Probably Approximately Correct Constrained Learning

Luiz F. O. Chamon
Dept. of Electrical and Systems Engineering
University of Pennsylvania
Pennsylvania, USA
luizf@seas.upenn.edu

Alejandro Ribeiro
Dept. of Electrical and Systems Engineering
University of Pennsylvania
Pennsylvania, USA
aribeiro@seas.upenn.edu

Abstract

As learning solutions reach critical applications in social, industrial, and medical domains, the need to curtail their behavior has become paramount. There is now ample evidence that without explicit tailoring, learning can lead to biased, unsafe, and prejudiced solutions. To tackle these problems, we develop a generalization theory of constrained learning based on the probably approximately correct (PAC) learning framework. In particular, we show that imposing requirements does not make a learning problem harder in the sense that any PAC learnable class is also PAC *constrained* learnable using a constrained counterpart of the empirical risk minimization (ERM) rule. For typical parametrized models, however, this learner involves solving a constrained non-convex optimization program for which even obtaining a feasible solution is challenging. To overcome this issue, we prove that under mild conditions the empirical dual problem of constrained learning is also a PAC constrained learner that now leads to a practical constrained learning algorithm based solely on solving unconstrained problems. We analyze the generalization properties of this solution and use it to illustrate how constrained learning can address problems in fair and robust classification.

1 Introduction

Learning has become a core component of the modern information systems we increasingly rely upon to select job candidates, analyze medical data, and control "smart" applications (home, grid, city). As these systems become ubiquitous, so does the need to curtail their behavior. Left untethered, they can fail catastrophically as evidenced by the growing number of reports involving biased, prejudiced models or systems prone to tampering (e.g., adversarial examples), unsafe behaviors, and deadly accidents [1–6]. Typically, learning is constrained by using domain expert knowledge to either construct models that *embed* the required properties (see, e.g., [7–13]) or *tune* the training objective so as to promote them (see, e.g., [14–17]). The latter approach, known as regularization, is ubiquitous in practice even though it need not yield feasible solutions [18]. In fact, existing results from classical learning theory guarantee generalization with respect to the regularized objective, which says nothing about meeting the requirements it may describe [19, 20]. While the former approach guarantees that the solution satisfies the requirements, the scale and opacity of modern machine learning (ML) systems render this model design impractical.

Since ML models are often trained using empirical risk minimization (ERM), an alternative solution is to explicitly add constraints to these optimization problems. Since requirements are often expressed as constraints in the first place, this approach overcomes the need to tune regularization parameters. What it more, any solution automatically satisfies the requirements. Nevertheless, this approach suffers from two fundamental drawbacks. First, its involves solving a constrained optimization problem that is non-convex for typical parametrizations (e.g., neural networks). Though gradient descent can often be used to obtain good minimizers for differentiable models, it does not guarantee

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

constraint satisfaction. Indeed, there is typically no straightforward way to project onto the feasibility set (e.g., the set of fair classifiers) and strong duality need not hold for non-convex programs [18]. Second, even if we could solve this constrained ERM, the issue remains of how its solutions generalize since classical learning theory is involved only with unconstrained problems [19, 20].

In this work, we address these issues in two steps. We begin by formalizing the concept of constrained learning using the probably approximately correct (PAC) framework. We prove that any hypothesis class that is unconstrained learnable is constrained learnable and that the constrained counterpart of the ERM rule is a PAC constrained learner. Hence, we establish that, from a learning theoretic perspective, *constrained learning is as hard as unconstrained (classical) learning*. This, however, does not resolve the practical issue of learning under requirements due to the non-convexity of the constrained ERM problem. To do so, we proceed by deriving an empirical saddle-point problem that is a (representation-independent) PAC constrained learner. We show that its approximation error depends on the richness of the parametrization and the difficulty of satisfying the learning constraints. Finally, we put forward practical constrained learning algorithm that we use to illustrate how constrained learning can address problems involving fairness and robustness.

2 Related work

Central to ML is the concept of ERM in which statistical quantities are replaced by their empirical counterparts, thus allowing learning problems to be solved from data, without prior knowledge of its underlying distributions. The set of conditions under which this is a sensible approach is known in learning theory as (agnostic) PAC learnability. More generally, the PAC framework formalizes what it means to solve a statistical learning problem and studies when it can be done [19–22]. While different learning models, such as structured complexity and PAC-Bayes, have been proposed, they are beyond the scope of this work.

The objects studied in (PAC) learning theory, however, are unconstrained statistical learning problem. Yet, there is a growing need to enable learning under constraints to tackle problems in fairness [23–29], robustness [30–32], safety [33–37], and semi-supervised learning [38–40], to name a few. While constraints have been used in statistics since Neyman-Pearson [41], generalization guarantees for constrained learning have been studied only in specific contexts, e.g., for coherence constraints or rate-constrained learning [23, 25, 29, 42]. Additionally, due to the non-convexity of typical learning problems, many of these results hold for randomized solutions, e.g., [23, 25, 27, 29]. In contrast, this work puts forward a formal constrained learning framework in which generalization results are derived for deterministic learners. A first step in that direction was taken in [43], albeit from an optimization perspective. This work also accounts for pointwise constraints, fundamental in the context of fairness, and provides a practical, guaranteed constrained learning algorithm (Sec. 5.2).

Due to these challenges, learning under requirements is often tackled using regularization, i.e., by integrating a fixed cost for violating the constraints into the training objective (see, e.g., [15–17, 31, 44, 45]). Selecting these costs, however, can be challenging, especially as the number of constraints grows. In fact, their values often depend on the problem instance, the objective value, and can interact in non-trivial ways [46–50]. In the case of convex optimization problems, a straightforward relation between constraints and regularization costs can be obtained due to strong duality. A myriad of primal-dual methods can then be used to obtain optimal, feasible solutions [51]. However, most modern parametrizations (e.g., CNNs) lead to non-convex programs for which a regularized formulation need not yield feasible solutions, all the more so good ones [18]. While primal-dual algorithms have been used in practice, no guarantees can be given for their outcome in general [30, 32, 52, 53].

3 Constrained Learning

Let $\mathfrak{D}_i$, $i = 0, \dots, m+q$, denote *unknown* probability distributions over the space of data pairs $(\boldsymbol{x}, y)$, with $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$. For a hypothesis class $\mathcal{H}$ of functions $\phi : \mathcal{X} \to \mathbb{R}^k$, define the

generic constrained statistical learning (CSL) problem as

$$\begin{aligned} P^\star = \min_{\phi\in\mathcal{H}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_0}\Big[\ell_0\big(\phi(\boldsymbol{x}),y\big)\Big] \\ \text{subject to} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_i}\Big[\ell_i\big(\phi(\boldsymbol{x}),y\big)\Big] \le c_i, \quad i=1,\dots,m, \\ & \ell_j\big(\phi(\boldsymbol{x}),y\big) \le c_j \quad \mathfrak{D}_j\text{-a.e.}, \qquad j=m+1,\dots,m+q, \end{aligned} \tag{P-CSL}$$

where $\ell_i : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ are performance metrics. In general, we think of $\mathfrak{D}_0$ as a nominal joint distribution over data pairs $(\boldsymbol{x}, y)$ corresponding to feature vectors $\boldsymbol{x}$ and responses y. The additional $\mathfrak{D}_i$ can be used to model different conditional distributions over which requirements are imposed either on average, through the losses ℓ_i, $i \le m$, or pointwise, through the losses ℓ_j, $j > m$. Note that the unconstrained version of (P-CSL), namely

$$P_{\text{U}}^\star = \min_{\phi\in\mathcal{H}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_0}\Big[\ell_0\big(\phi(\boldsymbol{x}),y\big)\Big], \tag{PI}$$

is at the core of virtually all of modern ML [20, 54].

Before tackling *if* and *how* we can learn under constraints, i.e., whether we can solve (P-CSL), we illustrate *what* constrained learning can enable. To make the discussion concrete, we present two constrained formulations of the learning problems we solve in Section 6.

Invariance and fair learning. Constrained learning is a natural way to formulate learning problems in which invariance is required. Consider a model ϕ whose output is a discrete distribution over k possible classes. Then, (P-CSL) can be used to write

$$\begin{aligned} \underset{\phi\in\mathcal{H}}{\text{minimize}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\ell_0\big(\phi(\boldsymbol{x}),y\big)\Big] \\ \text{subject to} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\,\mathrm{D}_{\mathrm{KL}}\Big(\phi(\boldsymbol{x}) \,\|\, \phi\big(\rho(\boldsymbol{x})\big)\Big)\Big] \le c, \end{aligned} \tag{PII}$$

where ρ is an input transformation we wish the model to be invariant to and $c > 0$ determines the sensitivity level. Formulation (PII) can be extended trivially to multiple transformations (see Sec. 6). When the average invariance in (PII) is not enough, a stricter, pointwise requirement can be imposed, by using

$$\mathrm{D}_{\mathrm{KL}}\Big(\phi(\boldsymbol{x}) \,\|\, \phi\big(\rho(\boldsymbol{x})\big)\Big) \le c \quad \mathfrak{D}\text{-a.e.}. \tag{1}$$

For instance, fairness can be seen as a form of invariance in which ρ induces an alternative distribution of a certain protected variable (e.g., a gender change) [23–26, 28, 29]. In this case, the constraint in (PII) is related to the average causal effect (ACE) and (1) to *counterfactual fairness* [24]. While fairness goes beyond invariance, our goal is not to litigate the merit of any fairness metrics, but to show how constrained learning may provide a natural way to encode them.

Robust learning. Another issue affecting ML models, especially CNNs, is robustness. It is straightforward to construct small input perturbations that lead to misclassification and there are now numerous methods to do so. While adversarial training has empirically been shown to improve robustness, it often results in classifiers with poor nominal performance [30, 31, 44, 52, 53, 55]. In [32], a constrained formulation involving an upper bound on the worst-case error was used to tackle this issue. Similarly, we can address this compromise using (P-CSL) by writing

$$\begin{aligned} \underset{\phi\in\mathcal{H}}{\text{minimize}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\ell_0\big(\phi(\boldsymbol{x}),y\big)\Big] \\ \text{subject to} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{A}}\Big[\ell_0\big(\phi(\boldsymbol{x}),y\big)\Big] \le c \end{aligned} \tag{PIII}$$

where $\mathfrak{A}$ is an adversarial data distributions. What is more, we can soften the worst-case requirements of robust optimization by taking $\mathfrak{A} \mid \varepsilon$ to be a distribution of adversarials with perturbation at most ε and pose a prior on ε (e.g., an exponential). This results in classifiers whose performance degrades smoothly with the perturbation magnitude. The theory and algorithms developed in this work give generalization guarantees on solutions of this problem obtained using samples of $\mathfrak{A}$, which can be accessed based on, e.g., adversarial attacks (Sec. 6). In other words, it establishes conditions under which a classifier that is accurate and robust during training is also accurate and robust during testing.

4 Probably Approximately Correct Constrained Learning

While (P-CSL) clearly addresses many of the issues discussed in Sec. 1, we cannot expect to solve it exactly without access to the $\mathfrak{D}_i$ against which expectations are evaluated. Additionally, solving the variational (P-CSL) is challenging unless $\mathcal{H}$ is finite. In this section, we address the first matter by settling, as in classical learning theory, on obtaining a *good enough* solution (Sec. 4.1). We then show that these solutions are not "harder" to get in constrained learning than they were in unconstrained learning (Sec. 4.2). We then proceed to tackle the algorithmic challenges by deriving and analyzing a practical constrained learning algorithm (Sec. 5.2).

4.1 From PAC to PACC

Let us begin by defining what it means to learn under constraints. To do so, we start by looking at the unconstrained case, which is addressed in learning theory under the PAC framework [19–22].

Definition 1 (PAC learnability)**.** *A hypothesis class $\mathcal{H}$ is* (agnostic) probably approximately correct (PAC) *learnable if for every $\epsilon, \delta \in (0,1)$ and every distribution $\mathfrak{D}_0$, a $\phi^\dagger \in \mathcal{H}$ can be obtained from $N \geq N_{\mathcal{H}}(\epsilon, \delta)$ samples of $\mathfrak{D}_0$ such that $\mathbb{E}\left[\ell_0\big(\phi^\dagger(\boldsymbol{x}), y\big)\right] \leq P_{\mathrm{U}}^\star + \epsilon$ with probability $1-\delta$.*

A classical result states that $\mathcal{H}$ is PAC learnable if and only if it has finite VC dimension and that the $\phi^\dagger$ from Def. 1 can be obtained by solving an ERM problem [19, 20]. This is, however, not enough to enable constrained learning since a PAC $\phi^\dagger$ may not be feasible for (P-CSL). In fact, feasibility often takes priority over performance in constrained learning problems. For instance, regardless of how good a fair classifier is, it serves no "fair" purpose in practice unless it meets fairness requirements [see, e.g., (PII)]. These observations lead us to the following definition.

Definition 2 (PACC learnability)**.** *A hypothesis class $\mathcal{H}$ is* probably approximately correct constrained (PACC) *learnable if for every $\epsilon, \delta \in (0,1)$ and every distribution $\mathfrak{D}_i$, $i = 0, \dots, m+q$, a $\phi^\dagger \in \mathcal{H}$ can be obtained based $N \geq N_{\mathcal{H}}(\epsilon, \delta)$ samples from each $\mathfrak{D}_i$ such that it is, with probability $1-\delta$,*

1) approximately optimal, i.e.,

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_0}\left[\ell_0\big(\phi^\dagger(\boldsymbol{x}), y\big)\right] \leq P^\star + \epsilon \quad \textit{and} \tag{2}$$

2) approximately feasible, i.e.,

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_i}\left[\ell_i\big(\phi^\dagger(\boldsymbol{x}), y\big)\right] \leq b_i + \epsilon, \qquad i = 1, \dots, m, \tag{3a}$$

$$\ell_j\big(\phi^\dagger(\boldsymbol{x}), y\big) \leq b_j, \ \textit{for all } (\boldsymbol{x}, y) \in \mathcal{K}_j, \quad j = m+1, \dots, m+q, \tag{3b}$$

where $\mathcal{K}_j \subseteq \mathcal{X} \times \mathcal{Y}$ are sets of $\mathfrak{D}_j$ measure at least $1-\epsilon$.

Note that every PACC learnable class is also PAC learnable since it satisfies (2). However, a PACC learner must also meet the probably approximate feasibility conditions in (3). The additional "C" in PACC is used to remind ourselves of this fact. Next, we show that the converse is also true, i.e., that PAC and PACC learning are equivalent problems.

4.2 PACC Learning is as Hard as PAC Learning

Having formalized what we mean by constrained learning (Sec. 4.1), we turn to the issue of when it can be done. To do so, we follow the unconstrained learning lead and put forward an empirical constrained risk minimization (ECRM) rule using N_i samples $(\boldsymbol{x}_{n_i}, y_{n_i}) \sim \mathfrak{D}_i$, namely

$$\begin{aligned} \hat{P}^\star = \min_{\phi\in\mathcal{H}} \quad & \frac{1}{N_0}\sum_{n_0=1}^{N_0} \ell_0\big(\phi(\boldsymbol{x}_{n_0}), y_{n_0}\big) \\ \text{subject to} \quad & \frac{1}{N_i}\sum_{n_i=1}^{N_i} \ell_i\big(\phi(\boldsymbol{x}_{n_i}), y_{n_i}\big) \leq c_i, \qquad i = 1, \dots, m \\ & \ell_j\big(\phi(\boldsymbol{x}_{n_j}), y_{n_j}\big) \leq c_j, \text{ for all } n_j, \quad j = m+1, \dots, m+q. \end{aligned} \tag{P-ECRM}$$

Notice that (P-ECRM) is a constrained version of the classical ERM problem that is ubiquitous in the solution of unconstrained learning problems [20, 54]. The next theorem shows that, under mild assumptions on the losses, if $\mathcal{H}$ is PAC learnable, then it is PACC learnable using (P-ECRM).

Theorem 1. *Let the ℓ_i, $i = 0, \dots, m+q$, be bounded on $\mathcal{X}$. The hypothesis class $\mathcal{H}$ is PACC learnable if and only if it is PAC learnable and* (P-ECRM) *is a PACC learner of $\mathcal{H}$. Explicitly, let $d_{\mathcal{H}} < \infty$ be the VC dimension of $\mathcal{H}$. If $N_i \geq C\zeta^{-1}(\epsilon, \delta, d_{\mathcal{H}})$, $i = 0, \dots, m+q$, for an absolute constant C and*

$$\zeta^{-1}(\epsilon, \delta, d) = \frac{d + \log(1/\delta)}{\epsilon^2}, \tag{4}$$

then any solution $\hat{\phi}^\star$ of (P-ECRM) *is a PACC solution of* (P-CSL).

Proof. See Appendix A in the extended version [56]. □

Theorem 1 shows that, from a learning theoretic point-of-view, constrained learning is as hard as unconstrained learning. Not only that, but notice the sample complexity of constrained described by (4) matches that of PAC learning [19, 20]. It is therefore not surprising that a constrained version of ERM is a PACC learner. A similar result appeared in [26] for a particular rate constraint and not in the context of PACC learning. Still, solving (P-ECRM) remains challenging. Indeed, while it addresses the statistical issue of (P-CSL), it remains, in most practical cases, an infinite dimensional (functional) problem. This issue is often addressed by leveraging a finite dimensional parametrization of (a subset of) $\mathcal{H}$, such as a kernel model or a (C)NN. Explicitly, we associate to each parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ a function $f_{\boldsymbol{\theta}} \in \mathcal{H}$, replacing (P-ECRM) by

$$\begin{aligned} \hat{P}_\theta^\star = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \quad & \frac{1}{N_0} \sum_{n_0=1}^{N_0} \ell_0\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n_0}), y_{n_0}\big) \\ \text{subject to} \quad & \frac{1}{N_i} \sum_{n_i=1}^{N_i} \ell_i\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n_i}), y_{n_i}\big) \leq c_i, \qquad i = 1, \dots, m \\ & \ell_j\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n_j}), y_{n_j}\big) \leq c_j, \text{ for all } n_j, \quad j = m+1, \dots, m+q. \end{aligned} \tag{PIV}$$

Even if (P-ECRM) is a convex program in ϕ, (PIV) typically is not a convex program in $\boldsymbol{\theta}$ (except, e.g., if the losses are convex and $f_{\boldsymbol{\theta}}$ is linear in $\boldsymbol{\theta}$). This issue also arises in unconstrained learning problems, but is exacerbated by the presence of constraints. Though it is sometimes possible to find good approximate minimizers of ℓ_0 using, e.g., gradient descent rules [57–61], even obtaining a feasible $\boldsymbol{\theta}$ may be challenging. Indeed, although good CNN classifiers can be trained using gradient descent, obtaining a good *fair/robust* classifier is considerably harder. Regularized formulations are often used to sidestep this issue by incorporating a linear combination of the constraints into the objective and solving the resulting unconstrained problem [15–17, 31, 44, 45]. Nevertheless, whereas the generalization guarantees of classical learning theory apply to this modified objective, they say nothing of the requirements it describes. Since strong duality need not hold for the non-convex (PIV), this procedure need not be PACC (Def. 2) and may lead to solutions that are either infeasible or whose performance is unacceptably poor [18].

While no formal connection can be drawn between (PIV) and its regularized formulation (due to the lack of strong duality [18]), its dual problem turns out to be related to (P-CSL). In the sequel, we prove that it provides (near-)PACC solutions for (P-CSL) with an approximation error in (2) that depends on the richness of the parametrization and how strict the learning constraints are (Sec. 5.1). In fact, we show that it is a (near-)PACC learner even if the parametrization is PAC learnable but $\mathcal{H}$ is not. Based on this result, we obtain a practical constrained learning algorithm (Sec. 5.2) that we use to solve the problems formulated in Sec. 3.

5 A (Near-)PACC Learning Algorithm

In this section, we derive a practical constrained learning algorithm by first analyzing the dual problem of (PIV) (Sec. 5.1) and then proposing an algorithm to solve it (Sec. 5.2). Although we know this dual problem is not related to (PIV), we prove that it is related directly to the original constrained learning problem (P-CSL) by showing it is a PACC learner except for an approximation error determined by the quality of the parametrization. We formalize this concept as follows:

Definition 3 (Near-PACC learnability)**.** *A class $\mathcal{H}$ is* (near-)PACC *learnable through a class $\mathcal{P}$ if there exists an $\epsilon_0 > 0$ such that for every $\epsilon, \delta \in (0, 1)$ and every distribution $\mathfrak{D}_i$, $i = 0, \dots, m+q$,*

an approximately feasible $\phi^\dagger \in \mathcal{P}$ [viz. (3)] can be obtained with probability $1-\delta$ based on $N \geq N_{\mathcal{P}}(\epsilon,\delta)$ samples from each $\mathfrak{D}_i$ and $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_0}\left[\ell_0\big(\phi^\dagger(\boldsymbol{x}),y\big)\right] \leq P^\star + \epsilon_0 + \epsilon$

In Def. 3, ϵ_0 characterizes the *approximation error*. In contrast to unconstrained learning, however, this error cannot be separated from the learning problem due to the constraints. Still, it is *fixed*, i.e., it is independent of the sample set, and affects neither the sample complexity nor the constraint satisfaction. Hence, the parametrized constrained learner sacrifices optimality, but not feasibility, which remains dependent only on the number of samples N (Def. 2). Finally, observe that the sample complexity does not depend on the original hypothesis class $\mathcal{H}$, but on the parametrized $\mathcal{P}$. Near-PACC is therefore related to representation-independent learning [20].

5.1 The Empirical Dual Problem of (P-CSL)

We begin by analyzing the gap between (P-CSL) and its (parametrized) empirical dual problem. Define the (parametrized) empirical Lagrangian of (P-CSL) as

$$\begin{aligned}\hat{L}(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\lambda}_j) = \frac{1}{N_0}\sum_{n_0=1}^{N_0}\ell_0\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n_0}),y_{n_0}\big) + \sum_{i=1}^{m}\mu_i\left[\frac{1}{N_i}\sum_{n_i=1}^{N_i}\ell_i\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n_i}),y_{n_i}\big) - c_i\right] \\ + \sum_{j=m+1}^{m+q}\left[\frac{1}{N_j}\sum_{n_j=1}^{N_j}\lambda_{j,n_j}\left(\ell_j\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n_j}),y_{n_j}\big) - c_j\right)\right],\end{aligned} \tag{5}$$

where $\boldsymbol{\mu} \in \mathbb{R}_+^m$ collects the dual variables μ_i relative to the average constraints and $\boldsymbol{\lambda}_j \in \mathbb{R}_+^{N_j}$ collects the dual variables λ_{j,n_j} relative to the j-th pointwise constraint. The empirical dual problem of (P-CSL) is then written as

$$\hat{D}^\star = \max_{\boldsymbol{\mu}\in\mathbb{R}_+^m,\ \boldsymbol{\lambda}_j\in\mathbb{R}_+^{N_j}}\ \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\hat{L}(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\lambda}_j), \tag{$\widehat{\text{D}}$-CSL}$$

Note that ($\widehat{\text{D}}$-CSL) is the dual problem of the parametrized ECRM (PIV). However, due to its non-convexity, its holds only that $\hat{D}^\star \leq \hat{P}_\theta^\star$ and, in general, a saddle-point of ($\widehat{\text{D}}$-CSL) is not related to a solution of (PIV) [18]. Still, ($\widehat{\text{D}}$-CSL) can be related directly to (P-CSL), which is why we refer to it as its empirical dual. This relation obtains under the following assumptions:

Assumption 1. The losses $\ell_i(\cdot,y)$, $i=0,\dots,m+q$, are $[0,B]$-valued, M-Lipschitz, convex functions for all $y\in\mathcal{Y}$. The loss ℓ_0 is additionally strongly convex.

Assumption 2. The hypothesis class $\mathcal{H}$ is convex, the parametrized $\mathcal{P} = \{f_{\boldsymbol{\theta}} \mid \boldsymbol{\theta}\in\mathbb{R}^p\} \subseteq \mathcal{H}$ is PAC learnable, and there is $\nu > 0$ such that for each $\phi \in \mathcal{H}$ there exists $f_{\boldsymbol{\theta}} \in \mathcal{P}$ for which $\sup_{\boldsymbol{x}\in\mathcal{X}}|f_{\boldsymbol{\theta}}(\boldsymbol{x}) - \phi(\boldsymbol{x})| \leq \nu$.

Assumption 3. There exists $\boldsymbol{\theta}' \in \mathbb{R}^p$ such that $f_{\boldsymbol{\theta}'}$ is strictly feasible for (P-CSL) with constraints $c_i - M\nu$ and $c_j - M\nu$ and for each datasets $\mathcal{S} = \big\{(\boldsymbol{x}_{n_i},y_{n_i})\big\}_{i=0,\dots,m+q}$ there exists a $\boldsymbol{\theta}''$ that is strictly feasible for (PIV).

In contrast to the unconstrained learning setting or the ECRM result in Theorem 1, we require that the losses ℓ_i and the hypothesis class $\mathcal{H}$ be convex. This, however, does not imply that ($\widehat{\text{D}}$-CSL) or (PIV) are convex problems since $\ell_i(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y)$ need not be convex in $\boldsymbol{\theta}$. Additionally, only the parametrized class $\mathcal{P}$ is required to be PAC learnable. Hence, $\mathcal{H}$ can be the space of continuous functions or a reproducing kernel Hilbert space (RKHS) and $f_{\boldsymbol{\theta}}$ can be a neural network [62–64] or a finite linear combinations of kernels [65, 66], both of which meet the uniform approximation assumption. This assumption can also be relaxed in the absence of pointwise constraints (Remark 1). Assumption 3 guarantees that the problem is well-posed, i.e., a feasible solution for (P-CSL) can be found in $\mathcal{P}$.

The main result of this section is collected in the following theorem.

Theorem 2. *Let $d_{\mathcal{P}}$ be the VC dimension of $\mathcal{P}$. Under Assumptions 1–3,* ($\widehat{\text{D}}$-CSL) *is a near-PACC learner of $\mathcal{H}$ with $N_{\mathcal{P}} = C\zeta^{-1}(\epsilon,\delta,d_{\mathcal{P}})$, for an absolute constant C and ζ^{-1} as in* (4)*, and*

$$\epsilon_0 = \left(1 + \left\|\boldsymbol{\mu}_p^\star\right\|_1 + \left\|\boldsymbol{\lambda}_p^\star\right\|_{L_1}\right) M\nu, \tag{6}$$

where $(\boldsymbol{\mu}_p^\star,\boldsymbol{\lambda}_p^\star)$ are dual variables of (P-CSL) *with constraints $c_i - M\nu$ for $i=1,\dots,m+q$.*

Algorithm 1 Primal-dual near-PACC learner

1: *Initialize*: $\boldsymbol{\theta}^{(0)} = 0$, $\boldsymbol{\mu}^{(0)} = \mathbb{1}$, $\boldsymbol{\lambda}_j^{(0)} = \mathbb{1}$

2: **for** $t = 1, \dots, T$

3: Obtain $\boldsymbol{\theta}^{(t-1)}$ such that $\hat{L}\Big(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}\Big) \leq \min\limits_{\boldsymbol{\theta} \in \mathbb{R}^p} \hat{L}\Big(\boldsymbol{\theta}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\lambda}_j^{(t-1)}\Big) + \rho$

4: Update dual variables

$$\mu_i^{(t)} = \left[\mu_i^{(t-1)} + \eta\left(\frac{1}{N_i}\sum_{n_i=1}^{N_i} \ell_i\big(f_{\boldsymbol{\theta}^{(t-1)}}(\boldsymbol{x}_{n_i}), y_{n_i}\big) - c_i\right)\right]_+$$

$$\lambda_{j,n_j}^{(t)} = \left[\lambda_{j,n_j}^{(t-1)} + \frac{\eta}{N_j}\Big(\ell_j\big(f_{\boldsymbol{\theta}^{(t-1)}}(\boldsymbol{x}_{n_j}), y_{n_j}\big) - c_j\Big)\right]_+$$

5: **end**

Proof. See Appendix B in the extended version [56]. □

Thus, the approximation error incurred by using the parametrization $f_{\boldsymbol{\theta}}$ is affected by (i) the difficulty of the learning problem and (ii) the richness of the parametrization. Indeed, under Assumptions 1–3, (P-CSL) is a strongly dual functional problem whose dual variables have a well-known sensitivity interpretation [67, Sec. 5.6]. So the bracketed quantity in (6) quantifies how stringent the learning constraints are in terms of how much performance could be gained by relaxing them. In addition, ϵ_0 is affected by the approximation capability ν of the parametrization. Since better parametrizations typically involve more parameters, which in turn affects the VC dimension of $\mathcal{P}$, a typical compromise between the approximation error and complexity arises. For small sample sets, the generalization error in Def. 3 is dominated by the estimation error ϵ, which improves for lower complexity classes. If there is abundance of data or the learning requirements are particularly stringent, the approximation error ϵ_0 dominates and more accurate, even if more complex, parametrizations should be used.

Note that the dual variables $(\boldsymbol{\mu}_p^\star, \boldsymbol{\lambda}_p^\star)$ may be hard to evaluate since they are related to a version of the statistical problem (P-CSL). While their norms can be estimated using classical results from optimization theory (see, e.g., [18, 68]), they often lead to loose, uninformative bounds. Notice, however, that only ϵ depends on the sample size.

Remark 1. When the constrained learning problem has no pointwise constraints [$q = 0$ in (P-CSL)], Assumption 2 can be relaxed from a uniform to a total variation approximation. Explicitly, Theorem 1 holds if for each $\phi \in \mathcal{H}$ there exist $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}_i}\big[|f_{\boldsymbol{\theta}}(\boldsymbol{x}) - \phi(\boldsymbol{x})|\big] \leq \nu$ for all i.

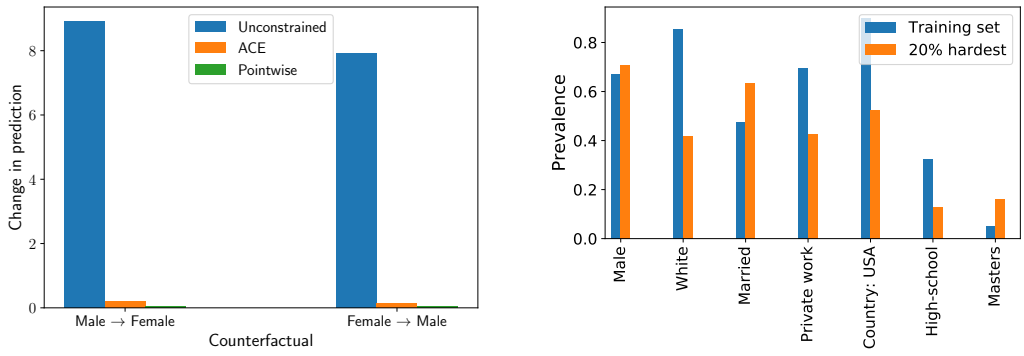
5.2 A Primal-Dual near-PACC Learner

We now proceed to introduce a practical algorithm to solve ($\widehat{\text{D}}$-CSL) based on a (sub)gradient primal-dual method. To do so, start by noting that the outer maximization is a convex optimization program. Indeed, the dual function $\hat{d}(\boldsymbol{\mu}, \boldsymbol{\lambda}_j) = \min_{\boldsymbol{\theta}} \hat{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ is the pointwise minimum of a set of affine functions and is therefore always concave [18]. Additionally, its (sub)gradients can be easily computed by evaluating the constraint slacks at the minimizer of $\hat{L}$ [51, Ch. 3]. Hence, the main challenge in ($\widehat{\text{D}}$-CSL) is the inner minimization.

Despite the Lagrangian (5) often being non-convex in $\boldsymbol{\theta}$, ($\widehat{\text{D}}$-CSL) is an unconstrained optimization problem. Hence, contrary to (PIV), it is often the case that good minimizers can be found, especially for differentiable losses and parametrizations (i.e., most common ML models). For instance, there is ample empirical and theoretical evidence that gradient descent can learn to good parameters for (C)NNs [57–61]. In that vein, we thus assume that we have access to the following oracle:

Assumption 4. There exists an oracle $\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ and $\rho > 0$ such that $\hat{L}\big(\boldsymbol{\theta}^\dagger(\boldsymbol{\mu}, \boldsymbol{\lambda}_j), \boldsymbol{\mu}, \boldsymbol{\lambda}_j\big) \leq \min_{\boldsymbol{\theta}} \hat{L}\big(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}_j\big) + \rho$ for all $\boldsymbol{\mu} \in \mathbb{R}_+^m$ and $\boldsymbol{\lambda}_j \in \mathbb{R}_+^{N_j}$, $j = m+1, \dots, m+q$.

Assumption 4 essentially states that we are able to (approximately) train regularized unconstrained learners using the parametrization $f_{\boldsymbol{\theta}}$. We can alternate between minimizing the Lagrangian (5) with



(a) (b)

Figure 1: Fair classification (Adult dataset): (a) classifier sensitivity and (b) prevalence of different groups among the 20% training set examples with largest dual variables.

respect to $\boldsymbol{\theta}$ for fixed $(\boldsymbol{\mu}, \boldsymbol{\lambda}_j)$ and updating the dual variables using the resulting minimizer. This procedure is summarized in Algorithm 1 and analyzed in the following theorem:

Theorem 3. *Fix $\beta > 0$ and consider Algorithm 1 with at least $C\zeta^{-1}(\epsilon, \delta, d_{\mathcal{P}})$ samples from each $\mathfrak{D}_j$, where C is an absolute constant, ζ^{-1} is as in* (4)*, and $d_{\mathcal{P}}$ is the VC dimension of $\mathcal{P}$. Under Assumptions 1–4, Algorithm 1 converges to the neighborhood*

$$P^\star - \rho - \beta - \eta S - \epsilon \leq \hat{L}\Big(\boldsymbol{\theta}^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\lambda}_j^{(T)}\Big) \leq P^\star + \rho + \epsilon_0 + \epsilon \tag{7}$$

with probability $1 - \delta$ after at most $T = O(1/\beta)$ for ϵ_0 as in (6) *and $S = O\big(B^2\big)$.*

Proof. See Appendix C in the extended version [56]. □

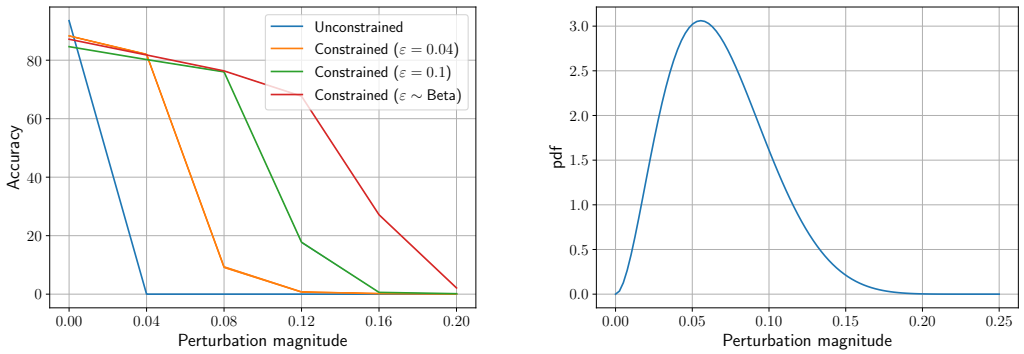
Theorem 3 bounds the suboptimality of Algorithm 1 with respect to the original learning problem (P-CSL). The size of this neighborhood depends polynomially on ϵ_0, ϵ, the oracle quality ρ, and the step size η. The number of iterations needed to reach this neighborhood is inversely proportional to the desired accuracy β. It is worth noting that this result applies to the deterministic outputs $(\boldsymbol{\theta}^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\lambda}_j^{(T)})$ of Algorithm 1 after convergence and not to a randomized solution obtained by sampling from $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\lambda}_j^{(t)})$, $t = 0, \ldots, T$ as in [23, 25, 29].

Underlying the oracle in Assumption 4 is often an iterative procedure, e.g., gradient descent, and the cost of running this procedure until convergence to obtain an approximate minimizer can be prohibitive. A common option then is to alternately update the primal variable $\boldsymbol{\theta}^{(t)}$ and the dual variables $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\lambda}_j^{(t)})$. This primal-dual method leads in fact to a classical convex optimization algorithm [69]. While the convergence guarantee of Theorem 3 no longer holds in this case, we observe good results by performing the primal and dual updates at different timescales, e.g., by performing step 3 once per epoch. This is exactly what we do in the next section where we illustrate the usefulness of this constrained learner.

6 Numerical experiments

Due to space constraints, we only provide highlights of the results obtained for the problems from Section 3. For more details and additional experiments, see Appendix D in the extended version [56].

Invariance and fair learning. In the Adult dataset [70], our goal is to predict whether an individual makes more than US$ 50,000.00 while being insensitive to gender. If left unconstrained, a small, one-hidden layer NN would change predictions on around 8% of the test samples had their genders been reversed (Fig. 1a). For step 3 of Algorithm 1, we use ADAM [71] with batch size 128 and learning rate 0.1. All other parameters were kept as in the original paper. After each epoch, we update the dual variables (step 4), also using ADAM with a step size of 0.01. All classifiers were trained over 300 epochs.



(a) (b)

Figure 2: Robust constrained learning (FMNIST): (a) Accuracy of classifiers under the PGD attack for different perturbation magnitudes and (b) distribution of ε used during training.

When constrained using the pointwise (1), the classifier becomes insensitive to the protected variable in over 99% of the test set. In such simple cases, invariant classifiers can be easily obtained by masking the training samples, although it can bring fairness issues of its own [26, 72]. But Algorithm 1 provides more than an invariant classifier. Due to the bound on the duality gap between (P-CSL) and ($\widehat{\text{D}}$-CSL), the dual variables have a sensitivity interpretation: the larger their value, the harder the constraint is to satisfy [18]. If we analyze the 20% of individuals with largest λ_n (Fig. 1b), we find that a significantly higher prevalence of non-white, non-US natives, married individuals. Clearly, while attempting to control for gender invariance, the constrained learner also had to overcome other prejudices correlated to sexism, a well-known challenge in fair classification [27]. Similar results can be derived when controlling for racial bias in the COMPAS dataset.

Robust learning. In this illustration, we use Algorithm 1 to train a ResNet18 [73] to classify images from the FMNIST dataset [74]. As in the previous example, we once again use the ADAM optimizer with the settings from [71]. The best accuracy over the validation set is achieved after 67 epochs, yielding a solution with test accuracy of 93.5% (Figure 2a). However, it fails to classify any of the test images when perturbed using a PGD attack with perturbation magnitude (ℓ_∞-norm of the perturbation) as low as $\varepsilon = 0.04$ [30]. The attack uses a step size of $\varepsilon/30$ for 50 iterations and we show the worst result over 10 restarts.

To overcome this issue, we use PGD to sample from a hypothetical "adversarial distribution" $\mathfrak{A}$ and constrain the performance of the solution against $\mathfrak{A}$ as in (PIII). To accelerate training, we use a much weaker attack running PGD without restarts for only 5 steps with step size $\varepsilon/3$. Notice that, as we increase ε, the model becomes increasingly more robust at the cost of nominal performance. Still, the performance degradation remains abrupt. As we argued before, smoother degradation can be obtained by training against a distribution of magnitudes, e.g., the one in Figure 2b. Doing so not only yields better performances under perturbation as well as a small loss of nominal accuracy.

7 Conclusion

We put forward a theory of learning under requirements by extending the PAC framework to constrained learning. We then prove that unconstrained and constrained learnability are equivalent by showing that a constrained version of the classical ERM rule is a PACC learner. To overcome the challenges in solving the optimization problem underlying this learner, we derive an alternative learner based on a parametrized empirical dual problem. We show that its approximation error is related to the richness of the parametrization as well as the difficulty of meeting the learning constraint and use it to propose a practical algorithm to learn under requirements. We expect that these generalization results can be used to theoretically ground techniques used in practice to address constrained learning problems beyond fairness and robustness. In particular, similar arguments can be used to develop a constrained theory for reinforcement learning [75]. We also believe that these results can be extended to non-convex losses using recent results on the strong duality of certain non-convex variational problems [68].

Broader Impact

As learning becomes an ubiquitous technological solution and begins to affect real societal impact, its shortcomings become more evident. A growing number of reports show that its solutions can be prejudiced and prone to tampering or unsafe behaviors [1–6]. Constrained learning allows requirements to be imposed during learning, so that the models and solutions obtained are guaranteed to behave in the desired way despite being learned fully from data. This work provides a framework under which to study learning under requirements and shows how and when it can be done. By providing generalization guarantees on the solutions, it enables learning to be used in critical applications in which there is little tolerance for failure. Naturally, solutions learned under constraints are not necessarily safe or fair. How the learning problem is formulated, i.e., which constraints are imposed, play a definite role on these outcomes and policies determining such requirements can be (and indeed are [76–78]) important sources of biases.

Acknowledgments and Disclosure of Funding

This work is supported by ARL DCIST CRA W911NF-17-2-0181.

References

[1] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.

[2] M. Kay, C. Matuszek, and S. A. Munson, "Unequal representation and gender stereotypes in image search results for occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 3819–3828. [Online]. Available: https://doi.org/10.1145/2702123.2702520

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] National Transportation Safety Board, "HWY18MH010: Preliminary report," National Transportation Safety Board, Tech. Rep., 2018. [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

[5] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[6] P. Bright, "Tay, the neo-Nazi millennial chatbot, gets autopsied," *Ars Technica*, 2016. [Online]. Available: https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/

[7] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[8] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 2016, pp. 2990–2999.

[9] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "Rotation equivariant vector field networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5058–5067.

[10] J. F. Henriques and A. Vedaldi, "Warped convolutions: Efficient invariance to spatial transformations," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 1461–1469.

[11] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3856–3866.

[12] M. Weiler, F. A. Hamprecht, and M. Storath, "Learning steerable filters for rotation equivariant CNNs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[13] L. Ruiz, F. Gama, A. G. Marques, and A. Ribeiro, "Invariance-preserving localized activation functions for graph neural networks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 127–141, 2020.

[14] J. Chen and L. Deng, "A primal-dual method for training recurrent neural networks constrained by the echo-state property," in *International Conference on Learning Representations*, 2014.

[15] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," in *Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[16] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. Van den Broeck, "A semantic loss function for deep learning with symbolic knowledge," in *International Conference on Machine Learning*, 2018.

[17] S. N. Ravi, T. Dinh, V. S. Lokhande, and V. Singh, "Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 4772–4779.

[18] D. Bertsekas, *Convex Optimization Theory*. Athena Scientific, 2009.

[19] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000.

[20] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2004.

[21] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[22] D. Haussler.

[23] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander, "Satisfying real-world goals with dataset constraints," in *Advances in Neural Information Processing Systems*, 2016, pp. 2415–2423.

[24] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[25] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*, 2018, pp. 60–69.

[26] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems*, 2018, pp. 2791–2801.

[27] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*, 2018, pp. 2564–2572.

[28] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *Journal of Machine Learning Research*, vol. 20, no. 75, pp. 1–42, 2019.

[29] A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals," *Journal of Machine Learning Research*, vol. 20, no. 172, pp. 1–59, 2019.

[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[31] A. Sinha, H. Namkoong, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," in *International Conference on Learning Representations*, 2018.

[32] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*, 2019, pp. 7472–7482.

[33] D. S. Kalogerias and W. B. Powell, "Recursive optimization of convex risk measures: Mean-semideviation models," *Extended Preprint, Arxiv*, 2018.

[34] C. Vitt, D. Dentcheva, and H. Xiong, "Risk-averse classification," *Arxiv*, 2018.

[35] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[36] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 22–31.

[37] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Learning safe policies via primal–dual methods," in *IEEE Conference on Decision and Control*, 2019.

[38] N. Nguyen and R. Caruana, "Improving classification with pairwise constraints: A margin-based approach," in *Machine Learning and Knowledge Discovery in Databases*, W. Daelemans, B. Goethals, and K. Morik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 113–124.

[39] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *J. Mach. Learn. Res.*, vol. 12, pp. 1501–1536, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021049

[40] F. Yu and M.-L. Zhang, "Maximum margin partial label learning," *Machine Learning*, vol. 106, no. 4, pp. 573–593, 2017. [Online]. Available: https://doi.org/10.1007/s10994-016-5606-4

[41] J. Neyman and E. S. Pearson, "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London*, vol. 231, no. 694–706, pp. 289–337, 1933.

[42] A. Garg and D. Roth, "Learning coherent concepts," in *Algorithmic Learning Theory*, 2001, pp. 135–150.

[43] L. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro, "The empirical duality gap of constrained statistical learning problems," in *International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2613–2616.

[44] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, 2014.

[45] S. Zhao, J. Song, and S. Ermon, "The information autoencoding family: A Lagrangian perspective on latent variable generative models," in *Conference on Uncertainty in Artificial Intelligence*, 2018.

[46] M. Ehrgott, *Multicriteria Optimization*. Springer, 2005.

[47] K. Miettinen, *Nonlinear Multiobjective Optimization*. Springer, 1998.

[48] A. Messac, A. Ismail-Yahaya, and C. Mattson, "The normalized normal constraint method for generating the Pareto frontier," *Structural and Multidisciplinary Optimization*, vol. 25[2], pp. 86–98, 2003.

[49] D. Mueller-Gritschneder, H. Graeb, and U. Schlichtmann, "A successive approach to compute the bounded Pareto front of practical multiobjective optimization problems," *SIAM Journal on Optimization*, vol. 20[2], pp. 915–934, 2009.

[50] T. Schaul, D. Borsa, J. Modayil, and R. Pascanu, "Ray interference: a source of plateaus in deep reinforcement learning," *arXiv preprint arXiv:1904.11455*, 2019.

[51] D. Bertsekas, *Convex optimization algorithms*. Athena Scientific, 2015.

[52] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, "Learning with a strong adversary," 2015.

[53] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, 2018.

[54] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[55] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[56] L. Chamon and A. Ribeiro, "Probably approximately correct constrained learning," *Advances in Neural Information Processing Systems (under review)*, 2020, https://arxiv.org/abs/2006.05487.

[57] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[58] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 233–242.

[59] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," *arXiv preprint arXiv:1711.00501*, 2017.

[60] A. Brutzkus and A. Globerson, "Globally optimal gradient descent for a convnet with gaussian inputs," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 605–614.

[61] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, 2018.

[62] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[63] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[64] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[65] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[66] C. A. Micchelli and M. Pontil, "Learning the kernel function via regularization," *Journal of machine learning research*, vol. 6, no. Jul, pp. 1099–1125, 2005.

[67] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[68] L. Chamon, Y. Eldar, and A. Ribeiro, "Functional nonlinear sparse models," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2449–2463, 2020.

[69] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*. Stanford University Press, 1958.

[70] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980v9*, 2017.

[72] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Innovations in Theoretical Computer Science Conference*, 2012, pp. 214–226.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[74] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[75] S. Paternain, L. F. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Advances in Neural Information Processing Systems*, 2019, pp. 7553–7563.

[76] P. Anand, "Algorithms of inequality," *The appeal*, 2020. [Online]. Available: https://theappeal.org/politicalreport/algorithms-of-inequality-covid-ration-care/

[77] C. Hadavas, "How automation bias encourages the use of flawed algorithms," *Slate*, 2020. [Online]. Available: https://slate.com/technology/2020/03/ice-lawsuit-hijacked-algorithm.html

[78] K. Loggins, "Here's what happens when an algorithm determines your work schedule," *Vice*, 2020. [Online]. Available: https://www.vice.com/en_us/article/g5xwby/heres-what-happens-when-an-algorithm-determines-your-work-schedule