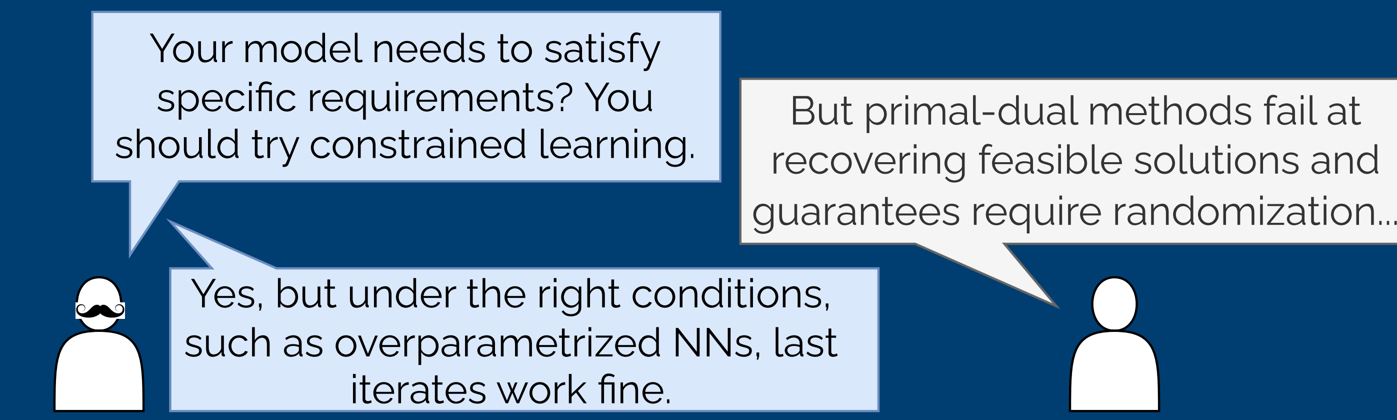
Near Optimal Solutions of Constrained Learning Problems

Juan Elenter [†], Luiz F.O. Chamon [*] and Alejandro Ribeiro [†]

Penn [†], University of Stuttgart [*]



Constrained Learning

Goal: Learn models that excel at their main task and also adhere to requirements.

$$P_p^\star = \min_{\theta\in\Theta} \quad \ell_0(f_\theta) := \mathbb{E}_{(x,y)}[\tilde{\ell}_0(f_\theta(x),y)] \qquad (P_p)$$
$$\text{s. to} \quad \ell_i(f_\theta) := \mathbb{E}_{(x,y)}[\tilde{\ell}_i(f_\theta(x),y)] \le 0, \quad i=1,..,m$$

- Typically the losses $\tilde{\ell}_i(\hat{y},y)$ are *not convex in* θ.
- There is no straightforward way to project onto the feasibility set.

Example: Counterfactual Fairness.

$$\min_{\theta\in\mathbb{R}^p} \mathbb{E}\left[-\log[f_\theta(x)]_y\right]$$
$$\text{s. to } \mathbb{E}\left[D_{\text{KL}}(f_\theta(\tilde{x},z) \,\|\, f_\theta(\tilde{x},\rho_i(z))\right] \le c, \quad i=1,\dots,m,$$

- The output is *near-invariant* to changes in the protected features z.
- I.e, $f_\theta(x_1,x_2,\mathit{Male},\dots,x_d)$ should be c−similar to $f_\theta(x_1,x_2,\mathit{Female},\dots,x_d)$

Learning in the Dual Domain

- Dual problem of (P_p): $D_p^\star = \max_{\lambda_p\ge 0} \underbrace{\min_{\theta\in\Theta} L(f_\theta,\lambda_p)}_{g_p(\lambda_p)} = \ell_0(f_\theta) + \lambda_p^T \ell(f_\theta))$

Dual Constrained Learning

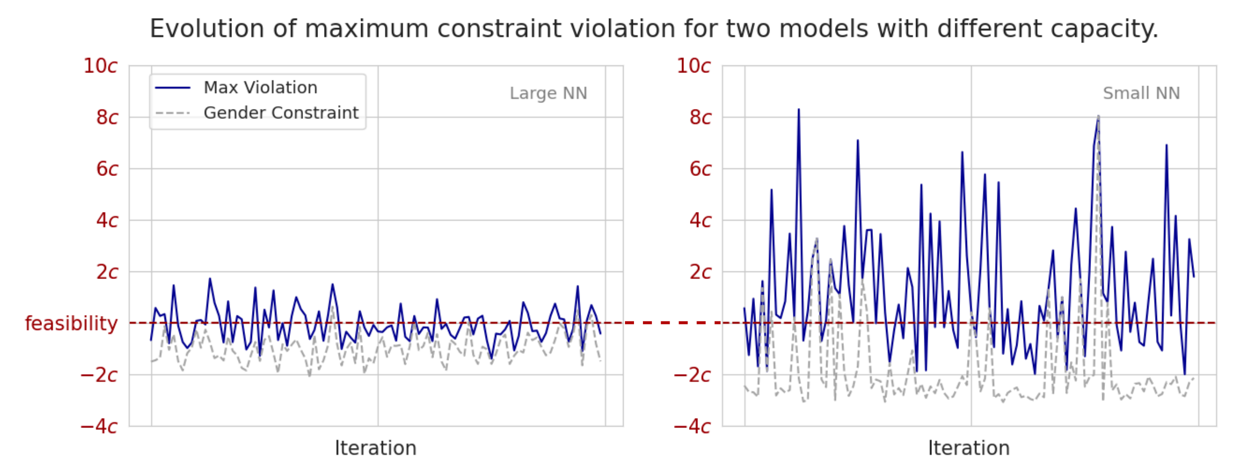
1: *Initialize*: $\lambda(1)$
2: **for** $t=1,\dots,T$ **do**
3: Obtain $f_\theta(t)$ such that
$$f_\theta(t) \in \arg\min_{\theta\in\Theta} L(f_\theta,\lambda(t))$$
4: Update dual variables
$$\lambda_i(t+1) = \left[\lambda_i(t) + \eta\,\ell_i(f_\theta(t))\right]_+$$
5: **end for**

- Involves solving a sequence of regularized ERM problems.
- Dual iterates $\lambda_p(t)$ move in ascent directions of the concave function g_p
- Dual convergence (stochastic): $\lim_{t\to\infty} g_p^{\text{best}}(t) \ge D_p^\star - O(\eta)$ a.s.
- PAC-C Learning Guarantees [Chamon. et al, 2023]

- Even though $D_p^\star$ - $P_p^\star \le \Gamma_1$ is bounded, the sequence of primal iterates $\{f_\theta(t)\}_{t=1}^T$ need not approach the set of solutions of (P_p).

The issue of primal recovery

- At least one constraint is violated on 82% of the iterations.
- We cannot stop the algorithm and expect to obtain a feasible solution.



⇒ Convergence guarantees require **randomization** over $\{f_\theta(t)\}_{t=1}^T$ △
See e.g, [Kearns, 2019; Cotter, 2019; Agarwal, 2018; Chamon, 2023]

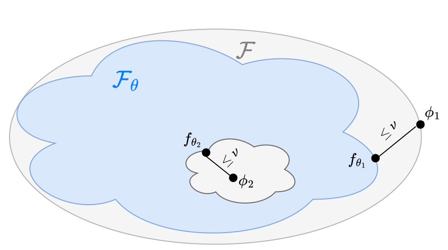
The unparametrized problem

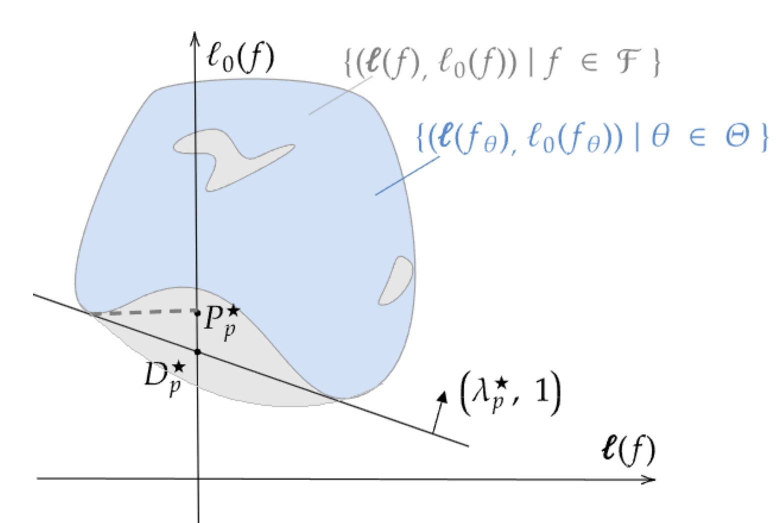
Key Observation : (P_p) is the parametrized version of a benign functional program

$$P_u^\star = \min_{\phi\in\mathcal{F}} \quad \ell_0(\phi) \quad \text{s.to} \;\; \ell_i(\phi) \le 0, \quad i=1,..,m, \qquad (P_u)$$

- $\mathcal{F}$ is a convex, compact subset of an L^2 space.
- (P_u) is convex and has a unique optimal (and feasible) solution $\phi^\star$ (sc, smooth)
- **Richness of $\mathcal{F}_\theta$:** For all $\phi\in\mathcal{F}$, there exists $\theta\in\Theta$ such that: $\|\phi - f_\theta\|_{L_2} \le \nu$.

E.g: $\mathcal{F}$ = { continuous functions on compact set } and
$\mathcal{F}_\theta$ = { 2-layer NN with $K(\nu)$ hidden neurons }





Near Optimal Solutions of Learning Problems

If the problem (P_u) is sufficiently benign, for any $f_\theta(\lambda_p^\star) \in \mathcal{F}(\lambda_p^\star)$ we have:

Near-Feasibility:

$$\|\ell(f_\theta(\lambda_p^\star)) - \ell(\phi^\star)\|_\infty \le M\,\left[1 + \kappa_1\kappa_0(1+\|\lambda_p^\star\|_1)\right]\sqrt{2m\frac{M\nu}{\mu_0}(1+\|\lambda_p^\star\|_1)} := \Gamma_2$$

Near-Optimality:

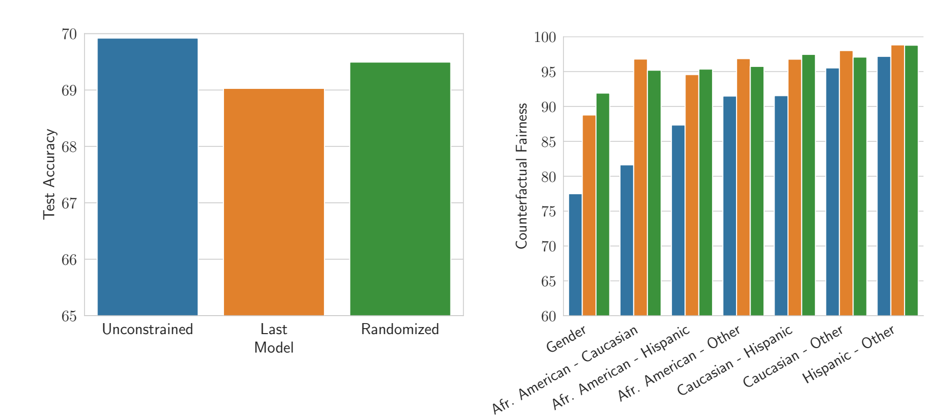
$$|P_p^\star - \ell_0(f_\theta(\lambda_p^\star))| \le (1+\|\lambda_p^\star\|_1)\, M\nu + \Gamma_1 + \|\lambda_p^\star\|_1\, \Gamma_2$$

Sensitivity of $P_p^\star$ - Expressivity of $\mathcal{F}_\theta$ - Condition nums: ℓ_0 (κ_0) and ℓ (κ_1)

- Primal iterates $f_\theta(\lambda_p^\star)$ associated to dual solutions $\lambda_p^\star$ approximate $\phi^\star$ in terms of objective and constraint values.
- $D_p^\star$ approximates $P_p^\star$ **and** provides approximate solutions for (P_p).
- Leads to a convergence guarantee for *best* primal iterates.

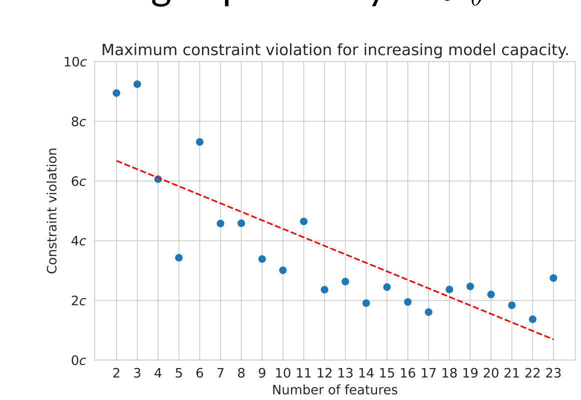
Empirical Performance of Last Iterates

Objective & Constraint Satisfaction



- Last and randomized predictors provide similar accuracy and constraint satisfaction.

Increasing expressivity of $\mathcal{F}_\theta$



- Maximum constraint violation decreases by up an order of magnitude as the richness of the parametrization increases.

elenter@seas.upenn.edu

University of Stuttgart Penn