NEAR-OPTIMAL SOLUTIONS OF CONSTRAINED LEARNING PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the widespread adoption of machine learning systems, the need to curtail their behavior has become increasingly apparent. This is evidenced by recent advancements towards developing models that satisfy robustness, safety and fairness requirements. Imposing these requirements leads to constrained learning problems, which can be tackled with dual ascent methods. However, convergence guarantees for dual ascent algorithms typically involve a randomized or averaged sequence of primal iterates. These solutions are impractical, since they require storing an ever growing sequence of models. Although it has been observed that final iterates perform well in practice, theoretical guarantees for their optimality and feasibility have remained elusive. In this work, we characterize the infeasibility of Lagrangian minimizers associated with optimal dual variables, which leads to a sub-optimality bound for best primal iterates. To do this, we leverage the fact that constrained learning problems are parametrized versions of convex functional programs. This bound sheds light on how the richness of the parametrization and the curvature of the objective impact the convergence of primal iterates. We empirically validate this finding in learning problems with fairness constraints.

1 INTRODUCTION

Machine learning (ML) has become a core technology of information systems, reaching critical applications from medical diagnostics (Engelhard et al., 2023) to autonomous driving (Kiran et al., 2021). Consequently, it has become paramount to develop ML models that not only excel at a main task, but also adhere to requirements such as fairness and robustness.

Since virtually all ML models are trained using Empirical Risk Minimization (ERM) (Vapnik, 1999), a natural way to impose requirements is to explicitly add constraints to these optimization problems (Fioretto et al., 2021; Velloso & Van Hentenryck, 2020; Cotter et al., 2018; Chamon et al., 2023). Recent works (Chamon & Ribeiro, 2020) have shown that from a PAC (Probably Approximately Correct) perspective, learning under requirements is essentially as hard as classical learning and that it can be done by means of *dual ascent methods*, which only involve solving a sequence of regularized, unconstrained ERM problems. This gave rise to applications across several areas such as federated learning (Shen et al., 2022), fairness (Cotter et al., 2019; Tran et al., 2021), active learning (Elenter et al., 2022), adversarial robustness (Robey et al., 2021) and data augmentation (Hounie et al., 2022).

Despite these statistical guarantees, duality-based algorithms suffer from a severe limitation in terms of recovering feasible solutions. Indeed, dual ascent iterates need not remain in the feasibility set or converge to a fixed point, displaying ciclostationary behaviour. For convex problems, this issue can be tackled using averaging (Nedić & Ozdaglar, 2009). In contrast, the non-convex case requires randomization (Kearns et al., 2018; Agarwal et al., 2018; Goh et al., 2016). These solutions are not only impractical, given the need to store an ever-growing sequence of primal iterates, but also raise ethical considerations. For instance, when selecting job candidates, giving medical diagnosis or deciding whether to give bank loans, randomizing over various predictors could be undesirable from an explainability standpoint.

In fact, this problem is even hard from an algorithmic complexity point of view (Daskalakis et al., 2021). While it has been observed that taking the last or best iterate, can perform well in practice (Cotter et al., 2018; Chamon et al., 2023; Robey et al., 2021; Elenter et al., 2022; Hounie et al., 2022;

Shen et al., 2022), these predictors can also fail miserably, severely violating the requirements that are being imposed.

This work adresses this gap between theory and practice by characterizing the infeasibility of primal iterates associated with optimal dual variables. To do so, we observe that many optimization problems can be seen as parametrized versions of very benign functional optimization problems. We show that for sufficiently rich parametrizations, duality-based solutions are able to closely approximate the feasibilibility of these functional solutions. This implies that dual ascent methods can yield solutions with guaranteed near-feasibility without randomization and despite non-convexity.

2 Problem Formulation

2.1 Constrained Learning

Constrained learning can be formulated as a statistical optimization problem, namely,

$$\begin{aligned} P_p^\star = &\min_{\theta\in\Theta} \ell_0(f_\theta) \\ &\text{s. to: } \ell_i(f_\theta) \leq 0, \quad i = 1, .., m \end{aligned} \tag{P_p}$$

where $f_\theta : \mathcal{X} \to \mathcal{Y}$ is a function associated with the parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$ and the hypothesis class $\mathcal{F}_\theta = \{f_\theta : \theta \in \Theta\}$ induced by these functions is a subset of a compact functional space $\mathcal{F} \subseteq L^2(\Omega)$. The use of the subscript p, for "parametrized", will later become evident. The functionals $\ell_i : \mathcal{F} \to \mathbb{R}$ denote expected risks: $\mathbb{E}_{(x,y)}[\tilde{\ell}_i(f_\theta(x), y)]$, where $\tilde{\ell}_i$ is typically a loss function such as mean squared error or cross entropy loss. In this setting, ℓ_0 can be interpreted as a top-line metric (e.g., accuracy), while the functional $\ell = (\ell_1(f_\theta), \cdots, \ell_m(f_\theta))$ encodes the statistical requirements that the solution must satisfy (see example below). It is important to note that the functionals ℓ_i are typically *non-convex* with respect to θ.

Learning under Counterfactual Fairness Constraints. In ProPublica's COMPAS dataset, the goal is to predict recidivism based on an individual's past offense data while controlling for gender and racial bias. Let $\tilde{\ell}_0$ denote the negative log-likelihood function: $-\log[f_\theta(x)]_y$. The problem of learning a predictor insensitive to perturbations of these protected variables can be formulated as an invariance constrained learning problem:

$$\begin{aligned} P_p^* = &\min_{\theta\in\mathbb{R}^p} \mathbb{E}\left[\tilde{\ell}_0\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)\right] \\ &\text{s. to: } \mathbb{E}_{(x,y)}\left[D_{KL}(f_\theta(x, \mathbf{z})||f_\theta(x, \rho_i(\mathbf{z}))\right] \leq c, \text{ for all } i \end{aligned}$$

where $\mathbf{z}$ contains the protected variables (gender and race), $c > 0$ determines the sensitivity level and the transformations ρ_i encompass all possible single variable modifications of z. This constrained optimization problem is non-convex for many parametrizations (e.g., neural networks) and there is typically no straightforward way to project onto the feasibility set (i.e. the set of classifiers insensitive to gender and race). In light of these challenges, we turn to Lagrangian Duality.

2.2 Dual Constrained Learning

The Lagrangian $L : \mathcal{F} \times \mathbb{R}^m \to \mathbb{R}$ is defined as:

$$L(\phi, \lambda) = \ell_0(\phi) + \lambda^T \ell(\phi), \tag{1}$$

with $\phi \in \mathcal{F}$ and $\lambda \in \mathbb{R}^m_+$. For reasons that will become apparent later we define L over $\mathcal{F}$ rather than $\mathcal{F}_\theta$. For a fixed dual variable λ, the Lagrangian $L(\phi, \lambda)$ is a regularized objective, where ℓ acts as a regularizing functional. This leads to the dual function of problem P_p:

$$g_p(\lambda_p) = \min_{\theta\in\Theta} L(f_\theta, \lambda_p), \tag{2}$$

and to the definition of the dual problem,

$$D_p^\star = \max_{\lambda_p \succeq 0} g_p(\lambda_p). \tag{D_p}$$

Algorithm 1 Dual Constrained Learning

1: *Inputs*: number of iterations $T \in \mathbb{N}$, step size $\eta > 0$.
2: *Initialize*: $\lambda(1) = 0$
3: **for** $t = 1, \dots, T$ **do**
4: Obtain $f_\theta(t)$ such that

$$f_\theta(t) \in \arg\min_{\theta \in \Theta} \ell_0(f_\theta) + \lambda(t)^T \ell(f_\theta)$$

5: Update dual variables

$$\lambda_i(t+1) = \max[0, \lambda_i(t) + \eta\, \ell_i(f_\theta(t))]$$

6: **end for**

This saddle-point problem can be viewed as a two-player game or as a regularized minimization, where the regularization weight is also an optimization variable, updated according to the degree of constraint satisfaction or violation.

The dual function g_p is concave, irrespective of whether P_p is convex. Indeed, it is the pointwise minimum of a family of affine functions on λ. As such, though g_p may not be differentiable, it can be equipped with supergradients, that may be used to determine ascent directions. Explicitly, a vector $s \in \mathbb{R}^m$ is a supergradient of the concave function $h : \mathbb{R}^m \to \mathbb{R}$ at a point x if $h(z) - h(x) \geq s^T(z - x)$ for all z. The set of all supergradients of h at x is called the superdifferential and is denoted $\partial h(x)$. When the losses ℓ_i are continuous, the superdifferential of g_p admits a simple description, namely:

$$\partial g_p(\lambda_p) = \text{conv}\{\ell(f_\theta(\lambda_p)) \ : \ f_\theta(\lambda_p) \in \mathcal{F}_\theta^\star(\lambda_p)\}.$$

where $\text{conv}(S)$ denotes the convex hull of the set S and $\mathcal{F}_\theta^\star(\lambda_p)$ denotes the set of Lagrangian minimizers $f_\theta(\lambda_p)$ associated to the multiplier λ_p:

$$\mathcal{F}_\theta^\star(\lambda_p) = \arg\min_{\theta \in \Theta} L(f_\theta, \lambda_p). \tag{3}$$

We can then proceed to obtain an algorithm for solving D_p, updating dual variables in the ascent direction indicated by a supergradient $\ell(f_\theta(\lambda_p))$ and projecting the resulting iterates into the non-negative orthant. This procedure, referred to as projected supergradient ascent (Polyak, 1987), is presented in Algorithm 1.

Observe that as dual iterates $\lambda_p(t)$ approach the set $\Lambda_p^\star = \arg\max_{\lambda_p \geq 0} g_p(\lambda_p)$ of solutions of D_p, a sequence of primal iterates $\{f_\theta(t) \in \mathcal{F}_\theta^\star(\lambda_p(t))\}_{t=1}^T$ is obtained as a by-product. In general, however, the Lagrangian minimizers are not unique. In particular, for an optimal dual variable $\lambda_p^\star \in \Lambda_p^\star$, the set $\mathcal{F}_\theta^\star(\lambda_p^\star)$ is typically not a singleton and could contain infeasible elements (i.e, $\ell_i(f_\theta(\lambda_p^\star)) > 0$ for some $i \geq 1$). Even more so, as $\lambda_p(t)$ approaches $\Lambda_p^\star$, the constraint satisfaction of primal iterates can exhibit pathological cyclostationary behaviour, where one or more constraints oscillate between feasibility and infeasibility. The experiment in Figure 1 showcases this behaviour and illustrates that, in general, one can not simply stop the dual ascent algorithm at an iteration t and expect the primal iterate $f_\theta(\lambda_p(t))$ to be feasible. This is why, in these type of non-convex problems, guarantees usually pertain a probability distribution over (a subset of) the sequence $\{f_\theta(t)\}_{t=1}^T$ (see e.g, (Agarwal et al., 2018) Theorem 2, (Kearns et al., 2018) Theorem 4.1, (Cotter et al., 2019) Theorem 2, (Chamon et al., 2023)).

3 Near-Optimal Solutions of Constrained Learning Problems

Final primal iterates obtained as a by-product of dual ascent methods can fail at solving problem P_p. However, constrained learning problems arise as parametrized versions of very benign convex functional programs, which are ammenable to a Lagrangian relaxation.

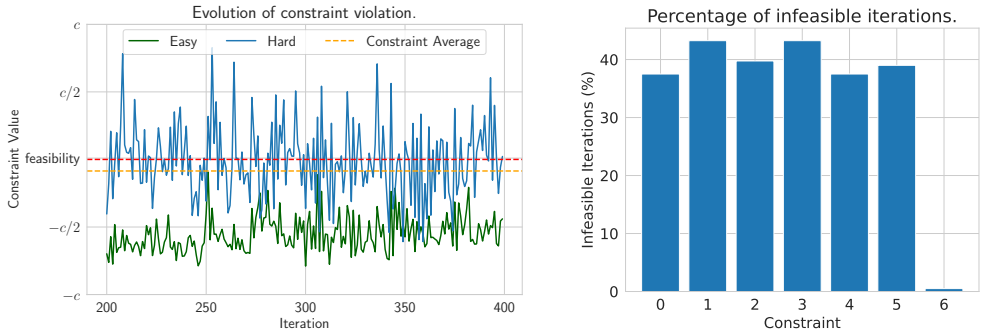

Figure 1: Feastibility of primal iterates in a constrained learning problem with fairness requirements. **Left**: Example of a hard constraint which oscillates between feasibiliy and infeasibility, and an easy constraint which remains feasible. The constraint average line, lying below feasibility threshold, illustrates that averaging primal iterates is beneficial in terms of constraint satisfaction. **Right**: After training accuracy has settled (around half of training epochs) all but the last constraint are infeasible 30-45 % of the time. In fact, on 85% of the iteratons, at least one constraint is violated. We cannot stop the algorithm and expect to arrive at a feasible solution. *This work analyzes the magnitude of feasibility oscillations.*

3.1 The Unparametrized Problem

The *unparametrized* constrained learning problem is defined as:

$$\begin{aligned} P_u^\star = \min_{\phi \in \mathcal{F}} \; & \ell_0(\phi) \\ \text{s.to}: \; & \ell_i(\phi) \leq 0 \quad i = 1, .., m \end{aligned} \tag{P_u}$$

where $\mathcal{F}$ is a convex, compact subset of an $L^2(\Omega)$ space, Ω being a Lebesgue measurable subset of $\mathbb{R}^d$. For instance, $\mathcal{F}$ can be the space of continuous functions or a reproducing kernel Hilbert space (RKHS) and $\mathcal{F}_\theta$ can be a neural network or a finite linear combinations of kernels, both of which meet the uniform approximation assumption (Hornik, 1991; Berlinet & Thomas-Agnan, 2011). Analogously to the definitions presented in section 2.1, $g_u(\lambda_u) := \min_{\phi \in \mathcal{F}} L(\phi, \lambda_u)$ denotes the unparametrized dual function, $\Phi^\star(\lambda_u) = \arg\min_{\phi \in \mathcal{F}} L(\phi, \lambda_u)$ denotes the set of unparametrized Lagrangian minimizers associated with λ_u and

$$D_u^\star = \max_{\lambda_u \succeq 0} \; g_u(\lambda_u) \tag{D_u}$$

is the unparametrized dual problem. We now present two assumptions that allow us to characterize the relation between the dual and primal solutions of problem D_u.

Assumption 3.1 *The functionals ℓ_i , $i = 0, \dots, m$, are convex and $M-$Lipschitz continuous in $\mathcal{F}$. Additionally, ℓ_0 is μ_0-strongly convex.*

Assumption 3.2 *There exists $\phi \in \mathcal{F}$ such that $\ell(\phi) \prec \min[\mathbf{0}, \ell(\phi(\lambda_p^\star)), \ell(f_\theta(\lambda_p^\star))]$, where the minimum is taken coordinate-wise.*

Note that we require convexity of the objective with respect to the functionals, but not model parameters, which holds for both mean squared error and cross-entropy loss. Assumption 3.2 is a stronger version of Slater's constraint qualification, since it requires strict feasibility in a set of perturbed versions of the unparametrized problem. It will later allow us to analyze the variations of the optimal value $P_u^\star$ as a function of the constraint tightness.

As will be explained in section 3.3, under these assumptions, the unparametrized Lagrangian minimizer is unique. This makes the superdifferential of the dual function a singleton at every λ_u: $\partial g_u(\lambda_u) = \{\ell(\phi(\lambda_u))\}$, which means that the dual function $g_u(\lambda_u)$ is differentiable. Let $\phi^\star$ be a solution of problem P_u. Assumptions 3.1 and 3.2 imply that strong duality (i.e, $P_u^* = D_u^\star$) holds in this problem, and that at $\lambda_u^\star$, there is a unique Lagrangian minimizer $\phi^\star(\lambda_u^\star) = \phi^\star$ which is, by definition, feasible.

Note that the only difference between problems P_p and P_u is the nature of the set over which the optimization is carried out. Thus, if the parametrization Θ is rich (e.g, deep neural networks), the set $\mathcal{F}_\theta$ is close to $\mathcal{F}$, and we can expect the properties of the solutions to problems D_p and D_u to be similar. This insight leads us to the $\nu-$near universality of the parametrization assumption.

Assumption 3.3 *For all $\phi \in \mathcal{F}$, there exists $\theta \in \Theta$ such that $\|\phi - f_\theta\|_{L_2} \leq \nu$.*

Given the properties of the problems presented, it is legitimate to ask: how close are $f_\theta(\lambda_p^\star)$ and $\phi^\star$ in terms of their constraint satisfaction ? Should a tight bound exist, then averaging and randomization would not be necessary.

3.2 Feasibility Approximation

We will characterize the constraint violation of the Lagrangian minimizers $f_\theta(\lambda_p^\star) \in \mathcal{F}_\theta(\lambda_p^\star)$ by comparing these primal variables with the solution of the unparametrized problem: $\phi^\star$. Since the problem is feasible, $\ell(\phi^\star)$ is non-positive and, due to complementary slackness, it is a null vector when all constraints are active.

The curvature of the unparametrtized dual function $g_u(\lambda_u)$ around the optimum is central in this analysis. We will first provide a result assuming this curvature is known, and will later describe its connection to the properties of Problem P_p. Let $\mathcal{B}_\lambda := \{\gamma\lambda_u^\star + (1-\gamma)\lambda_p^\star \,:\, \gamma \in [0,1]\}$ denote the segment connecting $\lambda_u^\star$ and $\lambda_p^\star$.

Assumption 3.4 *The dual function g_u is μ_g-strongly concave on $\mathcal{B}_\lambda$ and β_g-smooth.*

We now state the main result of this section, characterizing the constraint satisfaction of any Lagrangian minimizer of the constrained learning problem at an optimal dual variable $\lambda_p^\star \in \Lambda_p^\star$ with respect to that of the optimal, feasible solution of the unparametrized problem.

Theorem 3.5 *Under assumptions 3.1-3.4, for any $f_\theta(\lambda_p^\star) \in \mathcal{F}_\theta^\star(\lambda_p^\star)$, the distance between the unparametrized and parametrized constraint violations is bounded by:*

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi^\star)\|_2^2 \leq 2\beta_g M\nu(1 + \|\lambda_p^\star\|_1)\left(1 + \sqrt{\frac{\beta_g}{\mu_g}}\right)^2$$

The ratio $\frac{\beta_g}{\mu_g}$, which corresponds to the condition number of the Hessian of $g_u(\lambda)$, can be used to determine optimal step sizes in dual ascent methods (Polyak, 1987), and plays a crucial role in this bound. As will be shown in Section 3.3 if the dual function is steeply curved (i.e, μ_g is large) then $\lambda_p^\star$ and $\lambda_u^\star$ are not too far apart. However, as μ_g increases so does β_g, which increases the sensitivity of the optimum P_u^* with respect to constraint perturbations and loosens the bound.

The constant ν indicates how rich the parametrization is in terms of approximating the functions of $\mathcal{F}$. Thus, it is reasonable that as the model capacity increases and ν goes to 0, the distance between constraint violations decreases. In the extreme, if all functions in $\mathcal{F}$ can be exactly approximated by an element of $\mathcal{F}_\theta$ ($\nu = 0$), the problems P_p and P_u are equivalent and trivially $\phi^\star = f_\theta(\lambda_p^\star)$.

To better understand how the properties of problem P_p impact the feasibility approximation bound in Theorem 3.5, we now the relate the curvature of $g_u(\lambda_u)$ to the properties of the losses ℓ_i.

Assumption 3.6 *The functionals ℓ_i , $i = 0, \dots, m$ are β-smooth in $\mathcal{F}$.*

Assumption 3.7 *The Jacobian $D_\phi\ell(\phi^\star)$ is full-row rank at the optimum, i.e: $\exists\sigma > 0$ such that $\inf_{\|\lambda\|=1} \|\lambda^T D_\phi\ell(\phi^\star)\|_{L_2} \geq \sigma$, where $D_\phi\ell(\phi^\star)$ denotes the Frechet derivative of the functional ℓ at $\phi^\star$ (see Appendix A.1).*

Assumption 3.7, which lower-bounds the singular values of the constraint Jacobian, is customary in constrained optimization and is related to the Linear independence constraint qualification (LICQ).

Lemma 3.8 *Under assumptions 3.1, 3.2, 3.6 and 3.7, the unparametrized dual function $g_u(\lambda_u)$ is μ_g-strongly concave on $\mathcal{B}_\lambda$ and β_g-smooth with:*

$$\mu_g = \frac{\mu_0\, \sigma^2}{\beta^2(1+\Delta)^2}, \qquad \beta_g = \frac{\sqrt{m}M^2}{\mu_0} \tag{4}$$

where $\Delta = \max\{\|\lambda_u^\star\|_1, \|\lambda_p^\star\|_1\}$.

From Lemma 3.8, we have that $\frac{\beta_g}{\mu_g} = \sqrt{m}\frac{M^2}{\sigma^2}\frac{\beta^2}{\mu_0^2}(1+\Delta)^2$. Therefore, the condition numbers $\kappa_1 = \frac{M}{\sigma}$ and $\kappa_2 = \frac{\beta}{\mu_0}$ of the constraint Jacobian and the objective's Hessian impact how far the constraint violations of $f_\theta(\lambda_p^\star)$ are to that of $\phi^\star$. Combining this with Theorem 3.5, we obtain the following infinity norm bound:

Corollary 3.9 *Under assumptions 3.1, 3.2, 3.3, 3.6 and 3.7, the maximum distance between the parametrized and unparametrized contraints violations is characterized by:*

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi^\star)\|_\infty \leq M\ [1+\kappa_1\kappa_2(1+\Delta)]\ \sqrt{2m\frac{M\nu}{\mu_0}(1+\|\lambda_p^\star\|_1)} \tag{5}$$

with $\kappa_1 = \frac{M}{\sigma}$, $\kappa_2 = \frac{\beta}{\mu_0}$ *and* $\Delta = \max\{\|\lambda_u^\star\|_1, \|\lambda_p^\star\|_1\}$.

This bound can be split into three main components. The condition numbers $\kappa_1 = \frac{M}{\sigma}$ and $\kappa_2 = \frac{\beta}{\mu_0}$ of the constraint Jacobian and the objective's Hessian are present in the term $[1+\kappa_1\kappa_2(1+\Delta)]$. Thus, it can be thought of as the baseline effect, capturing how well-conditioned the problem is. Furthermore, the term $\sqrt{2m\frac{M\nu}{\mu_0}(1+\|\lambda_p^\star\|_1)}$ contains the approximation error in PACC learning (see (Chamon & Ribeiro, 2020), Theorem 2). Although the number of constraints m naturally appears in this bound, it can be integrated into the term M if one makes the stronger assumption that ℓ is $M-$Lipschitz, as opposed to assuming this of each individual ℓ_i.

3.3 Dual Variable and Hypothesis Class Perturbations

In this section, we give an outline on the results that build up to Theorem 3.5. We will focus on the properties that shed light on the nature of the Lagrangian minimizers $f_\theta(\lambda_p^\star) \in \mathcal{F}_\theta^\star(\lambda_p^\star)$. For clarity, we will sometimes write $\phi^\star$ as $\phi(\lambda_u^\star)$ to emphasize the dependence on $\lambda_u^\star$. We start by decomposing the distance between parametrized and unparametrized contraints violations using a triangle inequality:

$$\begin{aligned}\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2 &= \|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_p^\star)) + \ell(\phi(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2 \\ &\leq \|\ell(\phi(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2 + \|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_p^\star))\|_2\end{aligned}$$

The first term captures the impact of the perturbation of dual variables in the unparametrized problem. The second one captures the effect of parametrizating the hypothesis class for a fixed dual variable. For conciseness, technical definitions used in this section are deferred to Appendix A.1.

3.3.1 Dual variable perturbation

In the unparametrized problem the Lagrangian minimizer $\phi(\lambda_u)$ is unique at each λ_u, which implies that $g_u(\lambda)$ is everywhere differentiable with gradient $\nabla_\lambda g_u(\lambda) = \ell(\phi(\lambda))$ (see Appendix A.3). In this setting, analyzing the term $\ell(\phi(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))$ is equivalent to studying the distance between the gradients of g_u at $\lambda_p^\star$ and $\lambda_u^\star$. Furthermore, leveraging the $\nu-$near universality of the parametrization we can show that the maximizers of g_u and g_p cannot be too far apart. In fact, as shown in Appendix A.4, their distance is characterized by:

$$\|\lambda_p^\star - \lambda_u^\star\|_2^2 \leq 2\frac{M\nu}{\mu_g}(1+\|\lambda_p^\star\|_1). \tag{6}$$

Optimal dual variables indicate the sensitivity of the optimal value with respect to constraint perturbations. Thus, the term $(1+\|\lambda_p^\star\|_1)$ can be seen as an indicator of the sensitivity of the optimization

problem. Combining the bound in equation in equation 6 with the β_g-smoothness of g_u, we can characterize the impact of going from $\lambda_u^\star$ to $\lambda_p^\star$ in the unparametrized problem.

Proposition 3.10 *Under assumptions 3.1-3.4, the distance between the constraint violations $\phi(\lambda_p^\star)$ and $\phi(\lambda_u^\star)$ is bounded by:*

$$\|\ell(\phi(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2^2 \leq 2\frac{\beta_g^2}{\mu_g} M\nu(1 + \|\lambda_p^\star\|_1) \tag{7}$$

As the constant $\frac{\beta_g}{\mu_g}$ grows, and the Hessian of g_u becomes ill-conditioned, the bound becomes looser. As shown in Lemma 3.8, this occurs if the number of constraints grows, or if the curvature of the objective ℓ_0 decreases.

3.3.2 Hypothesis class perturbation

We now consider a perturbed version of the unparametrized problem. Its optimal value - or perturbation- function is defined as:

$$\begin{aligned} P^\star(\epsilon) = &\min_{\phi \in \mathcal{F}} \ell_0(\phi) \\ &\text{s.to : } \ell(\phi) + \epsilon \preceq 0 \end{aligned} \tag{P_ϵ}$$

for a perturbation $\epsilon \in \mathbb{R}^m$. Intuitively, increasing ϵ (coordinate-wise) tightens the constraint, making the feasible set smaller and potentially increasing $P^\star(\epsilon)$. Note that by setting $\epsilon = 0$, we recover the unparametrized problem: $P^*(0) = P_u^*$.

By focusing on particular instances of this problem with perturbations $\epsilon_u = -\ell(\phi(\lambda_p^\star))$ and $\epsilon_p = -\ell(f_\theta(\lambda_p^\star))$ we can bound the distance between the constraint violations of the Lagrangian minimizers associated to $\lambda_p^* \in \Lambda_p^*$ in the parametrized and unparametrized problems. To do this, we start by analyzing the variations of the optimal value function $P^\star(\epsilon)$ using a well-known result from conjugate duality: for every $\lambda \in \mathbb{R}_+^m$ we have that $P^\dagger(\lambda) = -g_u(\lambda)$, where $P^\dagger$ denotes the Fenchel conjugate of the perturbation function $P^*(\epsilon)$. In fact, the dual problem is sometimes defined as the maximization of $-P^\dagger(\lambda)$ (Rockafellar, 1974). This result is impactful, because it allows us to relate the curvature of $g_u(\lambda)$, which we know from Lemma 3.8, to the variations of $P^*(\epsilon)$. Specifically, we leverage the *duality between smoothness and strong convexity*.

Informally, a closed convex function h is strongly convex with constant μ if and only if its Fenchel conjugate $h^\dagger$ is $\frac{1}{\mu}-$smooth (Kakade et al., 2009). Since g_u is β_g-smooth, this implies that $P^*(\epsilon)$ is $1/\beta_g-$strongly convex in the region of interest (see Appendix A.8). However, for this result to hold, the strict feasibility assumption 3.2 is indispensable. Indeed, the fact that the perturbed problems are strictly feasible implies that perturbation function $P^\star(\epsilon)$ is closed, a property needed to characterize its variations (see Appendix A.8).

Furthermore, since we know that $f_\theta(\lambda_p^\star)$ is feasible for the problem with perturbation ϵ_p, we can use weak duality to bound the distance between $P^*(\epsilon_p)$ and $P^*(\epsilon_u)$ (see Appendix A.10). Combining these results, we can describe the impact of the parametrizion for a fixed dual variable $\lambda_p^\star \in \Lambda_p^\star$.

Proposition 3.11 *Under assumptions 3.1-3.4, the distance between constraint violation associated to the parametrization of the hypothesis class is given by:*

$$\|\ell(\phi(\lambda_p^\star)) - \ell(f_\theta(\lambda_p^\star))\|_2^2 \leq 2\beta_g M\nu(1 + \|\lambda_p^\star\|_1)$$

In contrast to Proposition 3.10, the strong concavity of the dual function is not present in Proposition 3.11, which means that the smoothness of the losses ℓ_i does not play a role in this bound.

4 Best Iterate Convergence

As described in section 2.1, in constrained learning, the objective ℓ_0 and the constraints ℓ are statistical in nature: $\ell_i(f_\theta) = \mathbb{E}_{(x,y)}[\tilde{\ell}_i(f(x), y)]$ where (x, y) is sampled from a distribution $\mathfrak{D}$. In practice,

we do not have access to $\mathfrak{D}$, but to a set of samples $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, assumed iid. We will denote by $\hat{\ell}_i(f_\theta)$ an estimate of $\ell_i(f_\theta)$ using the dataset D: $\hat{\ell}_i(f_\theta) := \frac{1}{n}\sum_{k=1}^{n} \tilde{\ell}_i(f_\theta(x_k), y_k)$.

Estimating expectations with sample means does not modify Algorithm 1 significantly. Since $\hat{\ell}_i(f_\theta(t))$ is an unbiased estimator of $\ell_i(f_\theta(t))$, a stochastic supergradient $\hat{s}(t) := \hat{\ell}_i(f_\theta(t))$ of g_p can be obtained using samples from D. Thus, in the stochastic version of the supergradient ascent algorithm (Shor, 2013), the dual update can be written as:

$$\lambda_p(t+1) = [\lambda_p(t) + \eta\, \hat{s}(t)]_+ \tag{8}$$

Observe that since $\mathbb{E}_{(x,y)}\{\hat{s}(t)|\lambda(t)\} \in \partial g_p(\lambda(t))$, dual variables will move, on average, towards the optimal set $\Lambda^\star$.

We now analyze the convergence of the best dual iterate, that is, the dual variable which evaluates to the largest dual function encountered so far. More precisely, we show that best dual iterates enter a near-optimality ball infinitely often. Since the dual function can be evaluated, this translates into a practical algorithm, analogous to a validation step in standard supervised learning. We then use Propositions 3.10 and 3.11 to obtain a bound on the infeasibility of primal variables associated to the best dual iterate.

Lemma 4.1 *Let $g_p^{best}(t|\lambda(t_0)) = \max_{s\in[t_0,t]} g_p(\lambda(s))$ be the maximum value of the parametrized dual function up to time t. Then,*

$$\lim_{t\to\infty} g_p^{best}(t|\lambda(t_0)) \geq D_p^\star - \frac{\eta S^2}{2} \quad a.s.$$

where $S^2 > \mathbb{E}[\|\hat{s}(t)\|^2|\lambda(t)]$ is an upper bound on the norm of the second order moment of the stochastic supergradients.

Observe that the existence of S^2 is guaranteed by the Lipschitz continuity of the losses ℓ_i and the boundedness of the set $\mathcal{F}_\theta \subseteq \mathcal{F}$. Since S^2 is finite, one can reduce the step size η to make g_p^{best} arbitrarily close to $D_p^\star$, with the potential cost of increasing the time of occurrence of such proximity.

Lemma 4.1 implies that for almost every realization, and arbitrary $\delta > 0$, as t grows $g_p^{\text{best}}(t|\lambda(t_0))$ is $\frac{\eta S^2}{2} + \delta$ close to $D_p^\star$ at least once. Since t_0 is arbitrary, this occurs infinitely often. Let λ^{best} be a dual iterate such that: $g_p(\lambda^{\text{best}}) \geq D_p^\star - (\frac{\eta S^2}{2} + \delta)$. The near-optimality of $\lambda^{\text{best}})$ and the results from section 3.3 allow us to derive a bound on the constraint violation of primal iterates associated to λ^{best} in the parametrized and unparametrized problems.

Proposition 4.2 *Let λ^{best} be a dual iterate such that: $g_p(\lambda^{best}) \geq D_p^\star - (\frac{\eta S^2}{2} + \delta)$. Under assumptions 3.1, 3.3, 3.3, 3.7 and assuming that there exists $\phi \in \mathcal{F}$ such that $\ell(\phi) \prec \min\{\mathbf{0}, \ell(\phi(\lambda^{best})), \ell(f_\theta(\lambda^{best}))\}$ we have:*

$$\|\ell(\phi^\star) - \ell(f_\theta(\lambda^{best}))\|_2^2 \leq 2\beta_g M\nu(1 + \|\lambda^{best}\|_1)\left(1 + (1 + \frac{\eta S^2}{2} + \delta)\sqrt{\frac{\beta_g}{\tilde{\mu}_g}}\right)^2$$

where $\tilde{\mu}_g = \frac{\mu_0\, \sigma^2}{\beta^2(1+\max\{\|\lambda_u^\star\|_1, \|\lambda^{best}\|_1\})^2}$

The main difference between Proposition 4.2 and the main Theorem 3.5 is that the condition number $\frac{\beta_g}{\tilde{\mu}_g}$ gets amplified by the sub-optimality of λ^{best} with respect to $\lambda_p^\star$.

5 Experimental Validation

To validate the theoretical findings of sections 3 and 4 we return to example 2.1, a constrained learning problem with counterfactual fairness requirements. In the COMPAS dataset, the protected variables gender and race can take the values ["Male", "Female"] and ["African American", "Hispanic", "Caucasian", "Other"] respectively. We use a two-layer neural network with 64 nodes and

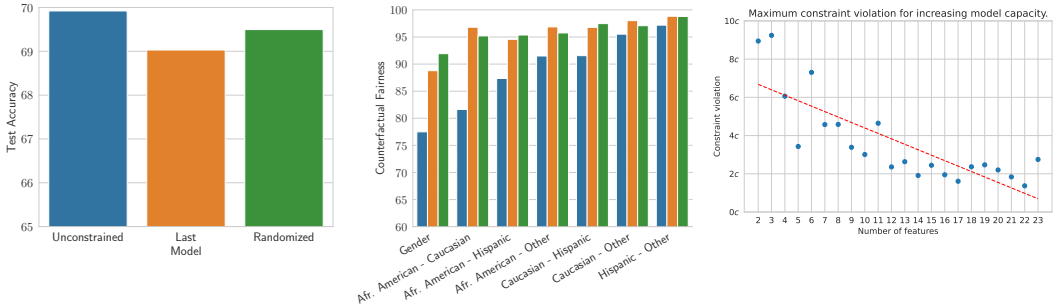

Figure 2: **Left:** the Unconstrained model performs better in terms of average test accuracy than both the Last and Randomized model. **Middle:** Both constrained models do better in terms of Counterfactual Fairness. The key point is that the last model is never far from the Randomized one. **Right**: As the richness of the parametrization increases the maximum constraint violation (i.e: size of the oscillations) decreases.

sigmoid activations. The training objective ℓ_0 is the negative log-likelihood and the constraint upper bound is set to 0.001. We train this model over $T = 400$ iterations using a ADAM, with batch size 256, primal learning rate 0.1, and dual variable learning rate 2. The objective ℓ_0 is the negative log-likelihood. We use the same data pre-processing steps as in (Chamon & Ribeiro, 2020).

In this setting, we compare the performance and constraint satisfaction of three predictors. An *Unconstrained* predictor: trained without contemplating the fairness requirements through empirical risk minimization. The *Last* predictor: corresponding to the final iterate $f_\theta(\lambda(T))$ of the stochastic dual supergradient ascent method. The *Randomized* predictor: which takes the sequence of primal iterates $\{f_\theta(\lambda(t))\}_{t=t_0}^{T}$ and samples a learner uniformly from this sequence in order to make a prediction. We take t_0 as the iteration where training accuracy settles, which corresponds to around half of the training iterations.

As shown in Figure 2, the unconstrained model is slightly better than the constrained ones, although this difference is small ($< 1\%$). Furthermore, the unconstrained model is significantly worse in terms of counterfactual fairness than both the Last and Randomized Predictors, which are always close, in agreement with Theorem 3.5. We also perform an ablation on how the richness of the parametrization impacts the maximum violation of primal iterates. To control the richness of the parametrization we project the input samples into a space of lower-dimension with a fixed, random linear map. These lower-dimensional vectors are then used to train the neural network. As the richness of the parametrization increases, ν decreases, reducing the upper bound in Theorem 3.5. This is illustrated by the right-most plot in Figure 2, where the maximum constraint violation (i.e: magnitude of the feasibility oscillations) decreases by an order of magnitude as we progressively increase the capacity of the model.

6 Conclusion

We analyzed the feasibility of primal iterates obtained from a dual ascent method when solving the Lagrangian dual of a primal non-convex constrained learning problem. The primal problem in question is the parametrized version of a convex functional program, which is amenable to a Lagrangian relaxation. Specifically, we characterized how far these predictors are from a solution of the unparametrized problem in terms of their constraint violations. This result led to a characterization of the infeasibility of best primal iterates and elucidated the role of the capacity of the model and the curvature of the objective. These guarantees bridge a gap between theory and practice in constrained learning, shedding light on when and why randomization is unnecessary.

The findings presented in this work can be extended in several ways. For instance, the estimation error incurred by using samples to estimate statistical losses can be included in the main analysis. Moreover, feasibility results studying the primal iterates directly, and not on their distance to the solution of the unparametrized problem can be obtained. Finally, it might be possible to lift one or more assumptions about the unparametrized problem.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

J. Frédéric Bonnans and Alexander Shapiro. Optimization problems with perturbations: A guided tour. *SIAM Review*, 40(2):228–264, 1998. ISSN 00361445. URL http://www.jstor.org/stable/2653333.

Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16722–16735. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c291b01517f3e6797c774c306591cc32-Paper.pdf.

Luiz F. O. Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2023. doi: 10.1109/TIT.2022.3187948.

Andrew Cotter, Maya R. Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Lutong Wang, Blake E. Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:49556538.

Andrew Cotter, Heinrich Jiang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20(172):1–59, 2019.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1466–1478, 2021.

Juan Elenter, Navid NaderiAlizadeh, and Alejandro Ribeiro. A lagrangian duality approach to active learning, 2022.

Matthew M. Engelhard, Ricardo Henao, Samuel I. Berchuck, Junya Chen, Brian Eichner, Darby Herkert, Scott H. Kollins, Andrew Olson, Eliana M. Perrin, Ursula Rogers, Connor Sullivan, YiQin Zhu, Guillermo Sapiro, and Geraldine Dawson. Predictive Value of Early Autism Detection Models Based on Electronic Health Record Data Collected Before Age 1 Year. *JAMA Network Open*, 6(2):e2254303–e2254303, 02 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.54303. URL https://doi.org/10.1001/jamanetworkopen.2022.54303.

Ferdinando Fioretto, Pascal Van Hentenryck, Terrence W. K. Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke (eds.), *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pp. 118–135, Cham, 2021. Springer International Publishing. ISBN 978-3-030-67670-4.

Rafal Goebel and R Tyrrell Rockafellar. Local strong convexity and local lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263, 2008.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. *Advances in neural information processing systems*, 29, 2016.

Vincent Guigues. Inexact stochastic mirror descent for two-stage nonlinear stochastic programs, 2020.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

Ignacio Hounie, Luiz F. O. Chamon, and Alejandro Ribeiro. Automatic data augmentation via invariance-constrained learning, 2022.

Sham Kakade, Shai Shalev-Shwartz, Ambuj Tewari, et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1):35, 2009.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Andrew J Kurdila and Michael Zabarankin. *Convex functional analysis*. Springer Science & Business Media, 2006.

Angelia Nedić and Asuman Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009. doi: 10.1137/070708111. URL https://doi.org/10.1137/070708111.

Boris T Polyak. Introduction to optimization. 1987.

Alejandro Ribeiro. Ergodic stochastic optimization algorithms for wireless communication and networking. *IEEE Transactions on Signal Processing*, 58(12):6369–6386, 2010. doi: 10.1109/TSP.2010.2057247.

Alexander Robey, Luiz Chamon, George J. Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial robustness with semi-infinite constrained learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6198–6215. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/312ecfdfa8b239e076b114498ce21905-Paper.pdf.

R Tyrrell Rockafellar. *Conjugate duality and optimization*. SIAM, 1974.

R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Xo0lbDt975.

N.Z. Shor. *Nondifferentiable Optimization and Polynomial Problems*. Nonconvex Optimization and Its Applications. Springer US, 2013. ISBN 9781475760156. URL https://books.google.com/books?id=_L_VBwAAQBAJ.

Victor Solo and Xuan Kong. Adaptive signal processing algorithms: Stability and performance. 1994. URL https://api.semanticscholar.org/CorpusID:61115048.

Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9932–9939, 2021.

V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10 (5):988–999, 1999. doi: 10.1109/72.788640.

Alexandre Velloso and Pascal Van Hentenryck. Combining deep learning and optimization for security-constrained optimal power flow. *arXiv preprint arXiv:2007.07002*, 2020.

A APPENDIX

A.1 ADDITIONAL DEFINITIONS

Definition A.1 *We say that a functional $\ell_i : \mathcal{F} \to \mathbb{R}$ is* Fréchet differentiable *at $\phi^0 \in \mathcal{F}$ if there exists an operator $D_\phi \ell_i(\phi^0) \in \mathfrak{B}(\mathcal{F}, \mathbb{R})$ such that:*

$$\lim_{h \to 0} \frac{|\ell_i(\phi^0 + h) - \ell_i(\phi^0) - \langle D_\phi \ell_i(\phi^0), h \rangle|}{\|h\|_{L_2}} = 0$$

where $\mathfrak{B}(\mathcal{F}, \mathbb{R})$ denotes the space of bounded linear operators from $\mathcal{F}$ to $\mathbb{R}$.

The space $\mathfrak{B}(\mathcal{F}, \mathbb{R})$, algebraic dual of $\mathcal{F}$, is equipped with the corresponding dual norm:

$$\|B\|_{L_2} = \sup \left\{ \frac{|\langle B, \phi \rangle|}{\|\phi\|_{L_2}} \ : \ \phi \in \mathcal{F}, \ \|\phi\|_{L_2} \neq 0 \right\}$$

which coincides with the L_2-norm through Riesz's Representation Theorem: there exists a unique $g \in \mathcal{F}$ such that $B(\phi) = \langle \phi, g \rangle$ for all ϕ and $\|B\|_{L_2} = \|g\|_{L_2}$.

Definition A.2 *A function $h : \mathcal{X} \to \mathbb{R}$ is said to be closed if for each $\alpha \in \mathbb{R}$, the sublevel set $\{h(x) \leq \alpha \ : \ x \in \mathcal{X}\}$ is a closed set.*

Definition A.3 *A convex function $h : \mathcal{X} \to \mathbb{R}$ is proper if $h(x) > -\infty$ for all $x \in \mathcal{X}$ and there exists $x_0 \in \mathcal{X}$ such that $h(x_0) < +\infty$.*

Definition A.4 *Let $\mathcal{X}$ be an Euclidean vector space. Given a convex function $h : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, its Fenchel conjugate $h^\dagger : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is defined as:*

$$h^\dagger(y) = \sup_{x \in \mathcal{X}} \langle x, y \rangle - h(x)$$

A.2 PROOF LEMMA A.5

Lemma A.5 *The point-wise distance between the parametrized and unparametrized dual functions is bounded by:*

$$0 \leq g_p(\lambda) - g_u(\lambda) \leq M\nu(1 + \|\lambda\|_1) \qquad \forall \quad \lambda \succeq 0 \tag{9}$$

As defined in section 2.1, $\phi(\lambda)$ denotes the Lagrangian minimizer associated to the multiplier λ in the unparametrized problem.

By the near-universality assumption, $\exists \, \tilde{\theta} \in \Theta$ such that $\|\phi(\lambda) - f_{\tilde{\theta}}\|_{L_2} \leq \nu$. Note that,

$$\begin{aligned} L(f_{\tilde{\theta}}, \lambda) - L(\phi(\lambda), \lambda) &= \ell_0(f_{\tilde{\theta}}) - \ell(\phi(\lambda)) + \lambda^T \left(\ell(f_{\tilde{\theta}}) - \ell(\phi(\lambda)) \right) \\ &\leq \|\ell_0(f_{\tilde{\theta}}) - \ell(\phi(\lambda))\|_2 + \sum_{i=1}^m [\lambda]_i \|\ell(f_{\tilde{\theta}}) - \ell(\phi(\lambda))\|_2 \end{aligned}$$

where we used the triangle inequality twice. Then, using the $M-$Lipschitz continuity of the functionals ℓ_i and the fact that $\|\phi(\lambda) - f_{\tilde{\theta}}\|_2 \leq \nu$, we obtain:

$$\begin{aligned} L(f_{\tilde{\theta}}, \lambda) - L(\phi(\lambda), \lambda) &\leq M\|f_{\tilde{\theta}} - \phi(\lambda)\|_{L_2} + M \sum_{i=1}^m [\lambda]_i \|f_{\tilde{\theta}} - \phi(\lambda)\|_{L_2} \\ &\leq M\nu + M\nu \sum_{i=1}^m [\lambda]_i = M\nu(1 + \|\lambda\|_1) \end{aligned}$$

Since $f_\theta(\lambda) \in \mathcal{F}_\theta^\star(\lambda)$ is a Lagrangian minimizer, we know that $L(f_\theta(\lambda), \lambda) \leq L(f_{\tilde{\theta}}, \lambda)$. Thus,

$$0 \leq L(f_\theta(\lambda), \lambda) - L(\phi(\lambda), \lambda) \leq L(f_{\tilde{\theta}}, \lambda) - L(\phi(\lambda), \lambda)$$

where the non-negativity comes from the fact that $\mathcal{F}_\Theta \subseteq \mathcal{F}$. This implies:

$$0 \leq g_p(\lambda) - g_u(\lambda) \leq M\nu(\|\lambda\|_1 + 1) \qquad \forall \quad \lambda \succeq 0$$

which conludes the proof.

A.3 LEMMA A.6: DIFFERENTIABILITY OF $g_u(\lambda_u)$

Lemma A.6 *Under assumption 3.1, the unparametrized dual function $g_u(\lambda)$ is everywhere differentiable with gradient $\nabla_\lambda g_u(\lambda) = \ell(\phi(\lambda))$.*

From assumption 3.1, $\ell(\phi)$ is strongly convex and $\lambda^T \ell(\phi)$ is a non-negative combination of convex functions. Thus, the Lagrangian $L(f, \lambda)$ is strongly convex on ϕ for any fixed dual variable $\lambda \in \mathbb{R}^m_+$.

The convexity and compactness of $\mathcal{F}$ imply that, in the unparametrized problem, the Lagrangian functional attains its minimizer $\phi(\lambda)$ for each λ. (see e.g, (Kurdila & Zabarankin, 2006) Theorem 7.3.1.) Then, by the strong convexity of $L(\phi, \lambda)$, this minimizer is unique.

Since $L(f, \lambda)$ is affine on λ, it is differentiable on λ. Then, by application of the Generalized Danskin's Theorem (see e.g: (Başar & Bernhard, 2008) Corollary 10.1) to $g_u(\lambda)$ and using that the set of minimizers $\phi(\lambda)$ of $L(f, \lambda)$ is a singleton, we obtain:

$$\nabla_\lambda g_u(\lambda) = \ell(\phi(\lambda)),$$

which completes the proof.

A.4 PROOF LEMMA A.7

Lemma A.7 *Under assumptions 3.1, 3.2, 3.3, 3.4, the proximity between the unparametrized and parametrized optimal dual variables is characterized by:*

$$\|\lambda_p^\star - \lambda_u^\star\|_2^2 \leq 2\frac{M\nu}{\mu_g}(1 + \|\lambda_p^\star\|_1) \tag{10}$$

Since $g_u(\lambda)$ is differentiable (see A.6) and μ_g-strongly concave for $\lambda \in \mathcal{B}_\lambda$:

$$g_u(\lambda) \leq g_u(\lambda_u^\star) + \nabla g_u(\lambda_u^\star)^T(\lambda - \lambda_u^\star) - \frac{\mu_g}{2}\|\lambda - \lambda_u^\star\|_2^2 \quad \forall \lambda \in \mathcal{B}_\lambda$$

From Lemma A.6 we have that $\nabla g_u(\lambda_u^\star) = \ell(\phi(\lambda_u^\star)))$, then evaluating at $\lambda_p^\star$ we obtain:

$$g_u(\lambda_p^\star) \leq g_u(\lambda_u^\star) + \ell(\phi(\lambda_u^\star))^T(\lambda_p^\star - \lambda_u^\star) - \frac{\mu_g}{2}\|\lambda_p^\star - \lambda_u^\star\|_2^2$$

By complementary slackness, $\ell(\phi(\lambda_u^\star))^T\lambda_u^\star = 0$. Then, since $\phi(\lambda_u^\star)$ is feasible and $\lambda_p^\star \geq 0$: $\ell(\phi(\lambda_u^\star))^T\lambda_p^\star \leq 0$. Thus,

$$g_u(\lambda_p^\star) \leq g_u(\lambda_u^\star) - \frac{\mu_g}{2}\|\lambda_p^\star - \lambda_u^\star\|_2^2$$

By Proposition 1: $g_p(\lambda_p^\star) - M\nu(1 + \|\lambda_p^\star\|_1) \leq g_u(\lambda_p^\star)$, which implies:

$$g_p(\lambda_p^\star) - M\nu(1 + \|\lambda_p^\star\|_1) \leq g_u(\lambda_u^\star) - \frac{\mu_g}{2}\|\lambda_p^\star - \lambda_u^\star\|_2^2$$

Thus,

$$\|\lambda_p^\star - \lambda_u^\star\|_2^2 \leq \frac{2}{\mu_g}\left[g_u(\lambda_u^\star) - g_p(\lambda_p^\star)\right] + \frac{2}{\mu_g}M\nu(1 + \|\lambda_p^\star\|_1) \tag{11}$$

Finally, since $\mathcal{F}_\theta \subseteq \mathcal{F}$ we have that : $g_u(\lambda) \leq g_p(\lambda) \quad \forall\, \lambda$. Evaluating at $\lambda_u^\star$ and using that $\lambda_p^\star$ maximizes g_p we obtain:

$$\begin{aligned} g_u(\lambda_u^\star) &\leq g_p(\lambda_u^\star) \\ &\leq g_p(\lambda_p^\star) \end{aligned}$$

Using this in equation 11 we obtain,

$$\|\lambda_p^\star - \lambda_u^\star\|_2^2 \leq \frac{2}{\mu_g}M\nu(1 + \|\lambda_p^\star\|_1)$$

A.5 PROOF THEOREM 3.10

The proof follows from straightforward applications of Lemma A.6, Proposition and Proposition A.7:

$$\begin{aligned}\|L(\phi(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2^2 &= \|\nabla_\lambda g_u(\lambda_p^\star) - \nabla_\lambda g_u(\lambda_u^\star)\|_2^2 \\ &\leq \beta_g^2 \|\lambda_p^\star - \lambda_u^\star\|_2^2 \\ &\leq 2\frac{\beta_g^2}{\mu_g} M\nu(1 + \|\lambda_p^\star\|_1)\end{aligned}$$

A.6 PROOF LEMMA 3.8

A.6.1 STRONG CONCAVITY CONSTANT μ_g

As argued in Lemma A.6, the unparametrized Lagrangian has a unique minimizer $\phi(\lambda)$ for each $\lambda \in \mathbb{R}_+^m$. Let $\lambda_1, \lambda_2 \in \mathcal{B}_\lambda$ and $\phi_1 = \phi(\lambda_1), \phi_2 = \phi(\lambda_2)$.

By convexity of the functions $\ell_i : \mathcal{F} \to \mathbb{R}$ for $i = 1, \ldots, m$, we have:

$$\begin{aligned}\ell_i(\phi_2) &\geq \ell_i(\phi_1) + \langle D_\phi \ell_i(\phi_1), \phi_2 - \phi_1 \rangle, \\ \ell_i(\phi_1) &\geq \ell_i(\phi_2) + \langle D_\phi \ell_i(\phi_2), \phi_1 - \phi_2 \rangle\end{aligned}$$

Multiplying the above inequalities by $[\lambda_1]_i \geq 0$ and $[\lambda_2]_i \geq 0$ respectively and adding them, we obtain:

$$-\langle \ell(\phi_2) - \ell(\phi_1), \lambda_2 - \lambda_1 \rangle \geq \langle \lambda_1^T D_\phi \ell(\phi_1) - \lambda_2^T D_\phi \ell(\phi_2), \phi_2 - \phi_1 \rangle \tag{12}$$

Since $\nabla g_u(\lambda) = L(\phi(\lambda))$, we have that:

$$-\langle \nabla g_u(\lambda_2) - \nabla g_u(\lambda_2), \lambda_2 - \lambda_1 \rangle \geq \langle \lambda_1^T D_\phi L(\phi_1) - \lambda_2^T D_\phi L(\phi_2), \phi_2 - \phi_1 \rangle \tag{13}$$

Moreover, first order optimality conditions yield:

$$\begin{aligned}D_\phi \ell_0(\phi_1) + \lambda_1^T D_\phi \ell(\phi_1) = 0, \\ D_\phi \ell_0(\phi_2) + \lambda_2^T D_\phi \ell(\phi_2) = 0\end{aligned} \tag{14}$$

where 0 denotes the null-opereator from $\mathcal{F}$ to $\mathbb{R}$ (see e.g: (Kurdila & Zabarankin, 2006) Theorem 5.3.1).

Combining equations 13 and 14 we obtain:

$$\begin{aligned}-\langle \nabla g_u(\lambda_2) - \nabla g_u(\lambda_2), \lambda_2 - \lambda_1 \rangle &\geq \langle D_\phi \ell_0(\phi_2) - D_\phi \ell_0(\phi_1), \phi_2 - \phi_1 \rangle \\ &\geq \mu_0 \|\phi_2 - \phi_1\|_{L_2}^2\end{aligned} \tag{15}$$

where we used the μ_0-strong convexity of the operator ℓ_0.

We will now obtain a lower bound on $\|\phi_2 - \phi_1\|_{L_2}$, starting from the $\beta-$smoothness of ℓ_0:

$$\begin{aligned}\|\phi_2 - \phi_1\|_2 &\geq \frac{1}{\beta} \|D_\phi \ell_0(\phi_2) - D_\phi \ell_0(\phi_1)\|_{L_2} \\ &= \frac{1}{\beta} \|\lambda_2^T D_\phi \ell(\phi_2) - \lambda_1^T D_\phi \ell(\phi_1)\|_{L_2} \\ &= \frac{1}{\beta} \|(\lambda_2 - \lambda_1)^T D_\phi \ell(\phi_2) - \lambda_1^T (D_\phi \ell(\phi_1) - D_\phi \ell(\phi_2))\|_{L_2}\end{aligned} \tag{16}$$

Then, second term in the previous equality can be characterized using assumption 3.7:

$$\|(\lambda_2 - \lambda_1)^T D_\phi \ell(\phi_2)\|_{L_2} \geq \sigma \|\lambda_2 - \lambda_1\|_2 \tag{17}$$

For the second term, using the $\beta-$smoothness of ℓ_i we can derive:

$$\begin{aligned}\|\lambda_1^T(D_\phi\ell(\phi_1) - D_\phi\ell(\phi_2))\|_{L_2} &= \|\sum_{i=1}^m [\lambda_1]_i(D_\phi\ell_i(\phi_1) - D_\phi\ell_i(\phi_2))\|_{L_2} \\ &\leq \sum_{i=1}^m [\lambda_1]_i\|D_\phi\ell_i(\phi_1) - D_\phi\ell_i(\phi_2)\|_{L_2} \\ &\leq \sum_{i=1}^m [\lambda_1]_i\beta\|\phi_1 - \phi_2\|_{L_2} \\ &= \beta\|\lambda_1\|_1\|\phi_1 - \phi_2\|_{L_2}\end{aligned} \tag{18}$$

Then, using the reverse triangle inequality:

$$\begin{aligned}\|(\lambda_2 - \lambda_1)^T D_\phi\ell(\phi_2) &- \lambda_1^T(D_\phi\ell(\phi_1) - D_\phi\ell(\phi_2))\|_{L_2} \\ &\geq \|(\lambda_2 - \lambda_1)^T D_\phi\ell(\phi_2)\|_{L_2} - \|\lambda_1^T(D_\phi\ell(\phi_1) - D_\phi\ell(\phi_2))\|_{L_2} \\ &\geq \sigma\|\lambda_2 - \lambda_1\|_2 - \beta\|\lambda_1\|_1\|\phi_2 - \phi_1\|_{L_2}\end{aligned} \tag{19}$$

Combining this with equation 16 we obtain:

$$\begin{aligned}\|\phi_2 - \phi_1\|_2 &\geq \frac{1}{\beta}\left(\sigma\|\lambda_2 - \lambda_1\|_2 - \beta\|\lambda_1\|_1\|\phi_2 - \phi_1\|_{L_2}\right) \\ \longrightarrow \|\phi_2 - \phi_1\|_{L_2} &\geq \frac{\sigma}{\beta(1 + \|\lambda_1\|_1)}\|\lambda_2 - \lambda_1\|_2\end{aligned} \tag{20}$$

This means that we can write equation 15 as:

$$-\langle\nabla g_u(\lambda_2) - \nabla g_u(\lambda_1), \lambda_2 - \lambda_1\rangle \geq \frac{\mu_0\,\sigma^2}{\beta^2(1 + \|\lambda_1\|_1)^2}\|\lambda_2 - \lambda_1\|_2^2$$

Letting $\lambda_2 = \lambda_u^\star$, we obtain that the strong concavity constant of g_u in $\mathcal{B}_\lambda$ is $\mu_g = \frac{\mu_0\,\sigma^2}{\beta^2(1+\max\{\|\lambda_u^\star\|_1, \|\lambda_p^\star\|_1\})^2}$. A similar proof in the finite dimensional case can be found in (Guigues, 2020).

A.6.2 SMOOTHNESS CONSTANT β_g

Set $\lambda_1, \lambda_2 \in \mathbb{R}_+^m$, and let $\phi_1 = \phi(\lambda_1)$ and $\phi_2 = \phi(\lambda_2)$ denote the Lagrangian minimizers associated to these multipliers.

Since the unparametrized Lagrangian is differentiable and μ_0-strongly convex we have:

$$\mathcal{L}(f, \lambda) \geq \mathcal{L}(\phi(\lambda), \lambda) + \langle D_\phi\mathcal{L}(\phi(\lambda), \lambda)), f - \phi(\lambda)\rangle + \frac{\mu_0}{2}\|f - \phi(\lambda)\|_{L_2}^2$$

Using that $\phi(\lambda)$ is a minimizer, we obtain (see e.g: (Kurdila & Zabarankin, 2006) Theorem 5.3.1) :

$$\mathcal{L}(\phi(\lambda), \lambda) \leq \mathcal{L}(f, \lambda) - \frac{\mu_0}{2}\|f - \phi(\lambda)\|_2^2, \forall f \in \mathcal{F}$$

Applying this to ϕ_2 and ϕ_1 we obtain:

$$\begin{aligned}\ell_0(\phi_2) + \lambda_2^T\ell(\phi_2) &\leq \ell_0(\phi_1) + \lambda_2^T\ell(\phi_1) - \frac{\mu_0}{2}\|\phi_2 - \phi_1\|_{L_2}^2 \\ \ell_0(\phi_1) + \lambda_1^T\ell(\phi_1) &\leq \ell_0(\phi_2) + \lambda_1^T\ell(\phi_2) - \frac{\mu_0}{2}\|\phi_2 - \phi_1\|_{L_2}^2\end{aligned}$$

Summing the above inequalities and applying Cauchy-Schwarz:

$$\begin{aligned}\mu_0\|\phi_2 - \phi_1\|_2^2 &\leq (\lambda_2 - \lambda_1)^T(\ell(\phi_1) - \ell(\phi_2)) \\ &\leq \|\lambda_2 - \lambda_1\|_2\|\ell(\phi_1) - \ell(\phi_2)\|_2 \\ &\leq \sqrt{m}M\|\lambda_2 - \lambda_1\|_2\|\phi_1 - \phi_2\|_{L_2}\end{aligned}$$

where the last inequality follows from assumption 3.1. Then, applying Lemma A.6 we obtain:

$$\begin{aligned}\|\nabla_\lambda g_u(\lambda_2) - \nabla_\lambda g_u(\lambda_1)\|_2 &= \|\ell(\phi_2) - \ell(\phi_1)\|_2 \\ &\leq M\|\phi_2 - \phi_1\|_{L_2} \\ &\leq \sqrt{m}\frac{M^2}{\mu_0}\|\lambda_2 - \lambda_1\|_2\end{aligned}$$

which means that g_u has a smoothness constant $\beta_g = \sqrt{m}\frac{M^2}{\mu_0}$.

A.7 Proof Lemma A.8

Lemma A.8 *Let $P^\dagger$ denote the Fenchel conjugate of the perturbation function $P^*(\epsilon)$. For every $\lambda \in \mathbb{R}^m_+$ we have that $P^\dagger(\lambda) = -g_u(\lambda)$.*

By definition of Fenchel conjugate:

$$P^\dagger(\lambda) = \sup_\epsilon \lambda^T\epsilon - P^\star(\epsilon) \tag{21}$$

Using the definition of $P^*(\epsilon)$:

$$\begin{aligned}P^\dagger(\lambda) = \sup_{\phi\in\mathcal{F},\epsilon} \lambda^T\epsilon - \ell_0(\phi) \\ \text{s.t: } \ell(\phi) + \epsilon \preceq 0\end{aligned} \tag{22}$$

Applying the change of variable $z = \ell(\phi) + \epsilon$, $P^\dagger(\lambda)$ can be written as:

$$\begin{aligned}P^\dagger(\lambda) = \sup_{\phi\in\mathcal{F},\mathbf{z}} \lambda^T\mathbf{z} - \lambda^T\ell(\phi) - \ell_0(\phi) \\ \text{s. to: } \mathbf{z} \preceq 0\end{aligned} \tag{23}$$

Since $\mathbf{z} \preceq 0$, the term $\lambda^T\mathbf{z}$ is unbounded above for $\lambda \prec 0$. Thus, we restrict the domain of $P^\dagger(\lambda)$ to $\lambda \succeq 0$. In this region, maximizing over $\mathbf{z} \in \mathbb{R}^m_-$ yields $\mathbf{z}^* = 0$. We can thus write $P^\dagger(\lambda)$ as:

$$\begin{aligned}P^\dagger(\lambda) &= \sup_{\phi\in\mathcal{F}} -\lambda^T\ell(\phi) - \ell_0(\phi), \quad \lambda \succeq 0 \\ &= -\inf_{\phi\in\mathcal{F}} \lambda^T\ell(\phi) + \ell_0(\phi), \quad \lambda \succeq 0\end{aligned} \tag{24}$$

Therefore,

$$P^\dagger(\lambda) = -g_u(\lambda), \quad \lambda \succeq 0.$$

See for instance: (Rockafellar, 1997), Section 28, (Guigues, 2020), Lemma 2.9 or (Rockafellar, 1974), Theorem 7.

A.8 Proof Corollary A.9

Corollary A.9 *Let $\mathcal{B}_\epsilon = \{\gamma\epsilon_\mathbf{u} + (1-\gamma)\epsilon_\mathbf{p} : \gamma \in [0,1]\}$ denote the segment connecting $\epsilon_\mathbf{u}$ and $\epsilon_\mathbf{p}$. The perturbation function $P^*(\epsilon)$ is $\mu_\epsilon-$strongly convex on $\mathcal{B}_\epsilon$ with constant: $\mu_\epsilon = 1/\beta_g$.*

Lemma A.10 *Let h be a closed convex function defined on a subset of the vector space $\mathcal{X}$; h is $\mu-$strongly convex if and only if $h^\dagger$ has $\mu-$Lipschitz continuous gradients. (See e.g, (**?**) or (Goebel & Rockafellar, 2008)).*

In order to apply Lemma A.10 we need to show that the perturbation function $P(\epsilon)$ is convex and closed in the region of interest.

Convexity of $P^\star(\epsilon)$ for convex functional programs is shown in (Bonnans & Shapiro, 1998) or (Rockafellar, 1997) Theorem 29.1. Now we will show that $P^\star(\epsilon)$ is proper and lower semi continuous in the region of interest, which implies that it is closed.

The functional ℓ_0, defined on the compact set $\mathcal{F}$, is smooth. Thus, it is bounded on $\mathcal{F}$. From assumption 3.2 we have that the problem is feasible for $\epsilon = 0$. Therefore, $P(0) < +\infty$. Moreover, by boundedness of ℓ_0, $P(\epsilon) > -\infty \quad \forall \epsilon$, implying that $P(\epsilon)$ is proper.

Now, fix $\epsilon^0 \in \mathcal{B}_\epsilon$. Assumption 3.2 implies that the perturbed problem with constraint: $L(f)+\epsilon^0 \preceq 0$ is strictly feasible. Since this perturbed problem is convex and strictly feasible, its perturbation function $\tilde{P}(\epsilon)$ is lower semi continuous at 0 (see (Bonnans & Shapiro, 1998) Theorem 4.2). Note that $\tilde{P}(\epsilon) = P^*(\epsilon + \epsilon^0)$. Thus, $P^\star(\epsilon)$ is lower semi continuous at ϵ^0.

We conclude that $P^\star(\epsilon)$ is proper and lower semi continuous for all $\epsilon \in \mathcal{B}_\epsilon$.

On the other hand, from Corollary 3.8 $P^\dagger(\lambda) = -g_u(\lambda)$ is β_g-smooth on $\mathbb{R}^m_+$. Thus, we are in the hypothesis of proposition A.8, which implies that $P^\star(\epsilon)$ is strongly convex on $\mathcal{B}_\epsilon$ with constant $\frac{1}{\beta_g}$.

A.9 Proof Proposition A.11

Proposition A.11 *Under assumptions 3.1 and 3.2, $\lambda_p^\star$ is a subgradient of the perturbation function at ϵ_u. That is, $\lambda_p^\star \in \partial P^*(\epsilon_u)$.*

The conjugate nature of the dual function $g_u(\lambda)$ and the perturbation function $P^\star(\epsilon)$ also establishes a dependence between their first order variations. This dependence is captured in the following lemma.

Lemma A.12 *If h is a closed convex function, the subdifferential $\partial h^\dagger$ is the inverse of ∂h in the sense of multivalued mappings (see (Rockafellar, 1997) Corollary 23.5.1):*

$$x \in \partial h^\dagger(y) \quad \iff \quad y \in \partial h(x)$$

On one hand, from Lemma A.6, we have that $\nabla_\lambda g_u(\lambda_p^\star) = \ell(\phi(\lambda_p^\star)) = -\epsilon_u$. On the other hand, from Lemma A.8, $P^\dagger(\lambda) = -g_u(\lambda)$ for all $\lambda \in \mathbb{R}^m_+$.

Taking the gradient with respect to λ and evaluating at $\lambda_p^\star$ we obtain: $\nabla_\lambda P^\dagger(\lambda_p^\star) = -\epsilon_u$. Then, Lemma A.12, yields the wanted sensitivity result:

$$\lambda_p^\star \in \partial P^*(\epsilon_u).$$

A.10 Proof Proposition A.13

Proposition A.13 *Under assumptions 3.3 and 3.1, the difference between the optimal values of problems perturbed by ϵ_p and ϵ_u is bounded:*

$$P^*(\epsilon_p) - P^*(\epsilon_u) \leq M\nu(1 + \|\lambda_p^\star\|) + \lambda_p^{*T}(\epsilon_p - \epsilon_u)$$

Recall that $\epsilon_u = -\ell(\phi(\lambda_p^\star))$ and $\epsilon_p = -\ell(f_\theta(\lambda_p^\star))$. We want to show that:

$$P^*(\epsilon_p) - P^*(\epsilon_u) \leq L\nu(1 + \|\lambda_p^\star\|) + \lambda_p^{*T}(\epsilon_p - \epsilon_u)$$

We start by showing that $P^*(\epsilon_p) \leq \ell_0(f_\theta(\lambda_p^\star))$. Note that $f_\theta(\lambda_p^\star)$ is feasible in the perturbed problem, since its constraint value is $-\epsilon_p$. Then,

$$P(\epsilon_p) = \min_f \{L_0(f) \ : \ L(f) + \epsilon_p \preceq 0\} \leq L_0(f_\theta(\lambda_p^\star))$$

Therefore,

$$P^*(\epsilon_p) - P^*(\epsilon_u) \leq \ell_0(f_\theta(\lambda_p^\star)) - P^*(\epsilon_u). \tag{25}$$

Note that the dual function of the problem perturbed by ϵ_u is $\tilde{g}_u(\lambda, \epsilon_u) := \min_{\phi \in \mathcal{F}} \{\ell_0(f) + \lambda^T(\ell(\phi) + \epsilon_u)\}$. Then, weak duality implies that $P^*(\epsilon_u) \geq \tilde{g}_u(\lambda, \epsilon_u)$ for all λ. Evaluating at $\lambda_p^\star$ we obtain:

$$\begin{aligned} P^*(\epsilon_u) &\geq \min_{\phi \in \mathcal{F}} \{L_0(f) + \lambda_p^{*T}(L(f) + \epsilon_u)\} \\ &= \min_{\phi \in \mathcal{F}} \{\ell_0(\phi) + \lambda_p^{*T}\ell(\phi)\} + \lambda_p^{*T}\epsilon_u \\ &= g_u(\lambda_p^\star) + \lambda_p^{*T}\epsilon_u \end{aligned} \tag{26}$$

Combining equations 26 and 25 we obtain:

$$\begin{aligned} P^*(\epsilon_p) - P^*(\epsilon_u) &\leq \ell_0(f_\theta(\lambda_p^\star)) - g_u(\lambda_p^\star) - \lambda_p^{*T}\epsilon_u \\ &= \ell_0(f_\theta(\lambda_p^\star)) \pm \lambda_p^{*T}\epsilon_p - g_u(\lambda_p^\star) - \lambda_p^{*T}\epsilon_u \end{aligned}$$

Using that $\epsilon_p = -\ell(f_\theta(\lambda_p^\star))$ we obtain:

$$P^*(\epsilon_p) - P^*(\epsilon_u) \leq g_p(\lambda_p^\star) - g_u(\lambda_p^\star) + \lambda_p^{*T}(\epsilon_p - \epsilon_u)$$

Finally, using Propostion we obtain:

$$P^*(\epsilon_p) - P^*(\epsilon_u) \leq M\nu(1 + \|\lambda_p^\star\|_1) + \lambda_p^{*T}(\epsilon_p - \epsilon_u),$$

which conludes the proof.

A.11 Proof Theorem 3.11

Let $\Delta\epsilon = \epsilon_p - \epsilon_u$, using the strong convexity constant obtained in Proposition A.9 we have that:

$$P^*(\epsilon_p) \geq P^*(\epsilon_u) + s^T\Delta\epsilon + \frac{1}{2\beta_g}\|\Delta\epsilon\|_2^2$$

where $s \in \partial P^*(\epsilon_u)$ is a subgradient of $P^*(\epsilon)$ at ϵ_u.

From Proposition A.11 we know that: $\lambda_p^\star \in \partial P^*(\epsilon_u)$. Thus,

$$P^*(\epsilon_p) \geq P^*(\epsilon_u) + \lambda_p^{*T}\Delta\epsilon + \frac{1}{2\beta_g}\|\Delta\epsilon\|_2^2$$

Using the bound on $P^*(\epsilon_p) - P^*(\epsilon_u)$ obtained in proposition A.13 we can write:

$$\begin{aligned} &M\nu(1 + \|\lambda_p^\star\|_1) + \lambda_p^{\star^T}\Delta\epsilon \geq \ \lambda_p^{\star^T}\Delta\epsilon + \frac{1}{2\beta_g}\|\Delta\epsilon\|_2^2 \\ \longrightarrow &M\nu(1 + \|\lambda_p^\star\|_1) \geq \frac{1}{2\beta_g}\|\Delta\epsilon\|_2^2 \end{aligned}$$

This implies:

$$\begin{aligned} &\|\Delta\epsilon\|_2^2 \leq 2\beta_g M\nu(1 + \|\lambda_p^\star\|_1) \\ \longrightarrow &\|\ell(\phi) - \ell(f_\theta(\lambda_p^\star))\|_2^2 \leq 2\beta_g M\nu(1 + \|\lambda_p^\star\|_1) \end{aligned}$$

which concludes the proof.

A.12 Proof Proposition 3.5

From Corollary 3.10 and Proposition 3.11 we have that:

$$\|\ell(\phi(\lambda_p^\star)) - \ell(\phi^\star)\|_2 \leq \sqrt{2\frac{\beta_g^2}{\mu_g}M\nu(1 + \|\lambda_p^\star\|_1)} \tag{27}$$

$$\|\ell(\phi(\lambda_p^\star)) - \ell(f_\theta(\lambda_p^\star))\|_2 \leq \sqrt{2\beta_g M\nu(1 + \|\lambda_p^\star\|_1)} \tag{28}$$

Combining the two equations above we obtain:

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2 \leq \sqrt{2\frac{\beta_g^2}{\mu_g}M\nu(1 + \|\lambda_p^\star\|_1)} + \sqrt{2\beta_g M\nu(1 + \|\lambda_p^\star\|_1)} \tag{29}$$

$$\leq \sqrt{2\beta_g M\nu(1 + \|\lambda_p^\star\|_1)}\left(1 + \sqrt{\frac{\beta_g}{\mu_g}}\right) \tag{30}$$

Taking squares on both sides yields the desired result.

A.13 PROOF COROLLARY 3.9

Proposition 3.8 characterizes the strong concavity μ_g and smoothness β_g of the dual function in terms of the properties of the losses ℓ_i and the functional space $\mathcal{F}$. The proof of this corollary stems from applying proposition 3.8 to the 2-norm bound in Theorem 3.5.

We start by observing that:

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_\infty \leq \|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_2 \tag{31}$$

$$\leq \sqrt{2\beta_g M\nu(1+\|\lambda_p^\star\|_1)}(1+\sqrt{\frac{\beta_g}{\mu_g}}) \tag{32}$$

From proposition 3.8, we have that $\mu_g = \frac{\mu_0\,\sigma^2}{\beta^2(1+\Delta)^2}$ and $\beta_g = \frac{\sqrt{m}M^2}{\mu_0}$. This implies that

$$\frac{\beta_g}{\mu_g} = \sqrt{m}\frac{M^2}{\sigma^2}\frac{\beta^2}{\mu_0^2}(1+\Delta)^2$$

where $\Delta = \max\{\|\lambda_u^\star\|_1, \|\lambda_p^\star\|_1\}$. Plugging this into equation 32, we obtain:

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_\infty \leq Mm^{1/4}\sqrt{2\frac{M\nu}{\mu_0}(1+\|\lambda_p^\star\|_1)}\left[1+m^{1/4}\frac{M}{\sigma}\frac{\beta}{\mu_0}(1+\Delta)\right]$$

$$\leq M\sqrt{2\frac{M\nu}{\mu_0}(1+\|\lambda_p^\star\|_1)}\left[1+\frac{M}{\sigma}\frac{\beta}{\mu_0}(1+\Delta)\right]\sqrt{m}$$

Finally, using the definitions of the condition numbers $\kappa_1 = \frac{M}{\sigma}$, $\kappa_2 = \frac{\beta}{\mu_0}$ we obtain:

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi(\lambda_u^\star))\|_\infty \leq M\ [1+\kappa_1\kappa_2(1+\Delta)]\ \sqrt{2m\frac{M\nu}{\mu_0}(1+\|\lambda_p^\star\|_1)} \tag{33}$$

which conludes the proof.

A.14 PROOF PROPOSITION 4.1

A similar proof in the context of resource allocation for wireless communications can be found in (Ribeiro, 2010), Theorem 2. To ease the notation, we will denote the value of the parametrized dual function at iteration t by $g(t) := g_p(\lambda(t))$. Similarly, $g^{\text{best}}(t)$ will denote the largest value of $g(t)$ encountered so far. As described in section 4, expected values are taken with respect to $(x,y) \sim \mathfrak{D}_i$.

We start by deriving a recursive inequality between the distances of iterates $\lambda(t)$ and an optimal dual variable $\lambda_p^\star \in \arg\max_{\lambda\succeq 0} g_p(\lambda)$.

Proposition A.14 *Consider the dual ascent algorithm described in Section 4 using a constant step size $\eta > 0$. Then,*

$$\mathbb{E}\{\|\lambda(t+1) - \lambda_p^\star\|^2 | \lambda(t)\} \leq \|\lambda(t) - \lambda_p^\star\|^2 + \eta^2 S^2 - 2\eta(D_p^\star - g(t)) \tag{34}$$

We can observe that as the optimality gap $D_p^\star - g(t)$ decreases, the term $2\eta(D_p^\star - g(t))$ eventually becomes smaller than the fixed term $\eta^2 S^2$, suggesting convergence of $\lambda(t)$ only to a neighborhood of $\lambda_p^\star$. In order to show this, the main obstacle is that Proposition A.14 bounds the expected value of $\|\lambda(t+1) - \lambda_p^\star\|^2$ and we wish to establish almost sure convergence. This can be addressed by leveraging the Supermartingale Convergence Theorem (see e.g, (Solo & Kong, 1994) Theorem E7.4), which we state here for completeness.

Theorem A.15 *Consider nonnegative stochastic processes $A(\mathbb{N})$ and $B(\mathbb{N})$ with realizations $\alpha(\mathbb{N})$ and $\beta(\mathbb{N})$ having values $\alpha(t) \geq 0$ and $\beta(t) \geq 0$ and a sequence of nested σ-algebras $\mathcal{A}(0:t)$ measuring at least $\alpha(0:t)$ and $\beta(0:t)$. If*

$$\mathbb{E}[\alpha(t+1) \mid \mathcal{A}(0:t)] \leq \alpha(t) - \beta(t) \tag{35}$$

the sequence $\alpha(t)$ converges almost surely and $\beta(t)$ is almost surely summable, i.e., $\sum_{u=1}^{\infty} \beta(u) < \infty$ a.s.

We define $\alpha(t)$ and $\beta(t)$ as follows,

$$\alpha(t) := \|\lambda(t) - \lambda_p^\star\|^2 \, \mathbb{I}\left\{ D_p^\star - g^{\text{best}}(t) > \frac{\eta S^2}{2} \right\}$$
$$\beta(t) := [2\eta(D_p^\star - g(t)) - \eta^2 S^2] \, \mathbb{I}\left\{ D_p^\star - g^{\text{best}}(t) > \frac{\eta S^2}{2} \right\}$$

Note that $\alpha(t)$ tracks $\|\lambda(t) - \lambda_p^\star\|^2$ until the optimality gap $D_p^\star - g^{\text{best}}(t)$ falls bellow the threshold $\frac{\eta S^2}{2}$ and is then set to 0. Similarly, $\beta(t)$ tracks $2\eta(D_p^\star - g(t)) - \eta^2 S^2$ until the optimality gap $D_p^\star - g^{\text{best}}(t)$ falls bellow the same threshold and is then set to 0.

It is clear that $\alpha(t) \geq 0$, since it is the product of a norm and an indicator function. The same holds for $\beta(t)$, since the indicator evaluates to 0 whenever $2\eta(D_p^\star - g(t)) - \eta^2 S^2 \leq 0$. We thus have, $\alpha(t), \beta(t) \geq 0$ for all t.

In what follows, we will leverage Theorem A.15 to show that $\beta(t)$ is almost surely summable, which will lead to the wanted result.

Let $\mathcal{A}(0:t)$ be a sequence of σ-algebras measuring $\alpha(0:t), \beta(0:t)$ and $\lambda(0:t)$. We will show that $\alpha(t)$ and $\beta(t)$ satisfy the hypothesis of Theorem A.15 with respect to $\mathcal{A}(0:t)$. Note that at each iteration, $\alpha(t)$ and $\beta(t)$ are fully determined by $\lambda(t)$. Therefore, conditioning on $\mathcal{A}(0:t)$ is equivalent to conditioning on $\lambda(t)$, i.e: $\mathbb{E}\{\alpha(t)|\mathcal{A}(0:t)\} = \mathbb{E}\{\alpha(t)|\lambda(t)\}$. Then we can write,

$$\begin{aligned}\mathbb{E}\{\alpha(t)|\mathcal{A}(0:t)\} =& \mathbb{E}\{\alpha(t)|\lambda(t), \alpha(t) = 0\}\mathbb{P}\{\alpha(t) = 0\} \\ &+ \mathbb{E}\{\alpha(t)|\lambda(t), \alpha(t) > 0\}\mathbb{P}\{\alpha(t) > 0\}\end{aligned} \tag{36}$$

From equation 36, we will derive that $\mathbb{E}\{\alpha(t)|\mathcal{A}(0:t)\} \leq \alpha(t) - \beta(t)$ which is the remaining hypothesis in Theorem A.15.

On one hand, observe that if $\alpha(t) = 0$ we have that $\mathbb{I}\{D_p^\star - g^{\text{best}}(t) \leq \frac{\eta S^2}{2}\} = 0$. This is because in the case where $\|\lambda(t) - \lambda_p^\star\|^2 = 0$, the indicator function also evaluates to 0. Therefore, if $\alpha(t) = 0$, it must be that $\beta(t) = 0$. Then, trivially, $\mathbb{E}\{\alpha(t)|\lambda(t), \alpha(t) = 0\} = \alpha(t) - \beta(t)$.

On the other hand, when $\alpha(t) > 0$:

$$\mathbb{E}[\alpha(t+1) \mid \lambda(t), \alpha(t) > 0] \tag{37}$$
$$= \mathbb{E}\left\{ \left\|\lambda(t+1) - \lambda_p^\star\right\|^2 \mathbb{I}\left\{ D_p^\star - g^{\text{best}}(t+1) > \frac{\eta \hat{S}^2}{2} \right\} \mid \lambda(t) \right\} \tag{38}$$
$$= \mathbb{E}\left\{ \left\|\lambda(t+1) - \lambda_p^\star\right\|^2 \mid \lambda(t) \right\} \tag{39}$$

where we used the definition of $\alpha(t+1)$ and the fact the the indicator function needs to evaluate to 1 since $\alpha(t) > 0$. Then, from proposition A.14 we have:

$$\mathbb{E}[\alpha(t+1) \mid \lambda(t), \alpha(t) > 0] \leq \left\|\lambda(t) - \lambda_p^\star\right\|^2 + \eta^2 S^2 - 2\eta(D_p^\star - g(t)) \tag{40}$$
$$= \alpha(t) - \beta(t). \tag{41}$$

where the last equality comes from the fact that $\alpha(t) > 0$ implies $\mathbb{I}\left\{ D_p^\star - g^{\text{best}}(t+1) > \frac{\eta \hat{S}^2}{2} \right\} = 1$.

This means that we can write equation 36 as:

$$\begin{aligned}\mathbb{E}\{\alpha(t)|\mathcal{A}(0:t)\} &\leq [\alpha(t) - \beta(t)](\mathbb{P}\{\alpha(t) = 0\} + \mathbb{P}\{\alpha(t) > 0\}) \\ &= \alpha(t) - \beta(t)\end{aligned} \tag{42}$$

which shows that $\alpha(t)$ and $\beta(t)$ satisfy the hypothesis of Theorem A.15. Then, we have that $\beta(t)$ is almost surely summable, which implies,

$$\liminf_{t\to\infty} \left[2\eta(D_p^\star - g(t)) - \eta^2 S^2\right] \mathbb{I}\{D_p^\star - g^{\text{best}}(t) > \eta\hat{S}^2/2\} = 0 \text{ a.s.}$$

This is true if either $D_p^\star - g^{\text{best}}(t) \leq \frac{\eta S^2}{2}$ for some t, or if $\liminf_{t\to\infty} \left[2\eta(D_p^\star - g(t)) - \eta^2 S^2\right] = 0$, which concludes the proof.

A.14.1 Proof proposition A.14

We want to show that

$$\mathbb{E}\{\|\lambda(t+1) - \lambda_p^\star\|^2 | \lambda(t)\} \leq \|\lambda(t) - \lambda_p^\star\|^2 + \eta^2 S^2 - 2\eta(D_p^\star - g(t)) \tag{43}$$

We start from the definition of $\lambda(t+1)$:

$$\begin{aligned} \|\lambda(t+1) - \lambda_p^\star\|^2 &= \|[\lambda(t) + \eta\hat{s}(t)]_+ - \lambda_p^\star\|^2 \\ &\leq \|\lambda(t) - \lambda_p^\star + \eta\hat{s}(t)\|^2 \\ &= \|\lambda(t) - \lambda_p^\star\|^2 + \eta^2\|\hat{s}(t)\|^2 + 2\eta\hat{s}(t)^T(\lambda(t) - \lambda_p^\star) \end{aligned} \tag{44}$$

where we used the fact that setting the negative components of $\lambda(t)+\eta\hat{s}(t)$ to 0 decreases its distance to the positive vector $\lambda_p^\star$ and then expanded the square.

Note that for a given $\lambda(t)$, the relations in 44 hold for all realizations of $\hat{s}(t)$. Thus, the expectation of $\|\lambda(t+1) - \lambda_p^\star\|$, conditioned on $\lambda(t)$ satisifes:

$$\mathbb{E}\{\|\lambda(t+1) - \lambda_p^\star\|^2 | \lambda(t)\} \leq \|\lambda(t) - \lambda_p^\star\|^2 + \eta^2\mathbb{E}\{\|\hat{s}(t)\|^2 | \lambda(t)\} + 2\eta\mathbb{E}\{\hat{s}(t)|\lambda(t)\}(\lambda(t) - \lambda_p^\star) \tag{45}$$

Finally, recall that $\mathbb{E}\{\hat{s}(t)|\lambda(t)\}$ is a supergradient of the concave dual function $g_p(\lambda)$, that is:

$$\mathbb{E}\{\hat{s}(t)|\lambda(t)\}(\lambda(t) - \lambda) \leq g(t) - g_p(\lambda)\,. \tag{46}$$

Evaluating the previous inequality at $\lambda_p^\star$ and combining it with equation 45 we obtain:

$$\mathbb{E}\{\|\lambda(t+1) - \lambda_p^\star\|^2 | \lambda(t)\} \leq \|\lambda(t) - \lambda_p^\star\|^2 + \eta^2 S^2 + 2\eta(g(t) - D_p^\star) \tag{47}$$

which concludes the proof.

A.15 Proof Proposition 4.2

We will bound the distance between $\ell(\phi(\lambda_u^\star))$ and $L(f_\theta(\lambda^{\text{best}})))$ by partioning it into terms that are similar to those we have previously analyzed in Corollary 3.10 and Proposition 3.11:

$$\begin{aligned} \|\ell(\phi(\lambda_u^\star)) - \ell(f_\theta(\lambda^{\text{best}})))\|_2 \leq\ & \|\ell(\phi(\lambda_u^\star)) - \ell(\phi(\lambda^{\text{best}}))\|_2 \\ &+ \|\ell(\phi(\lambda^{\text{best}})) - \ell(f_\theta(\lambda^{\text{best}}))\|_2 \end{aligned}$$

The first term is of the same nature as the one analyzed in Corollary 3.10, since it is characterizes a perturbation in dual variables in the unparametrized problem. Thus, using the characterization of the curvature of the dual function from proposition A.7 and the sub-optimality of λ^{best} with respect to $\lambda_p^\star$, this term can be bounded.

We will denote by $\mathcal{B}_{\lambda^{\text{best}}}$ the segment connecting λ^{best} and $\lambda_u^\star$ and by $\tilde{\mu}_g$ the strong concavity constant of g_u in $\mathcal{B}_{\lambda^{\text{best}}}$. Proceeding exactly as in the proof of Propositon A.7 we obtain:

$$\begin{aligned} \|\lambda^{\text{best}} - \lambda_u^\star\|_2^2 &\leq \frac{2}{\tilde{\mu}_g}(g_u(\lambda_u^\star) - g_u(\lambda^{\text{best}})) \\ &\leq \frac{2}{\tilde{\mu}_g}(g_p(\lambda_p^\star) - \big(g_p(\lambda^{\text{best}}) - M\nu(1 + \|\lambda^{\text{best}}\|_1)\big) \end{aligned}$$

where we used Lemma A.5 and the fact that $g_p(\lambda_p^\star) \geq g_u(\lambda_u^\star)$.

Then, leveraging the almost sure convergence shown in Proposition 4.1 we have:

$$\|\lambda^{\text{best}} - \lambda_u^\star\|_2^2 \leq \frac{2}{\tilde{\mu}_g}\left(M\nu(1 + \|\lambda^{\text{best}}\|_1) + \frac{\eta S^2}{2} + \delta\right) \tag{48}$$

Thus,

$$\|\ell(\phi(\lambda^{\text{best}})) - \ell(\phi(\lambda_u^\star))\|_2^2 = \|\nabla_\lambda g_u(\lambda^{\text{best}}) - \nabla_\lambda g_u(\lambda_u^\star)\|_2^2 \tag{49}$$

$$\leq \beta_g^2 \|\lambda^{\text{best}} - \lambda_u^\star\|_2^2 \tag{50}$$

$$\leq \frac{2\beta_g^2}{\tilde{\mu}_g}\left(M\nu(1 + \|\lambda^{\text{best}}\|_1) + \frac{\eta S^2}{2} + \delta\right) \tag{51}$$

which completes the first part of the proof.

The term $\|\ell(\phi(\lambda^{\text{best}})) - \ell(f_\theta(\lambda^{\text{best}}))\|_2$ captures a perturbation in the function class for a fixed dual variable, and can be analyzed similarly to Proposition 3.11. Let $\tilde{\epsilon}_u = -\ell(\phi(\lambda^{\text{best}}))$ and $\tilde{\epsilon}_p = -\ell(f_\theta(\lambda^{\text{best}}))$. Using the same arguments as in Lemma A.11, Proposition A.13 and Corollary A.9 we have that:

1. $P^*(\epsilon)$ is strongly convex with constant $\frac{1}{\beta_g}$ on $\mathcal{B}_{\lambda^{\text{best}}}$
2. $\lambda^{\text{best}} \in \partial P(\tilde{\epsilon}_u)$
3. $P^*(\tilde{\epsilon}_p) - P^*(\tilde{\epsilon}_u) \leq M\nu(1 + \|\lambda^{\text{best}}\|_1) + {\lambda^{\text{best}}}^T(\tilde{\epsilon}_p - \tilde{\epsilon}_u)$

Let $\Delta\tilde{\epsilon} = \tilde{\epsilon}_p - \tilde{\epsilon}_u$. Combining the aforementioned properties as done in Theorem 3.11 yields:

$$M\nu(1 + \|\lambda^{\text{best}}\|_1) + {\lambda^{\text{best}}}^T\Delta\tilde{\epsilon} \geq {\lambda^{\text{best}}}^T\Delta\tilde{\epsilon} + \frac{1}{2\beta_g}\|\Delta\tilde{\epsilon}\|_2^2 \tag{52}$$

which implies:

$$\|\Delta\tilde{\epsilon}\|_2^2 \leq 2\beta_g M\nu(1 + \|\lambda^{\text{best}}\|_1) \tag{53}$$

Combining the bounds in equations 51 and 53 we obtain:

$$\|\ell(\phi(\lambda_u^\star)) - \ell(f_\theta(\lambda^{\text{best}}))\|_2 \tag{54}$$

$$\leq \sqrt{2\beta_g M\nu(1 + \|\lambda^{\text{best}}\|_1)} + \sqrt{\frac{2\beta_g^2}{\tilde{\mu}_g}\left(M\nu(1 + \|\lambda^{\text{best}}\|_1) + \frac{\eta S^2}{2} + \delta\right)} \tag{55}$$

$$= \sqrt{2\beta_g M\nu(1 + \|\lambda^{\text{best}}\|_1)}\left(1 + (1 + \frac{\eta S^2}{2} + \delta)\sqrt{\frac{\beta_g}{\mu_g}}\right) \tag{56}$$

Taking squares on both sides concludes the proof.