda}\right\|_1 \leq 1\right\}$ w.r.t. the 1-norm (note that we could instead consider only the positive orthant, which would improve the constant in the overall result). There exists a radius-$r/2$ covering $\tilde{C}_r \subseteq \mathbb{R}^m$ of $\tilde{B}$ with $\left|\tilde{C}_r\right| \leq (1+4/r)^m \leq (5/r)^m$ (Bartlett, 2013).

Define $C_r \subseteq \mathbb{R}^{m+1}$ as:

$$C_r = \left\{\begin{bmatrix}\tilde{\lambda} \\ 1 - \left\|\tilde{\lambda}\right\|_1\end{bmatrix} : \tilde{\lambda} \in \tilde{C}_r\right\}$$

Notice that we do not necessarily have that $C_r \subseteq \Delta^m$, i.e. this will be an *external* covering.

From any $\lambda \in \Delta^m$, we can define $\lambda' \in \mathbb{R}^m_+$ by dropping the last coordinate of λ, and we'll have that $\|\lambda'\|_1 \leq \|\lambda\|_1 = 1$, so there will exist a $\tilde{\lambda} \in \tilde{C}_r$ such that $\left\|\tilde{\lambda} - \lambda'\right\|_1 \leq r/2$, which implies that the corresponding element of C_r is r-far from λ, showing that C_r is a radius-r covering of Δ^m w.r.t. the 1-norm. □

Lemma 2. *Let S be an* i.i.d. *sample from a distribution $\mathcal{D}$ supported on $\mathcal{X}$, and $\hat{\Theta} \subseteq \Theta$ the finite set of permitted model parameters, which defines a a finite function class ($\hat{\Theta}$ may be a random variable, but must be independent of S). Suppose that $\ell : \mathcal{X} \times \Theta \to [a,b]$ with $B_\ell := b - a$. Then:*

$$\left|\frac{1}{|S|}\sum_{x\in S}\ell\left(x;\theta\right) - \mathbb{E}_{x\sim\mathcal{D}}\left[\ell\left(x;\theta\right)\right]\right| < B_\ell\sqrt{\frac{\ln\left(2\left|\hat{\Theta}\right|/\delta\right)}{2\left|S\right|}}$$

for all $\theta \in \hat{\Theta}$, with probability at least $1-\delta$ over the sampling of S.

Proof. Allowing $\hat{\Theta}$ to be a random variable independent of S, instead of a constant set, doesn't significantly change the standard proof (e.g. Srebro, 2016) . By Hoeffding's inequality:

$$\Pr\left\{\left|\frac{1}{|S|}\sum_{x\in S}\ell\left(x;\theta\right) - \mathbb{E}_{x\sim\mathcal{D}}\left[\ell\left(x;\theta\right)\right]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2\left|S\right|\epsilon^2}{B_\ell^2}\right)$$

the above holding for any $\theta \in \Theta$. Since $\hat{\Theta}$ is independent of S, we can apply the union bound over all $\theta \in \hat{\Theta}$:

$$\Pr\left\{\exists\theta \in \hat{\Theta}.\left|\frac{1}{|S|}\sum_{x\in S}\ell\left(x;\theta\right) - \mathbb{E}_{x\sim\mathcal{D}}\left[\ell\left(x;\theta\right)\right]\right| \geq \epsilon\right\} \leq 2\left|\theta\right|\exp\left(-\frac{2\left|S\right|\epsilon^2}{B_\ell^2}\right)$$

Rearranging terms yields the claimed result. □

E.1. Algorithm 1

Lemma 3. *If we take* $\eta_\lambda := \sqrt{(m+1)\ln(m+1)/TB_\Delta^2}$*, then the result of Algorithm 1 satisfies the conditions of Theorem 3 with:*

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

where $B_{\tilde{\ell}} \geq |\ell(x,\theta)|$ *for all* $\ell \in \left\{\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m\right\}$*, and* $B_\Delta \geq \max_{t\in[T]} \left\|\Delta_\lambda^{(t)}\right\|_\infty$ *is a bound on the gradients.*

Proof. Since C_r is a radius-r external covering of $\Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$ w.r.t. the 1-norm, we must have that $\left\|\tilde{\lambda}^{(t)} - \lambda^{(t)}\right\|_1 \leq r$, which implies by Definition 1 that:

$$\left|\hat{\mathcal{L}}_\theta\left(\theta, \tilde{\lambda}^{(t)}\right) - \hat{\mathcal{L}}_\lambda\left(\theta, \lambda^{(t)}\right)\right| \leq rB_{\tilde{\ell}}$$

the above holding for all $\theta \in \Theta$, and all t. In particular, this implies that:

$$\begin{aligned}\hat{\mathcal{L}}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right) \leq& \hat{\mathcal{L}}_\theta\left(\theta^{(t)}, \tilde{\lambda}^{(t)}\right) + rB_{\tilde{\ell}}\\ \leq& \inf_{\theta^*\in\Theta} \hat{\mathcal{L}}_\theta\left(\theta^*, \tilde{\lambda}^{(t)}\right) + \rho + rB_{\tilde{\ell}}\\ \leq& \inf_{\theta^*\in\Theta} \hat{\mathcal{L}}_\theta\left(\theta^*, \lambda^{(t)}\right) + \rho + 2rB_{\tilde{\ell}}\end{aligned}$$

Therefore:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \hat{\mathcal{L}}_\theta\left(\theta^{(t)}, \lambda^{(t)}\right) \leq& \frac{1}{T}\sum_{t=1}^T \inf_{\theta^*\in\Theta} \hat{\mathcal{L}}_\theta\left(\theta^*, \lambda^{(t)}\right) + \rho + 2rB_{\tilde{\ell}}\\ \leq& \inf_{\theta^*\in\Theta} \frac{1}{T}\sum_{t=1}^T \hat{\mathcal{L}}_\theta\left(\theta^*, \lambda^{(t)}\right) + \rho + 2rB_{\tilde{\ell}}\end{aligned}$$

so the first condition of Theorem 3 is satisfied with the claimed ϵ_θ.

The second condition, on ϵ_λ, follows immediately from Lemma 8 of Appendix C.1 of Cotter et al. (2019), taking $\tilde{m} = m+1$. □

Lemma 4. *If we take* $\hat{\Theta} := \left\{\theta^{(1)}, \ldots, \theta^{(T)}\right\}$ *as in Theorem 3, where* $\theta^{(1)}, \ldots, \theta^{(T)}$ *are the result of Algorithm 1, then with probability* $1-\delta$ *over the sampling of* $S^{(\mathrm{val})}$*:*

$$G^{(\mathrm{val})}(\hat{\Theta}) < B_\ell\sqrt{\frac{\ln(2m|C_r|/\delta)}{2\left|S^{(\mathrm{val})}\right|}}$$

where $B_\ell \geq \max_{i\in[m]}(b_i - a_i)$ *assuming that the range of each* ℓ_i *is the interval* $[a_i, b_i]$*.*

Proof. Since each $\theta^{(t)}$ is uniquely associated with a $\tilde{\lambda}^{(t)} \in C_r$ (Definition 3), we will have that $\hat{\Theta} \subseteq \Theta_{C_r}$, where:

$$\Theta_{C_r} := \left\{\mathcal{O}_\rho\left(\hat{\mathcal{L}}_\theta\left(\cdot, \tilde{\lambda}\right)\right) : \tilde{\lambda} \in C_r\right\}$$

Because the oracle call defining Θ_{C_r} depends only on $\hat{\mathcal{L}}_\theta$, which itself depends only on $S^{(\mathrm{trn})}$, we can apply Lemma 2 to Θ_{C_r}, yielding that, for each $i \in [m]$, the following holds with probability δ/m for all $\theta \in \Theta_{C_r}$:

$$\left|\frac{1}{\left|S^{(\mathrm{val})}\right|}\sum_{x\in S^{(\mathrm{val})}} \ell_i(x;\theta) - \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i(x;\theta)\right]\right| < B_\ell\sqrt{\frac{\ln(2m|\Theta_{C_r}|/\delta)}{2\left|S^{(\mathrm{val})}\right|}}$$

The claimed result on $G^{(\mathrm{val})}(\hat{\Theta})$ then follows from the union bound and the facts that $\hat{\Theta} \subseteq \Theta_{C_r}$ and $|\Theta_{C_r}| = |C_r|$. □

Theorem 1. *Given any $\epsilon > 0$, there exists a covering C_r such that, if we take $T \geq 4B_\Delta^2 (m+1)\ln(m+1)/\epsilon^2$ and $\eta_\lambda = \sqrt{(m+1)\ln(m+1)/TB_\Delta^2}$, where $B_\Delta \geq \max_{t\in[T]} \left\|\Delta_\lambda^{(t)}\right\|_\infty$ is a bound on the gradients, then the following hold, where $\hat{\Theta} := \left\{\theta^{(1)}, \ldots, \theta^{(T)}\right\}$ is the set of results of Algorithm 1.*

Optimality and Feasibility: *Let $\bar{\theta}$ be a random variable taking values from $\hat{\Theta}$, defined such that $\bar{\theta} = \theta^{(t)}$ 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:*

$$\mathbb{E}_{\bar{\theta},x\sim\mathcal{D}}\left[\ell_0\left(x;\bar{\theta}\right)\right] \leq \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0\left(x;\theta^*\right)\right] + \frac{1}{\bar{\lambda}_1}\left(\rho + 2\epsilon + 2\tilde{G}^{(\mathrm{trn})}(\Theta) + G^{(\mathrm{val})}(\hat{\Theta})\right) \tag{3}$$

where θ^* minimizes $\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0(x;\cdot)\right]$ subject to the proxy-constraints $\mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{\ell}_i(x;\theta^*)\right] \leq 0$. It is also nearly-feasible:

$$\max_{i\in[m]} \mathbb{E}_{\bar{\theta},x\sim\mathcal{D}}\left[\ell_i\left(x;\bar{\theta}\right)\right] \leq \frac{\epsilon}{\bar{\lambda}_1} + G^{(\mathrm{val})}(\hat{\Theta}) \tag{4}$$

Additionally, if there exists a $\theta' \in \Theta$ that satisfies all of the constraints with margin γ (i.e. $\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i(x;\theta')\right] \leq -\gamma$ for all $i \in [m]$), then:

$$\bar{\lambda}_1 \geq \frac{1}{\gamma + B_{\ell_0}}\left(\gamma - \rho - 2\epsilon - 2\tilde{G}^{(\mathrm{trn})}(\Theta) - G^{(\mathrm{val})}(\hat{\Theta})\right) \tag{5}$$

where $B_{\ell_0} \geq \sup_{\theta\in\Theta}\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0(x;\theta)\right] - \inf_{\theta\in\Theta}\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0(x;\theta)\right]$ is a bound on the range of the objective loss.

Generalization: *With probability $1-\delta$ over the sampling of $S^{(\mathrm{val})}$:*

$$G^{(\mathrm{val})}(\hat{\Theta}) < B_\ell\sqrt{\frac{m\ln\left(10B_{\tilde{\ell}}/\epsilon\right) + \ln(2m/\delta)}{2\left|S^{(\mathrm{val})}\right|}} \tag{6}$$

where $B_{\tilde{\ell}} \geq |\ell(x,\theta)|$ for all $\ell \in \left\{\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m\right\}$, and $B_\ell \geq \max_{i\in[m]}(b_i - a_i)$ assuming that the range of each ℓ_i is the interval $[a_i, b_i]$.

Proof. The particular values we choose for T and η_λ come from Lemma 3, taking $r = \epsilon/2B_{\tilde{\ell}}$, $\epsilon_\theta = \rho + 2rB_{\tilde{\ell}} = \rho + \epsilon$, and $\epsilon_\lambda = \epsilon$. The optimality and feasibility results then follow from Theorem 3.

For the bound on $G^{(\mathrm{val})}(\hat{\Theta})$, notice that by Lemma 1, there exists a radius-r covering C_r w.r.t. the 1-norm with $|C_r| \leq (5/r)^m = \left(10B_{\tilde{\ell}}/\epsilon\right)^m$. Substituting this, and the definition of r, into the bound of Lemma 4 yields the claimed bound. □

E.2. Algorithm 2

Lemma 5. *Suppose that Θ is compact and convex, and that $\ell(x;\theta)$ is μ-strongly convex in θ for all $\ell \in \left\{\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m\right\}$. If we take $\eta_\lambda := \sqrt{(m+1)\ln(m+1)/T_\lambda B_\Delta^2}$, then the result of Algorithm 2 satisfies the conditions of Theorem 3 with:*

$$\epsilon_\theta = \frac{B_{\check{\Delta}}^2(1+\ln T_\theta)}{2\mu T_\theta}$$
$$\epsilon_\lambda = 2B_\Delta\sqrt{\frac{(m+1)\ln(m+1)}{T_\lambda}}$$

where $B_{\check{\Delta}} \geq \max_{s,t\in[T_\theta]\times[T_\lambda]}\left\|\check{\Delta}_\theta^{(t,s)}\right\|_2$ is a bound on the subgradients, and $B_\Delta \geq \max_{t\in[T_\lambda]}\left\|\Delta_\lambda^{(t)}\right\|_\infty$ is a bound on the gradients.

Proof. By Lemma 1 of Shalev-Shwartz et al. (2011), the fact that $\hat{\mathcal{L}}_\theta(\theta,\lambda)$ is μ-strongly convex in θ (because $\ell(x;\theta)$ is μ-strongly convex in θ for $\ell \in \left\{\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m\right\}$, and $\lambda \in \Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$), and Jensen's inequality:

$$\hat{\mathcal{L}}_\theta\left(\theta^{(t)},\lambda^{(t)}\right) \leq \frac{1}{T_\theta}\sum_{s=1}^{T_\theta}\hat{\mathcal{L}}_\theta\left(\tilde{\theta}^{(t,s)},\lambda^{(t)}\right) \leq \min_{\theta^*\in\Theta}\hat{\mathcal{L}}_\theta\left(\theta^*,\lambda^{(t)}\right) + \frac{B_{\check{\Delta}}^2(1+\ln T_\theta)}{2\mu T_\theta} \tag{13}$$

the above holding for all t. Therefore:

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{L}}_\theta\left(\theta^{(t)},\lambda^{(t)}\right) \leq \min_{\theta^*\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{L}}_\theta\left(\theta^*,\lambda^{(t)}\right) + \frac{B_\Delta^2\left(1+\ln T_\theta\right)}{2\mu T_\theta}$$

so the first condition of Theorem 3 is satisfied with the claimed ϵ_θ.

As in the proof of Lemma 3, the second condition, on ϵ_λ, follows immediately from Lemma 8 of Appendix C.1 of Cotter et al. (2019), taking $\tilde{m} = m+1$. □

Lemma 6. *In addition to the conditions of Lemma 5, suppose that $\ell(x;\theta)$ is L-Lipschitz continuous in θ for all $\ell \in \{\ell_1,\ldots,\ell_m\}$. If we take $\hat{\Theta} := \left\{\theta^{(1)},\ldots,\theta^{(T)}\right\}$ as in Theorem 3, where $\theta^{(1)},\ldots,\theta^{(T)}$ are the result of Algorithm 2, then with probability $1-\delta$ over the sampling of $S^{(\mathrm{val})}$:*

$$G^{(\mathrm{val})}(\hat{\Theta}) < 2L\sqrt{\frac{4rB_{\tilde{\ell}}}{\mu}} + \frac{2LB_{\tilde{\Delta}}}{\mu}\sqrt{\frac{1+\ln T_\theta}{T_\theta}} + B_\ell\sqrt{\frac{\ln\left(2m\left|C_r\right|/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}}$$

where $B_{\tilde{\ell}} \geq \left|\ell(x,\theta)\right|$ for all $\ell \in \left\{\ell_0,\tilde{\ell}_1,\ldots,\tilde{\ell}_m\right\}$, $B_\ell \geq \max_{i\in[m]}(b_i-a_i)$ assuming that the range of each ℓ_i is the interval $[a_i,b_i]$, and C_r is a radius-r covering of $\Lambda := \Delta^m \subseteq \mathbb{R}_+^{m+1}$ w.r.t. the 1-norm.

Proof. Define $\tilde{\lambda}^{(t)} := \mathrm{argmin}_{\tilde{\lambda}\in C_r}\left\|\lambda^{(t)}-\tilde{\lambda}\right\|$ for all t. Since C_r is a radius-r covering of $\Lambda := \Delta^m$ w.r.t. the 1-norm, we must have that $\left\|\tilde{\lambda}^{(t)}-\lambda^{(t)}\right\|_1 \leq r$, which implies by Definition 1 that:

$$\left|\hat{\mathcal{L}}_\theta\left(\theta,\tilde{\lambda}^{(t)}\right) - \hat{\mathcal{L}}_\lambda\left(\theta,\lambda^{(t)}\right)\right| \leq rB_{\tilde{\ell}}$$

the above holding for all $\theta\in\Theta$, and all t. Take $\theta^{(t*)} := \mathrm{argmin}_{\theta^*\in\Theta}\hat{\mathcal{L}}_\theta\left(\theta^*,\lambda^{(t)}\right)$ and $\tilde{\theta}^{(t*)} := \mathrm{argmin}_{\tilde{\theta}^*\in\Theta}\hat{\mathcal{L}}_\theta\left(\tilde{\theta}^*,\tilde{\lambda}^{(t)}\right)$. Then, by the above result and the triangle inequality:

$$\begin{aligned}\left|\hat{\mathcal{L}}_\theta\left(\theta^{(t*)},\lambda^{(t)}\right) - \hat{\mathcal{L}}_\theta\left(\tilde{\theta}^{(t*)},\tilde{\lambda}^{(t)}\right)\right| &\leq rB_{\tilde{\ell}}\\ \left|\hat{\mathcal{L}}_\theta\left(\theta^{(t*)},\tilde{\lambda}^{(t)}\right) - \hat{\mathcal{L}}_\theta\left(\tilde{\theta}^{(t*)},\tilde{\lambda}^{(t)}\right)\right| &\leq 2rB_{\tilde{\ell}}\end{aligned}$$

so by the fact that $\hat{\mathcal{L}}_\theta(\theta,\lambda)$ is μ-strongly convex in θ for all λ:

$$\left\|\theta^{(t*)}-\tilde{\theta}^{(t*)}\right\|_2 \leq \sqrt{\frac{4rB_{\tilde{\ell}}}{\mu}}$$

Again by strong convexity, but applied to Equation 13 in the proof of Lemma 5:

$$\left\|\theta^{(t)}-\theta^{(t*)}\right\|_2 \leq \sqrt{\frac{B_\Delta^2\left(1+\ln T_\theta\right)}{\mu^2 T_\theta}}$$

so by the triangle inequality:

$$\left\|\theta^{(t)}-\tilde{\theta}^{(t*)}\right\|_2 \leq \sqrt{\frac{4rB_{\tilde{\ell}}}{\mu}} + \sqrt{\frac{B_\Delta^2\left(1+\ln T_\theta\right)}{\mu^2 T_\theta}}$$

and by L-Lipschitz continuity:

$$\left|\ell_i\left(x;\theta^{(t)}\right) - \ell_i\left(x;\tilde{\theta}^{(t*)}\right)\right| \leq L\left(\sqrt{\frac{4rB_{\tilde{\ell}}}{\mu}} + \sqrt{\frac{B_\Delta^2\left(1+\ln T_\theta\right)}{\mu^2 T_\theta}}\right) \tag{14}$$

the above holding for all t, and all $i\in[m]$.

Define $\tilde{\Theta} := \left\{\tilde{\theta}^{(1*)}, \ldots, \tilde{\theta}^{(T*)}\right\}$. Observe that since $\tilde{\theta}^{(t*)}$ is uniquely associated with a $\tilde{\lambda}^{(t)} \in C_r$, we will have that $\tilde{\Theta} \subseteq \Theta_{C_r}$, where:

$$\Theta_{C_r} := \left\{ \operatorname*{argmin}_{\tilde{\theta}^* \in \Theta} \hat{\mathcal{L}}_\theta \left(\tilde{\theta}^*, \tilde{\lambda}\right) : \tilde{\lambda} \in C_r \right\}$$

Because the argmins defining Θ_{C_r} depend only on $\hat{\mathcal{L}}_\theta$, which itself depends only on $S^{(\mathrm{trn})}$, we can apply Lemma 2 to Θ_{C_r}, yielding that, for each $i \in [m]$, the following holds with probability δ/m for any $\theta \in \Theta_{C_r}$:

$$\left| \frac{1}{\left|S^{(\mathrm{val})}\right|} \sum_{x \in S^{(\mathrm{val})}} \ell_i\left(x;\theta\right) - \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i\left(x;\theta\right)\right] \right| < B_\ell \sqrt{\frac{\ln\left(2m\left|\Theta_{C_r}\right|/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}}$$

By the union bound, we could instead take the above to hold uniformly for all $i \in [m]$ with probability $1-\delta$. Substituting Equation 14 and using the facts that $\tilde{\Theta} \subseteq \Theta_{C_r}$ and $|\Theta_{C_r}| = |C_r|$ yields the claimed result. □

Theorem 2. *Suppose that Θ is compact and convex, and that $\ell\left(x;\theta\right)$ is μ-strongly convex in θ for all $\ell \in \left\{\ell_0, \tilde{\ell}_1, \ldots, \tilde{\ell}_m\right\}$. Given any $\epsilon > 0$, if we take $T_\theta \geq \left(B_{\tilde{\Delta}}^2/\mu\epsilon\right)\ln\left(B_{\tilde{\Delta}}^2/\mu\epsilon\right)$, $T_\lambda \geq 4B_\Delta^2\left(m+1\right)\ln\left(m+1\right)/\epsilon^2$ and $\eta_\lambda = \sqrt{\left(m+1\right)\ln\left(m+1\right)/T_\lambda B_\Delta^2}$, where B_Δ is as in Theorem 1 and $B_{\tilde{\Delta}} \geq \max_{s,t\in[T_\theta]\times[T_\lambda]} \left\|\check{\Delta}_\theta^{(t,s)}\right\|_2$ is a bound on the subgradients, then the following hold, where $\hat{\Theta} := \left\{\theta^{(1)}, \ldots, \theta^{(T_\lambda)}\right\}$ is the set of results of Algorithm 1.*

Optimality and Feasibility: *Let $\bar{\theta}$ be a random variable taking values from $\hat{\Theta}$, defined such that $\bar{\theta} = \theta^{(t)}$ with probability $\lambda_1^{(t)} / \sum_{s=1}^T \lambda_1^{(s)}$, and let $\bar{\lambda} := \left(\sum_{t=1}^T \lambda^{(t)}\right)/T_\lambda$. Then $\bar{\theta}$ is nearly-optimal in expectation:*

$$\mathbb{E}_{\bar{\theta},x\sim\mathcal{D}}\left[\ell_0\left(x;\bar{\theta}\right)\right] \leq \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0\left(x;\theta^*\right)\right] + \frac{1}{\bar{\lambda}_1}\left(2\epsilon + 2\tilde{G}^{(\mathrm{trn})}(\Theta) + G^{(\mathrm{val})}(\hat{\Theta})\right)$$

where θ^* minimizes $\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0\left(x;\cdot\right)\right]$ subject to the proxy-constraints $\mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{\ell}_i\left(x;\theta^*\right)\right] \leq 0$. It is also nearly-feasible:

$$\max_{i\in[m]} \mathbb{E}_{\bar{\theta},x\sim\mathcal{D}}\left[\ell_i\left(x;\bar{\theta}\right)\right] \leq \frac{\epsilon}{\bar{\lambda}_1} + G^{(\mathrm{val})}(\hat{\Theta})$$

Additionally, if there exists a $\theta' \in \Theta$ that satisfies all of the constraints with margin γ (i.e. $\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i\left(x;\theta'\right)\right] \leq -\gamma$ for all $i \in [m]$), then:

$$\bar{\lambda}_1 \geq \frac{1}{\gamma + B_{\ell_0}}\left(\gamma - 2\epsilon - 2\tilde{G}^{(\mathrm{trn})}(\Theta) - G^{(\mathrm{val})}(\hat{\Theta})\right)$$

where B_{ℓ_0} is as in Theorem 1.

Generalization: *If, in addition to the above requirements, $\ell\left(x;\theta\right)$ is L-Lipschitz continuous in θ for all $\ell \in \{\ell_1, \ldots, \ell_m\}$, then with probability $1-\delta$ over the sampling of $S^{(\mathrm{val})}$:*

$$G^{(\mathrm{val})}(\hat{\Theta}) < B_\ell\sqrt{\frac{2m}{\left|S^{(\mathrm{val})}\right|}\max\left\{1, \ln\left(\frac{160L^2B_{\tilde{\ell}}\left|S^{(\mathrm{val})}\right|}{m\mu B_\ell^2}\right)\right\}} + B_\ell\sqrt{\frac{\ln\left(2m/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}} + 2L\epsilon\sqrt{\frac{2}{\mu}} \tag{7}$$

where $B_{\tilde{\ell}}$ and B_ℓ are as in Theorem 1.

Proof. The particular values we choose for T_θ, T_λ and η_λ come from Lemma 5, taking $\epsilon_\theta = \epsilon_\lambda = \epsilon$. The optimality and feasibility results then follow from Theorem 3.

For the bound on $G^{(\mathrm{val})}(\hat{\Theta})$, notice that by Lemma 1, there exists a radius-r external covering C_r w.r.t. the 1-norm with $|C_r| \leq \max\left\{1, (5/r)^m\right\}$. Substituting into the bound of Lemma 6:

$$\begin{aligned} G^{(\mathrm{val})}(\hat{\Theta}) <& 2L\sqrt{\frac{4rB_{\tilde{\ell}}}{\mu}} + \frac{2LB_{\check{\Delta}}}{\mu}\sqrt{\frac{1+\ln T_\theta}{T_\theta}} + B_\ell\sqrt{\frac{m\max\left\{0,\ln\left(5/r\right)\right\} + \ln\left(2m/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}} \\ <& 2L\sqrt{\frac{4rB_{\tilde{\ell}}}{\mu}} + B_\ell\sqrt{\frac{m\max\left\{0,\ln\left(5/r\right)\right\}}{2\left|S^{(\mathrm{val})}\right|}} + \frac{2LB_{\check{\Delta}}}{\mu}\sqrt{\frac{1+\ln T_\theta}{T_\theta}} + B_\ell\sqrt{\frac{\ln\left(2m/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}} \end{aligned}$$

Taking $r := \left(m\mu B_\ell^2\right) / \left(32L^2 B_{\tilde{\ell}} \left|S^{(\mathrm{val})}\right|\right)$:

$$\begin{aligned} G^{(\mathrm{val})}(\hat{\Theta}) <& 2B_\ell \sqrt{\frac{m \max\left\{1, \ln\left(5/r\right)\right\}}{2\left|S^{(\mathrm{val})}\right|}} + \frac{2LB_{\check{\Delta}}}{\mu}\sqrt{\frac{1+\ln T_\theta}{T_\theta}} + B_\ell\sqrt{\frac{\ln\left(2m/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}} \\ <& B_\ell \sqrt{\frac{2m}{\left|S^{(\mathrm{val})}\right|} \max\left\{1, \ln\left(\frac{160L^2 B_{\tilde{\ell}}\left|S^{(\mathrm{val})}\right|}{m\mu B_\ell^2}\right)\right\}} \\ &+ B_\ell\sqrt{\frac{\ln\left(2m/\delta\right)}{2\left|S^{(\mathrm{val})}\right|}} + \frac{2LB_{\check{\Delta}}}{\mu}\sqrt{\frac{1+\ln T_\theta}{T_\theta}} \end{aligned}$$

Substituting the definition of T_θ then yields the claimed result. □

F. One-dataset Lagrangian Baseline Approach

In this appendix, we'll analyze the most natural theoretical baseline for our proposed approach, namely using a single training dataset, and optimizing the empirical Lagrangian:

$$\hat{\mathcal{L}}\left(\theta, \lambda\right) := \frac{1}{\left|S^{(\mathrm{trn})}\right|} \sum_{x \in S^{(\mathrm{trn})}} \left(\ell_0\left(x;\theta\right) + \sum_{i=1}^m \lambda_i \ell_i\left(x;\theta\right)\right) \tag{15}$$

This is essentially a minor extension of the approach proposed by Agarwal et al. (2018)—who proposed using the Lagrangian formulation in the particular case of fair classification—to the slightly more general setting of inequality constrained optimization.

Theorem 1 of Cotter et al. (2019) characterizes the optimality and feasibility properties of Nash equilibria of the in-expectation Lagrangian (Equation 2). The analogous result for the empirical Lagrangian of Equation 15 is given in the following theorem:

Theorem 4. *Define $\Lambda = \left\{\lambda \in \mathbb{R}_+^m : \left\|\lambda\right\|_1 \leq R\right\}$, let $\tilde{\ell}_i := \ell_i$ for all $i \in [m]$, and consider the empirical Lagrangian of Equation 15. Let $\theta^{(1)}, \ldots, \theta^{(T)} \in \Theta$ and $\lambda^{(1)}, \ldots, \lambda^{(T)} \in \Lambda$ be sequences satisfying:*

$$\max_{\lambda^* \in \Lambda} \frac{1}{T}\sum_{t=1}^T \hat{\mathcal{L}}\left(\theta^{(t)}, \lambda^*\right) - \inf_{\theta^* \in \Theta} \frac{1}{T}\sum_{t=1}^T \hat{\mathcal{L}}\left(\theta^*, \lambda^{(t)}\right) \leq \epsilon \tag{16}$$

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:

$$\mathbb{E}_{\bar{\theta}, x\sim\mathcal{D}}\left[\ell_0\left(x;\bar{\theta}\right)\right] \leq \inf_{\theta^* \in \Theta : \forall i. \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i\left(x;\theta^*\right)\right] \leq 0} \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0\left(x;\theta^*\right)\right] + \epsilon + 2\tilde{G}^{(\mathrm{trn})}(\Theta)$$

and nearly-feasible:

$$\max_{i\in[m]} \mathbb{E}_{\bar{\theta}, x\sim\mathcal{D}}\left[\ell_i\left(x;\bar{\theta}\right)\right] \leq \frac{\epsilon}{R - \left\|\bar{\lambda}\right\|_1} + \tilde{G}^{(\mathrm{trn})}(\Theta)$$

Additionally, if there exists a $\theta' \in \Theta$ that satisfies all of the constraints with margin γ (i.e. $\mathbb{E}_{x\sim\mathcal{D}}\left[\ell_i\left(x;\theta'\right)\right] \leq -\gamma$ for all $i \in [m]$), then:

$$\left\|\bar{\lambda}\right\|_1 \leq \frac{\epsilon + B_{\ell_0}}{\gamma - \tilde{G}^{(\mathrm{trn})}(\Theta)}$$

assuming that $\gamma > \tilde{G}^{(\mathrm{trn})}(\Theta)$, where $B_{\ell_0} \geq \sup_{\theta\in\Theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0\left(x;\theta\right)\right] - \inf_{\theta\in\Theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\ell_0\left(x;\theta\right)\right]$ is a bound on the range of the objective loss.

Proof. The empirical Lagrangian is nothing but the in-expectation Lagrangian over the finite training sample $S^{(\mathrm{trn})}$, so by Theorem 1 of Cotter et al. (2019), $\bar{\theta}$ is nearly-optimal in expectation:

$$\mathbb{E}_{\bar{\theta}, x\sim S^{(\mathrm{trn})}}\left[\ell_0\left(x;\bar{\theta}\right)\right] \leq \inf_{\theta^* \in \Theta : \forall i. \mathbb{E}_{x\sim S^{(\mathrm{trn})}}\left[\ell_i\left(x;\theta^*\right)\right] \leq 0} \mathbb{E}_{x\sim S^{(\mathrm{trn})}}\left[\ell_0\left(x;\theta^*\right)\right] + \epsilon$$

and nearly-feasible:

$$\max_{i\in[m]} \mathbb{E}_{\bar{\theta},x\sim S^{(\mathrm{trn})}}\left[\ell_i\left(x;\bar{\theta}\right)\right] \leq \frac{\epsilon}{R-\left\|\bar{\lambda}\right\|_1}$$

Since θ' satisfies the constraints with margin γ in expectation, it will satisfy them with margin $\gamma - \tilde{G}^{(\mathrm{trn})}(\Theta)$ on the training dataset, so the same theorem gives the claimed upper-bound $\left\|\bar{\lambda}\right\|_1 \leq (\epsilon + B_{\ell_0})/(\gamma - \tilde{G}^{(\mathrm{trn})}(\Theta))$ when $\gamma > \tilde{G}^{(\mathrm{trn})}(\Theta)$.

Notice that the above expectations are taken over the finite training sample $S^{(\mathrm{trn})}$, rather than the data distribution $\mathcal{D}$. To fix this, we need only define $\tilde{\ell}_i = \ell_i$, and appeal to the definition of $\tilde{G}^{(\mathrm{trn})}(\Theta)$ (Definition 2), yielding the claimed results. $\square$

Here, R is the maximum allowed 1-norm of the vector of Lagrange multipliers (such a bound is necessary for Cotter et al. (2019)'s proof to work out). Notice that we have assumed that $\tilde{\ell}_i := \ell_i$ for all i. This is purely for notational reasons (the Lagrangian does not involve proxy constraints *at all*)—it allows us to re-use the definition of $\tilde{G}^{(\mathrm{trn})}(\Theta)$ (Definition 2) in the above Theorem, and causes γ to have the same definition here, as it did in Theorems 1 and 2.