A Lagrangian Dual Framework for Deep Neural Networks with Constraints

Ferdinando Fioretto[1] , **Terrence WK Mak**[2] , **Federico Baldo**[3] ,
Michele Lombardi[3] and **Pascal Van Hentenryck**[2]

[1]Syracuse University , [2]Georgia Institute of Technology , [3]University of Bologna

ffiorett@syr.edu, wmak@gatech.edu, federico.baldo2,@unibo.it,
michele.lombardi2@unibo.it, pvh@isye.gatech.edu

arXiv:submit/3019298 [cs.LG] 26 Jan 2020

Abstract

A variety of computationally challenging constrained optimization problems in several engineering disciplines are solved repeatedly under different scenarios. In many cases, they would benefit from fast and accurate approximations, either to support real-time operations or large-scale simulation studies. This paper aims at exploring how to leverage the substantial data being accumulated by repeatedly solving instances of these applications over time. It introduces a deep learning model that exploits Lagrangian duality to encourage the satisfaction of hard constraints. The proposed method is evaluated on a collection of realistic energy networks, by enforcing non-discriminatory decisions on a variety of datasets, and on a transprecision computing application. The results illustrate the effectiveness of the proposed method that dramatically decreases constraint violations by the predictors and, in some applications, increases the prediction accuracy.

1 Introduction

Deep Neural Networks, in conjunction with progress in GPU technology and the availability of large data sets, have proven enormously successful at a wide array of tasks, including image classification [Krizhevsky *et al.*, 2012], speech recognition [Amodei *et al.*, 2016], and natural language processing [Collobert and Weston, 2008], to name but a few examples. More generally, deep learning has achieved significant success on a variety of regression and classification tasks. On the other hand, the application of deep learning to aid computationally challenging constrained optimization problems has been more sparse, but is receiving increasing attention, such as the efforts in jointly training prediction and optimization models [Vinyals *et al.*, 2015; Khalil *et al.*, 2017; Kool *et al.*, 2018] and incorporating optimization algorithms into differentiable systems [Donti *et al.*, 2017; Amos and Kolter, 2017; Wilder *et al.*, 2019].

This research originated in an attempt to apply deep learning to fundamentally different application areas: The learning of constrained optimization problems and, in particular, optimization problems with hard physical and engineering constraints. These constrained optimization problems arise in numerous contexts including in energy systems, mobility, resilience, and disaster management. Indeed, these applications must capture physical laws such as Ohm's law and Kirchhoff's law in electrical power systems, the Weymouth equation in gas networks, flow constraints in transportation models, and the Navier-Stoke's equations for shallow water in flood mitigation. Moreover, they often feature constraints that represent good engineering and operational practice to protect various devices. For instance, they may include thermal limits, voltage and pressure bounds, as well as generator and pump limitations, when the domain is that of energy systems. Direct applications of deep learning to these applications may result in predictions with severe constraint violations, as shown in Section 5.

There is thus a need to provide deep learning architectures with capabilities that would allow them to capture constraints directly. Such models can have a transformative impact in many engineering applications by providing high-quality solutions in real-time and be a cornerstone for large planning studies that run multi-year simulations. To this end, this paper proposes a *Lagrangian Dual Framework for Deep Learning* that addresses the challenge of enforcing constraints during learning: Its key idea is to transfer Lagrangian Duality, which is widely used to obtain tight bounds in optimization, to deep learning. This paper presents the theoretical foundations of the Lagrangian Dual Framework and demonstrates its practical potential with applications in electricity and gas networks.

Interestingly, the proposed Lagrangian Dual Framework can also be applied to another class of applications: Constrained Learning Problems, that requires specific properties to hold on the predictor itself. For instance, transprecision computing is a technique that achieves energy savings by adjusting the precision of power-hungry algorithms. An important challenge in this area is to predict the error resulting from a loss in accuracy and the error should be monotonically decreasing with increases in accuracy. As a result, the learning task may impose constraints over the samples used during the predictor training. Other applications in constrained learning may impose fairness constraints on predictors, e.g., a constraint ensuring no disparate impact in a classifier. This paper shows that the Lagrangian Dual Framework also provides a tool to address these constrained learning problems

and reports empirical results on applications in transprecision computing and fair classification.

The Lagrangian Dual Framework is evaluated extensively on a variety of real benchmarks in power system optimization and gas compression optimization, that present hard engineering and operational constraints. Additionally, the proposed method is tested on several datasets that enforce non-discriminatory decisions and on a realistic transprecision computing application, that requires constraints to be enforced on the predictors themselves. The results present a dramatic improvement in the number of constraint violations reduction, and often result in substantial improvements in the prediction in energy optimization problems.

2 Preliminaries: Lagrangian Duality

Consider the optimization problem

$$\mathcal{O} = \operatorname*{argmin}_y f(y) \text{ subject to } g(y) \leqslant 0. \tag{1}$$

In *Lagrangian relaxation*, some or all the problem constraints are relaxed into the objective function using *Lagrangian multipliers* to capture the penalty induced by violating a constraint. When all the constraints are relaxed, the *Lagrangian function* becomes

$$f_\lambda(y) = f(y) + \lambda g(y)$$

where $\lambda \geqslant 0$ are the Lagrangian multipliers. Note that, in this formulation, $g(y)$ can be positive or negative. An alternative formulation, used in augmented Lagrangian methods [Hestenes, 1969] and constraint programming [Fontaine *et al.*, 2014], uses the following Lagrangian function

$$f_\lambda(y) = f(y) + \lambda \max(0, g(y))$$

where the expression $\max(0, g(y))$ captures a quantification of the constraint violations. This paper abstracts these formulations by using a function $\nu(\cdot)$ that returns either the constraint satisfiability or violation degree of a constraint.

When using a Lagrangian function, the optimization problem becomes

$$LR_\lambda = \operatorname*{argmin}_y f_\lambda(y)$$

and it satisfies $f(LR_\lambda) \leqslant f(\mathcal{O})$.

Finally, to obtain the strongest Lagrangian relaxation of $\mathcal{O}$, the *Lagrangian dual* can be used to find the best Lagrangian multipliers, i.e.,

$$LD = \operatorname*{argmax}_{\lambda \geqslant 0} f(LR_\lambda).$$

For various classes of problems, the Lagrangian dual is a strong approximation of $\mathcal{O}$. Moreover, its optimal solutions can often be translated into high-quality feasible solutions by a post-processing step.

3 Constrained Optimization Problems

This section describes how to use the Lagrangian Dual Framework for approximating constrained optimization problems. It first reviews two fundamental applications that serve as motivation.

3.1 Motivating Applications

Many energy applications require solving challenging (non-convex, non-linear) optimization problems in order to derive the best system operational controls to serve the energy demands of the customers. Power grid and gas pipeline systems are two examples of such applications. While these problems can be solved using effective optimization solvers, their resolution relies on *accurate* predictions of the energy demands. The increasing penetration of renewable energy sources, including those behind the meter (e.g., solar panels on roofs), has rendered accurate predictions more challenging. In turn, predictions need to be performed at minute time scales to ensure sufficient accuracy. Finding optimal solutions for these underlying optimization problems in these reduced time scales thus becomes computationally challenging, opening opportunities for machine-learning approaches. The next paragraphs review two energy applications that motivate the proposed framework.

Optimal Power Flow The *Optimal Power Flow* (OPF) problem determines the best generator dispatch ($y = S^g$) of minimal cost ($\mathcal{O} = \min_{S^g} \text{cost}(S^g)$) that meets the demands ($d = S^d$) while satisfying the physical and engineering constraints ($g(y)$) of the power system [Chowdhury and Rahman, 1990], where S^g and S^d denote the vectors (in complex domain) of generator dispatches and power demands. Typical constraints include the non-linear non-convex AC power flow equations, Kirchhoff's current laws, voltage bounds, and thermal limits. The OPF problem is a fundamental building bock of many applications, including security-constrained OPFs [Monticelli *et al.*, 1987]), optimal transmission switching [Fisher *et al.*, 2008], capacitor placement [Baran and Wu, 1989], and expansion planning [Niharika *et al.*, 2016].

Optimal Gas Compressor Optimization The Optimal Gas Compressor Optimization (OGC) problem aims at determining the best compression controls ($y = R$) with minimum compression costs ($\mathcal{O} = \min_R \text{cost}(R)$) to meet gas demands ($d = q^d$) while satisfying the physical and operational limits ($g(y)$) of the natural gas pipeline systems [Herty *et al.*, 2010]. Therein, R and q^d are compressors control values and gas demands. Typical constraints include: the non-linear gas flow equations describing pressure losses, the flow balance equations, the non-linear non-convex compressor objectives $\mathcal{O}$, and the pressure bounds. Similar to the OPF problem, the OGC problem is a non-convex non-linear optimization problem with physical and engineering constraints and a fundamental building block for many gas pipeline problems.

The next section shows how to approximate OPFs and OGCs, by viewing them as parametric optimization problems, using the proposed Lagrangian Dual Framework.

3.2 The Learning Task

The learning task estimates a parametric version of optimization problem (1), defined as

$$\mathcal{O}(d) = \operatorname*{argmin}_y f(y, d) \text{ subject to } g(y, d) \leqslant 0 \tag{2}$$

with a set of samples $\{(d_l, y_l = \mathcal{O}(d_l))\}_{l=1}^n$. More precisely, given a parametric model $\mathcal{M}[w]$ with weights w and a loss

function $\mathcal{L}$, the learning task must solve the following optimization problem

$$w^* = \operatorname*{argmin}_w \sum_{l=1}^{n} \mathcal{L}(\mathcal{M}[w](d_l), y_l) \tag{3a}$$

$$\text{subject to } g(\mathcal{M}[w](d_l), d_l) \leqslant 0 \quad (1 \leqslant l \leqslant n) \tag{3b}$$

to obtain the approximation $\widetilde{\mathcal{O}} = \mathcal{M}[w^*]$ of $\mathcal{O}$.

The main difficulty lies in the constraints $g(y, d) \leqslant 0$, which can represent physical and operational constraints, as mentioned in the motivating applications. In addition, observe that the model weights must be chosen so that the constraints are satisfied for all samples, which makes the learning particularly challenging. A naive approach to the learning task is thus likely to result in predictors that violate these constraints significantly, as demonstrated in Section 5, producing a model that would not be useful in practice.

3.3 The Lagrangian Dual Framework

To learn constrained optimization problems, the paper proposes a *Lagrangian dualilty* framework to the learning task. The framework relies on the notion of *Augmented Lagrangian* [Hestenes, 1969] used for solving constrained optimization problems [Fontaine *et al.*, 2014].

In more details, the Lagrangian duality framework exploits a Lagrangian dual approach in the learning task to approximate minimizer $\mathcal{O}$. Given multipliers λ, consider the Lagrangian loss function

$$\mathcal{L}_\lambda(\tilde{y}_l, y_l, d_l) = \mathcal{L}(\tilde{y}_l, y_l) + \lambda\, \nu(g(\tilde{y}_l, d_l) \leqslant 0).$$

For multipliers λ, solving the optimization problem

$$w^*(\lambda) = \operatorname*{argmin}_w \sum_{l=1}^{n} \mathcal{L}_\lambda(\mathcal{M}[w](d_l), y_l, d_l) \tag{4}$$

produces an approximation $\widetilde{\mathcal{O}}_\lambda = \mathcal{M}[w^*(\lambda)]$ of $\mathcal{O}$. The Lagrangian dual computes the optimal multipliers, i.e.,

$$\lambda^* = \operatorname*{argmax}_\lambda \min_w \sum_{l=1}^{n} \mathcal{L}_\lambda(\mathcal{M}[w](d_l), y_l, d_l) \tag{5}$$

to obtain $\widetilde{\mathcal{O}}^* = \mathcal{M}[w^*(\lambda^*)]$, i.e., the strongest Lagrangian relaxation of $\mathcal{O}$.

Learning $\widetilde{\mathcal{O}}^*$ relies on an iterative scheme that interleaves the learning of a number of Lagrangian relaxations (for various multipliers) with a subgradient method to learn the best multipliers. The Lagrangian dual framework, described in Equations (4) and (5), can be summarized as the iteration of the following three steps:

(6a) Learn $\widetilde{\mathcal{O}}_{\lambda^k}$

(6b) Let $y_l^k = \widetilde{\mathcal{O}}_{\lambda^k}(d_l)$ $(1 \leqslant l \leqslant n)$

(6c) $\lambda^{k+1} = \lambda^k + s_k \sum_{l=1}^{n} \nu(g(y_l^k, d_l) \leqslant 0)$

where k denotes the iteration index and the multipliers update in step (6c) use a step size s_k.

4 Constrained Learning

This section describes how to use the Lagrangian Duality Framework for Constrained Learning Problems. It starts with two motivating applications.

4.1 Motivating Applications

Several applications require to enforce constraints on the learning process itself to attain desirable properties of the predictor. These constraints impose conditions on subsets of the samples that must be satisfied. For instance, assume that there is a partial order $\preceq$ on the optimization inputs and the following property holds:

$$d_1 \preceq d_2 \Rightarrow f(\mathcal{O}(d_1), d_1) \leqslant f(\mathcal{O}(d_2), d_2).$$

The predictor should ideally satisfy these constraints as well:

$$d_1 \preceq d_2 \Rightarrow f(\widetilde{\mathcal{O}}(d_1), d_1) \leqslant f(\widetilde{\mathcal{O}}(d_2), d_2).$$

Transprecision computing Transprecision computing is the idea of reducing energy consumption by reducing the precision (a.k.a. number of bits) of the variables involved in a computation [Malossi *et al.*, 2018]. It is especially important in low-power embedded platforms, which arise in many contexts such as smart wearable and autonomous vechicles. Increasing precision typically reduces the error of the target algorithm. However, it also increases the energy consumption, which is a function of the maximal number of used bits. The objective is to design a *configuration* d_l, i.e., a mapping from input computation to the precision for the variables involved in the computation. The sought configuration should balance *precision* and *energy consumption*, given a bound to the error produced by the loss in precision when the highest precision configuration is adopted.

However, given a configuration, computing the corresponding error can be very time-consuming and the task considered in this paper seeks to learn a mapping between configurations and error. This learning task is non-trivial, since the solution space precision-error is non-smooth and non-linear [Malossi *et al.*, 2018]. The samples (d_l, y_l) in the dataset represent, respectively, a configuration d_l and its associated error y_l obtained by running the configuration d_l for a given computation. The problem $\mathcal{O}(d_l)$ specifies the error obtained when using configuration d_l.

Importantly, transcomputing expects a *monotonic* behavior: Higher precision configurations should generate more accurate results (i.e., a smaller error). Therefore, the structure of the problem imposes the learning task to require a dominance relation $\preceq$ between instances of the dataset. More precisely, $d_2 \preceq d_2$ holds if

$$\forall i \in [N] :\ x_{1_i} \leqslant x_{2_i}$$

where N is the number of variables involved in the computation and x_{1_i}, x_{2_i} are the precision values for the variables in d_1 and d_2 respectively.

Fair Classifier The second motivating application considers the task of building a classifier that minimizes *disparate treatment* [Zafar *et al.*, 2017]. Disparate Treatment arises when a classifier provides different results for groups of individuals with similar *non-sensitive* features but different *sensitive* features. Consider a classifier that maps feature vectors $d \in \mathcal{D} \subseteq \mathbb{R}^n$ with associated sensitive values $d^s \in \mathcal{D}_s$ and non-sensitive values $d^p \in \mathcal{D}_p$ to labels $y \in \mathcal{Y}$. A classifier *does not* suffer from disparate treatment if:

$$\Pr(y|d^s, d^p) = \Pr(y|d^p) \quad \forall y \in \mathcal{Y}, d \in \mathcal{D},$$

that is, the probability of returning a particular value y does not depend on the sensitive features d^s. To construct an estimator that minimizes the disparate treatment discrimination, the paper considers $|\mathcal{D}_s|$ estimators $\mathcal{M}_1, \ldots, \mathcal{M}_m$, each associated with a dataset partition $D_{|s_i} = \{(d_l, y_l) | d_l^s = s_i\}$ that marginalizes for a particular (combination of) value(s) of the protected feature(s), in addition to the classical estimator $\mathcal{M}$ that is trained over the entire dataset D. Thus, the learning process is defined by the following objective:

$$\min_{w, w_1, \ldots, w_m} \mathcal{L}\left(\mathcal{M}[w](D)\right) + \sum_i \mathcal{L}\left(\mathcal{M}_i[w_i](D_{|s_i})\right)$$

$$\text{such that } \mathcal{M}_i[w_i](D_{s_i}) = \mathcal{M}[w](D) \;\; \forall i \in [m],$$

that enforces a constraint on the output of the classifiers $\mathcal{M}_i$, trained on data D_{s_i} to be equivalent to that of the output of the classifier $\mathcal{M}$, which is trained on the whole dataset D.

The next section will specify how to encode such type of constraints as well as how to express and enforce dominance relations in the proposed constrained learning framework.

4.2 The Learning Task

Consider a set $\mathcal{S} = \{S_1, \ldots, S_m\}$ where S_i is a subset of the inputs that must satisfy the associated constraint

$$h_i(\{\mathcal{O}(d_l)\}_{l \in S_i}, \{d_l\}_{l \in S_i}).$$

In this context, the learning task is defined by the following optimization problem

$$\operatorname*{argmin}_w \sum_{i=1}^{n} \mathcal{L}(\mathcal{M}[w](d_l), y_l) \tag{7a}$$

$$\text{subject to } g(\mathcal{M}[w](d_l), d_l) \leqslant 0 \;\; (1 \leqslant l \leqslant n) \tag{7b}$$

$$h_i(\{\mathcal{M}[w](d_l)\}_{l \in S_i}, \{d_l\}_{l \in S_i}) \;\; (1 \leqslant i \leqslant m). \tag{7c}$$

4.3 The Lagrangian Duality Framework

To approximate Problem (7), the learning task considers Lagrangian loss functions of the form

$$\mathcal{L}_{\mu,\lambda}(\tilde{y}_l, y_l, d_l) = \mathcal{L}_\lambda(\tilde{y}_l, y_l, d_l) + \mu \, \nu(h(\tilde{y}, \mathcal{S})),$$

where $\tilde{y}_l = \mathcal{M}[w](d_l)$. It learns approximations of the Lagrangian relaxations $\widetilde{\mathcal{O}}_{\lambda,\mu}$ of the form

$$w^*(\mu, \lambda) = \operatorname*{argmin}_w \sum_{l=1}^{n} \mathcal{L}_{\mu,\lambda}(\mathcal{M}[w](d_l), y_l, d_l), \tag{8}$$

as well as the Lagrangian duals of Equation (8) of the form

$$\lambda^*(\mu) = \operatorname*{argmax}_\lambda \min_w \sum_{l=1}^{n} \mathcal{L}_{\mu,\lambda}(\mathcal{M}[w](d_l), y_l, d_l), \tag{9}$$

and, finally, the Lagrangian dual of the Lagrangian duals (Equation (9)) as

$$\mu^* = \operatorname*{argmax}_\mu \max_\lambda \min_w \sum_{l=1}^{n} \mathcal{L}_{\mu,\lambda}(\mathcal{M}[w](d_l), y_l, d_l) \tag{10}$$

to obtain the best estimator $\widetilde{\mathcal{O}}^* = \mathcal{M}[w^*]$, where

$$w^* = \operatorname*{argmin}_w \sum_{l=1}^{n} \mathcal{L}_{\mu^*,\lambda^*(\mu^*)}(\mathcal{M}[w](d_l), y_l, d_l).$$

The learning algorithm interleaves the learning of the Lagrangian duals with the subgradient optimization of the multipliers μ, i.e.,

(11a) Learn $\lambda^*(\mu^k)$ to obtain $\widetilde{\mathcal{O}}_{\mu^k,\lambda^*(\mu^k)}$ (as in (6))

(11b) Let $y_l^k = \widetilde{\mathcal{O}}_{\mu^k,\lambda^*(\mu^k)}(d_l)$ $(1 \leqslant l \leqslant n)$

(11c) $\mu^{k+1} = \mu^k + t_k \sum_{i=1}^{m} \nu(h(\{y_l^k\}_{l \in S_i}, \{d_l\}_{l \in S_i}))$

where t_k, in step (11c), is a step size.

5 Experiments

This section evaluates the predictive accuracy of the proposed Lagrangian Duality Framework on constrained optimization problems for energy and gas networks and on constrained learning problems–that enforce constraints on the predictors– for applications in transprecision computing and fairness.

5.1 Constrained Optimization Problems

Data set Generation The experiments examine the proposed models on a variety of power networks from the NESTA library [Coffrin *et al.*, 2014] and natural gas benchmarks from [Mak *et al.*, 2019] and GasLib [Pfetsch *et al.*, 2015]. The ground truth data are constructed as follows: For each power and gas network, different benchmarks are generated by altering the amount of nominal demands $d = S^d$ (for power networks) and $d = q^d$ (for gas network gas) within a $\pm 20\%$ range. The resulting 4000 demand vectors are used to generate solutions to the OPF/OGC problem. Increasing loads causes heavily congestions to the system, rendering the computation of optimal solutions challenging. A network value that constitutes a dataset entry $(d_l, y_l = \mathcal{O}(d))$ is a feasible solution obtained by solving the AC-OPF problem [Chowdhury and Rahman, 1990] for electricity networks or the OGC for gas networks [Herty *et al.*, 2010]. The experiments use a $80/20$ train-test split and report results on the test set.

Learning Models The experiments use a baseline feed-forward network model, with 5 layers, $\mathcal{M}$ which minimizes the Mean Squared Error (MSE) loss $\mathcal{L}$ to predict to active power $\hat{p}$, voltage magnitude $\hat{v}$, and voltage angle $\hat{\theta}$, for energy networks, and compression ratios $\hat{R}, pressure\hat{p}$, and gas flows $\hat{q}$, for gas networks.

This baseline model is compared with a model $\mathcal{M}_\text{C}$ that exploits the problem constraints and minimizes the loss: $\mathcal{L} + \lambda\nu(\cdot)$, with multiplier values λ fixed to $\mathbf{1}$. Finally, $\mathcal{M}_\text{C}^\text{D}$ extends model $\mathcal{M}_\text{C}$ by learning the Lagrangian multipliers using the Lagrangian duality framework introduced in Section 3.3. The constrained learning model for power systems also exploits the hot-start techniques used in [Fioretto *et al.*, 2020], with states differing by at most 1%. Results with larger percentages (up to 10%) are essentially similar.

The training uses the Adam optimizer with learning rate $(\alpha = 0.001)$ and β values $(0.9, 0.999)$ and was performed for 80 epochs using batch sizes $b = 64$. Finally, the Lagrangian step size ρ is set to 10^{-4}.

Prediction Errors

Table 1 and 2 report the average L_1-distance between a subset of predicted variables y (marked with $\hat{y}$) on both the power and gas benchmarks and their original ground-truth

Test Case	Type	$\mathcal{M}$	$\mathcal{M}_C$	$\mathcal{M}_C^D$
30_ieee	$\hat{p}$	3.3465	0.3052	**0.0055**
	$\hat{v}$	14.699	0.3130	**0.0070**
	$\hat{\theta}$	4.3130	0.0580	**0.0041**
	$\tilde{p}^f$	27.213	0.2030	**0.0620**
118_ieee	$\hat{p}$	0.2150	0.0380	**0.0340**
	$\hat{v}$	7.1520	0.1170	**0.0290**
	$\hat{\theta}$	4.2600	1.2750	**0.2070**
	$\tilde{p}^f$	38.863	0.6640	**0.4550**
300_ieee	$\hat{p}$	0.0838	0.0174	**0.0126**
	$\hat{v}$	28.025	3.1130	**0.0610**
	$\hat{\theta}$	12.137	7.2330	**2.5670**
	$\tilde{p}^f$	125.47	26.905	**1.1360**

Table 1: Mean Prediction Errors (%) on OPF Benchmarks.

Test Case	Type	$\mathcal{M}$	$\mathcal{M}_C$	$\mathcal{M}_C^D$
24-pipe	$\hat{R}$	0.0052	0.0079	**0.0025**
	$\hat{p}$	**0.0057**	0.0068	**0.0057**
	$\hat{q}$	0.0029	0.0592	**0.0007**
40-pipe	$\hat{R}$	0.0009	0.0103	**0.0006**
	$\hat{p}$	0.0011	0.0025	**0.0006**
	$\hat{q}$	0.0006	0.0329	**0.0004**
135-pipe	$\hat{R}$	0.0206	0.0317	**0.0199**
	$\hat{p}$	0.0260	**0.0209**	0.0225
	$\hat{q}$	0.0223	0.0572	**0.0222**

Table 2: Mean Prediction Errors (%) on OGC Benchmarks.

quantities. Active generation dispatches $\hat{p}^g = Re(S^g)$, voltage magnitudes $\hat{v}$, voltage angles $\hat{\theta}$, and active transmission line (including transformers) flows $\tilde{p}^f$ are included for power network benchmarks. Power flows $\tilde{p}^f$ are not directly predicted but computed from the predicted quantities through the Ohm's laws. For the gas network benchmarks, compression ratios $\hat{R}$, pressure values $\hat{p}$, and gas flows $\hat{q}$ are included. Let y be a quantity to be measured. The error for y is reported in percentage by $100\frac{\|\hat{y}-y\|_1}{\|y\|_1}$. The best results are highlighted in bold. A clear trend appears: The prediction errors decrease with the increase in model complexity. In particular, model $\mathcal{M}_C$, which exploits the problem constraints, predicts much better voltage quantities and power flows than $\mathcal{M}$, for OPF problems. The prediction errors on OGC benchmarks, instead remain on the same order of magnitude as those obtained by the baseline model $\mathcal{M}$. *Finally, the Lagrangian dual framework that finds the best multipliers ($\mathcal{M}_C^D$) consistently improves the baseline model on OGC benchmarks, and further improves $\mathcal{M}_C$ predictions by up to an additional order of magnitude, for OPF problems.* The accuracy gains appear more pronounced on OPF problems since the baseline model $\mathcal{M}$ produces already extremely accurate results on OGC benchmarks.

Minimum Operational Adjustments

This section simulates the prediction results in an operational environment, by measuring the minimum required adjustments in order to satisfy the operational limits and physical constraints. The minimum distance can be found by running least-square problems on the predicted controls: generator dispatch and voltage set points for power systems and compression ratios for gas systems[1]. Table 3 reports the minimum distance required (in normalized % metric) to satisfy the operational limits and physical constraints, and the best results are highlighted in bold. The adjustment required decrease with the increase in model complexity. These results provide a proxy to evaluate the degree of constraint violations of a model. *They show that the Lagrangian Duality Framework can drastically reduce the effort required by a post-processing step to satisfy the problem constraints.*

Test Case		$\mathcal{M}$	$\mathcal{M}_C$	$\mathcal{M}_C^D$
30_ieee	p^g	2.0793	0.1815	**0.0007**
	v	83.138	0.0944	**0.0037**
118_ieee	p^g	0.1071	0.0043	**0.0038**
	v	3.4391	0.0956	**0.0866**
300_ieee	p^g	0.0447	0.0091	**0.0084**
	v	31.698	0.2383	**0.1994**
24-pipe	R	0.1012	0.1033	**0.0897**
40-pipe	R	0.0303	0.0277	**0.0207**

Table 3: Average distances (in normalized % metric) for the active power p^g, voltage magnitude v, and compressor ratios R of the simulated solutions w.r.t. the corresponding predictions.

5.2 Constrained Learning

This section examines the Constrained Learning model discussed in Section 4.1 on Transprecision computing and fairness application domains.

Transprecision computing

The benchmark considers training a neural network to predict the error of transprecision configurations. The monotonicity property is expressed as a constraint exploiting the relation of dominance among configurations of the train set, i.e. $\nu_i = \max(0, \mathcal{M}(x_1) - \mathcal{M}(x_2))$ if $x_1 \preceq x_2$ for every pair (x_1, x_2). This approach is particularly suited for instances of training with scarce data points with a high rate of violated constraints, since it guides the network in the learning process towards a more general approximation of the target function. In order to explore different scenarios the experiments use 3 different train sets of increasing size, i.e. 100, 250 and 1000. The test set size is fixed to 1000 samples. The data sets are constructed by generating random configuration (d_i) and computing errors (y) by measuring the performance loss obtained when running the configuration d_i on the target algorithm.

Table 4 illustrates the results comparing a model that minimizes the MSE prediction error $\mathcal{M}$, one that include the Lagrangian loss functions $\mathcal{L}_\lambda$ associated to each constraint and where all λ are fixed to value 1.0 ($\mathcal{M}_C$), and the proposed model that uses the Lagrangian dual framework to find the optimal Lagrangian weights ($\mathcal{M}_C^D$). All prediction model are implemented classical feed-forward neural network with 3 layers and 10 hidden units and minimize the Mean Squared Error as loss function. The training uses 150 epochs, Lagrangian step size $s_k = 10^{-4}$ and learning rate 10^{-3}.

The table clearly illustrates the positive effect of adding the Lagrangian constraints on reducing the number of constraint

[1]Non-convex least square problems can be challenging for interior point solvers. 135-pipe is skipped due to convergence issues.

Train Size	$\mathcal{M}$		$\mathcal{M}_C$		$\mathcal{M}_C^D$	
	MAE	VC	MAE	VC	MAE	VC
100	0.199	1136	0.219	**301**	**0.198**	665
250	0.199	637	0.255	**65**	**0.198**	124
1000	**0.146**	233	0.208	**31**	**0.146**	66

Table 4: Mean Absolute Error (MAE) and number of constraints violations (VC). Best results are highlighted in bold.

Dataset	$\mathcal{M}$		$\mathcal{M}_C$		$\mathcal{M}_C^D$	
	Acc.	DTDI	Acc.	DTDI	Acc.	DTDI
Adult	**0.8488**	0.3517	0.8439	0.3281	0.8350	**0.2453**
Default	0.8204	0.1216	**0.8224**	0.1168	0.8202	**0.1066**
COMPAS	**0.9681**	1.6012	0.9663	1.5959	0.9432	**1.4120**

Table 5: Classification accuracy (Acc.) and fairness score (DTDI). Best results are highlighted in bold.

violations. Both versions of the predictors that use the Lagrangian framework reduce substantially the number of violated constraints, with model $\mathcal{M}_C$ producing fewer constraint violations, on average, then $\mathcal{M}_C^D$. However, $\mathcal{M}_C$ also increases the MAE score substantially, while $\mathcal{M}_C^D$ finds a good tradeoff between precision and number of constraints violations. The most significant contribution was obtained on training sets with fewer data points, confirming that *exploiting the Lagrangian Duals of the Constraint Violations can be an important tool for constrained learning* .

Fairness Constraints

The benchmark considers building a classifier that minimizes *disparate treatment* [Zafar *et al.*, 2017]. The paper considers the disparate treatment discrimination (DTDI) index, introduced by Aghaei *et al.* (2019), to quantify the disparate treatment in a dataset. Given a dataset of samples $D = (x_i, y_i)_{i\in[n]}$, this index is defined as:

$$\text{DTDI}(D) = \sum_{y\in\mathcal{Y}, s\in\mathcal{X}_s, j\in[n]} \Big| \frac{\sum_{i\in[n]} d(x_i^p, x_j^p) I(y_i = y)}{\sum_{i\in[n]} d(x_i^p, x_j^p)} - \frac{\sum_{i\in[n]} d(x_i^p, x_j^p) I(y_i = y \cap x_i^s = s)}{\sum_{i\in[n]} d(x_i^p, x_j^p) I(x_i^s = s)} \Big|.$$

where $d(x_i^p, x_j^p)$ is the edit distance and I is the characteristic function. The idea is to use a locally weighted average to estimate the conditional expectation. The higher is the DTDI score for a dataset D, the more it suffer from disparate treatment, with DTDI $= 0$ meaning that the dataset does not suffer from disparate treatment.

The effect of the Lagrangian Dual framework on reducing disparate treatment was evaluated on three datasets: The *Adult* dataset [Kohavi, 1996], containing 30,000 samples and 23 features, in which the prediction task is that of assessing whether an individual earns more than $50K$ per year and protected attribute is *race*. The *Default* of Taiwanese credit card users [Yeh and Lien, 2009], containing 45,000 samples and 13 features, in which the task is to predict whether an individual will default and the protected attribute is gender. Finally, the *COMPAS* dataset [Angwin *et al.*, 2016], containing 10,500 samples and 14 features, where the task is to predict whether an individual will re-commit a crime and the protected attribute is *race*. The experiments use a 80/20 train/test split and executes a 5-fold cross-validation to evaluate the accuracy and the fairness score (DTDI) of the predictors.

Table 5 illustrates the results comparing model $\mathcal{M}$ that minimizes the Binary Cross Entropy (BCE) loss, model $\mathcal{M}_C$ that includes the Lagrangian loss functions $\mathcal{L}_\lambda$ associated with each constraint and where all λ are fixed to value 1.0, and the proposed model $\mathcal{M}_C^D$ that uses the Lagrangian dual framework to find the optimal Lagrangian weights. All prediction models use a classical feed-forward neural network with 3 layers and 10 hidden units and minimize the Mean Squared Error as loss function. The training uses 100 epochs, Lagrangian step size $s_k = 10^{-4}$ and learning rate 10^{-3}.

The table clearly illustrates the positive effect of the Lagrangian constraints on reducing the DTDI score, and shows that the proposed Lagrangian Dual model outperforms the other models in terms of DTDI score reduction. Importantly, such reduction comes at a contained cost of accuracy loss.

6 Related Work

The application of Deep Learning to constrained optimization problems is receiving increasing attention. Approaches which embed optimization components in neural networks include [Vinyals *et al.*, 2015; Khalil *et al.*, 2017; Kool *et al.*, 2018]. These approaches typically rely on problems exhibiting properties like convexity or submodularity. Another line of work leverages explicit optimization algorithms as a differentiable layer into neural networks [Amos and Kolter, 2017; Donti *et al.*, 2017; Wilder *et al.*, 2019].

Different from these proposal, this paper proposes a framework that exploits key ideas in Lagrangian duality to encourage the satisfaction of generic constraints within a neural network learning cycle. This paper builds on the recent results that were dedicated to learning and optimization in power systems [Fioretto *et al.*, 2020].

7 Conclusions

This paper proposed a Lagrangian dual framework to encourage the satisfaction of constraints in deep learning. It was motivated by a desire to learn parametric constrained optimization problems that feature complex physical and engineering constraints. The paper showed how to transfer Lagrangian dual from optimization to deep learning to obtain predictors that minimize constraint violations. Moreover, it showed that the proposed approach can be applied to constrained learning problems where the learning task imposes constraints on the predictor itself. The Lagrangian Dual Framework for deep learning was evaluated on a collection of realistic energy networks, by enforcing non-discriminatory decisions on a variety of datasets, and on a transprecision computing application. The results demonstrated the effectiveness of the proposed method that dramatically decreases constraint violations (e.g. up to 80% in transprecision computing) committed by the predictors and, in some applications, as in those in

energy optimization, increases the prediction accuracy by up to two orders of magnitude.

References

[Aghaei *et al.*, 2019] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *AAAI*, pages 1418–1426, 2019.

[Amodei *et al.*, 2016] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *ICML*, pages 173–182, 2016.

[Amos and Kolter, 2017] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *ICML*, pages 136–145. JMLR. org, 2017.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica. 2016.

[Baran and Wu, 1989] M. E. Baran and F. F. Wu. Optimal capacitor placement on radial distribution systems. *IEEE Transactions on Power Delivery*, 4(1):725–734, Jan 1989.

[Chowdhury and Rahman, 1990] B. H. Chowdhury and S. Rahman. A review of recent advances in economic dispatch. *IEEE Transactions on Power Systems*, 5(4):1248–1259, Nov 1990.

[Coffrin *et al.*, 2014] Carleton Coffrin, Dan Gordon, and Paul Scott. NESTA, the NICTA energy system test case archive. *CoRR*, abs/1411.0359, 2014.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.

[Donti *et al.*, 2017] Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *NIPS*, pages 5484–5494, 2017.

[Fioretto *et al.*, 2020] Ferdinando Fioretto, Terrence W. K. Mak, and Pascal Van Hentenryck. Predicting ac optimal power flows: Combining deep learning and lagrangian dual methods. In *AAAI*, page to appear, 2020.

[Fisher *et al.*, 2008] E. B. Fisher, R. P. O'Neill, and M. C. Ferris. Optimal transmission switching. *IEEE Transactions on Power Systems*, 23(3):1346–1355, Aug 2008.

[Fontaine *et al.*, 2014] Daniel Fontaine, Michel Laurent, and Pascal Van Hentenryck. Constraint-based lagrangian relaxation. In *Principles and Practice of Constraint Programming*, pages 324–339, 2014.

[Herty *et al.*, 2010] M. Herty, J. Mohring, and V. Sachers. A new model for gas flow in pipe networks. *Mathematical Methods in the Applied Sciences*, 33(7):845–855, 2010.

[Hestenes, 1969] Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

[Khalil *et al.*, 2017] Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, and Le Song. Learning combinatorial optimization algorithms over graphs. In *NIPS*, pages 6348–6358, 2017.

[Kohavi, 1996] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

[Kool *et al.*, 2018] Wouter Kool, Herke Van Hoof, and Max Welling. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*, 2018.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Mak *et al.*, 2019] Terrence W. K. Mak, Pascal Van Hentenryck, Anatoly Zlotnik, and Russell Bent. Dynamic compressor optimization in natural gas pipeline systems. *INFORMS Journal on Computing*, 31(1):40–65, 2019.

[Malossi *et al.*, 2018] A Cristiano I Malossi, Michael Schaffner, and et al. The transprecision computing paradigm: Concept, design, and applications. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018*, pages 1105–1110. IEEE, 2018.

[Monticelli *et al.*, 1987] A Monticelli, MVF Pereira, and S Granville. Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Transactions on Power Systems*, 2(1):175–180, 1987.

[Niharika *et al.*, 2016] Niharika, S. Verma, and V. Mukherjee. Transmission expansion planning: A review. In *International Conference on Energy Efficient Technologies for Sustainability*, pages 350–355, April 2016.

[Pfetsch *et al.*, 2015] M. Pfetsch, A. Fügenschuh, B. Geißler, N. Geißler, R. Gollmer, B. Hiller, J. Humpola, T. Koch, T. Lehmann, A. Martin, A. Morsi, J. Rövekamp, L. Schewe, M Schmidt, R. Schultz, R. Schwarz, J. Schweiger, C. Stangl, M. Steinbach, S. Vigerske, and B. Willert. Validation of nominations in gas network optimization: Models, methods, and solutions. *Optimization Methods and Software*, 30(1):15–53, 2015.

[Vinyals *et al.*, 2015] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *NIPS*, pages 2692–2700, 2015.

[Wilder *et al.*, 2019] Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *AAAI*, volume 33, pages 1658–1665, 2019.

[Yeh and Lien, 2009] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

[Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *International Conference on WWW*, 2017.