We thank the reviewers for their time and for consideration. They will find responses to their specific points below.

Reviewer 2. While we use fairness and robustness to illustrate the use of constraints in learning, the main results of the paper (Thm. 1–3) are independent of these applications or the experimental settings. Still, we believe they are relevant. Models invariant to gender, e.g., can be used for fair decision-making and, in a more data analysis context, to seek patterns masked by gender or identify, through the dual variables, the features (phenotypes) more sensitive to gender. We illustrate such analyses in Section D of the appendices. As the reviewer notes, this framework has applications beyond model structure and preliminary results suggest that pointwise, per data point constraints can be used to improve fit. However, this is beyond the scope of this work that focuses on understanding what can be achieved when learning under constraints. Due to space limitations we could not address these points in the main body of the paper, but we intend to use the extra page in the final manuscript to bring experimental details from the appendices. If this does not fully address the reviewer concerns about the broader impact of the manuscript, we can work on expanding that section for the camera-ready.

Reviewer 3. We agree that, being a new area of research, constrained learning theory still has many open questions, such as understanding the effect of constraints on other learning models (e.g., structured complexity and PAC-Bayes) and learning forms (e.g., reinforcement learning). We will include these discussions in the camera-ready. Still, this work already shows that constrained learning is fundamentally different than PAC learning, especially dual constrained learning where the dependence on $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ appears. Nevertheless, they affect only the *parametrization error* ϵ_0, which is independent of the sample size (Def. 3). Hence, ϵ_0 is an intrinsic limitation of the parametrized learning problem and sample sizes can always be estimated relative to this lower error bound. Alternatively, there is a well-known upper bound from optimization theory, which for the case of $[0, B]$-bounded objective yields $\|\boldsymbol{\mu}^\star\|_1 \leq Bs^{-1}$, where $s = \min_i \mathbb{E}\big[\ell_0\big(f_{\boldsymbol{\theta}'}(\boldsymbol{x}), y\big)\big] - c_i$ is the smallest slack for any strictly feasible solution $\boldsymbol{\theta}'$ (i.e., such that $s > 0$). If the output $\boldsymbol{\theta}^{(T)}$ of Alg. 1 is sufficiently in the interior of the feasible set, bounds for the proof of Thm. 2 yield $s \leq \min_i \left[c_i - M\nu - B\zeta(N_i) - \frac{1}{N_i}\sum_{n_i=1}^{N_i} \ell_i\big(f_{\boldsymbol{\theta}^{(T)}}(\boldsymbol{x}_{n_i}), y_{n_i}\big)\right]_+$. While more tractable, we do not think that this result is convenient as it requires the constraints to be quite loose for the bound not to vanish. We therefore did not include these details in the manuscript, but will discuss this limitation more clearly in the camera-ready. Also, we use CSL to mean "Constrained Statistical Learning" and write $\|\cdot\|_1$ to denote the ℓ_1-norm (i.e., on a space of sequences) and $\|\cdot\|_{L_1}$ to denote the L_1-norm (i.e., on a space of functions). This distinction occurs because the dual variable of the j-th pointwise constraint is actually a continuous (as opposed to discrete) function.

Reviewer 4. The reviewer raises good points to improve the clarity of the paper. We intend to use the extra page of the camera-ready to address them as well as bring part of the numerical experiments discussion from the appendices. In particular, we will emphasize how the contributions of the paper provide generalization guarantees and therefore ensure that an accurate, robust classifier trained using our constrained method will also perform well in testing (as evidenced by Figures 4 and 5). Addressing the reviewer's questions more specifically, the indices in (3) refer to the constraints in (P-CSL), i.e., (3a) relates to the i-th average constraint of the learning problem and (3b) relates to the j-th pointwise constraint. The reason we write "PACC Learning is as Hard as…" is because Thm. 1 is both necessary and sufficient. That PAC is not harder than PACC is trivial since PACC includes PAC in (2). But Thm. 1 also proves the converse. We then say these forms of learning are equivalent in the sense that a hypothesis class is PACC learnable if and only if it is PAC learnable (Thm. 1). As for the relation between (PIV) and its dual ($\hat{\text{D}}$-CSL), recall that (PIV) is non-convex so that its dual only provides a lower bound on its value (weak duality). However, Thm. 2 shows that ($\hat{\text{D}}$-CSL) provides both lower and upper bounds for the original (P-CSL). In Def. 3, ϵ and ϵ_0 are not related: ϵ is the *estimation error* while ϵ_0 is the *parametrization error*. Indeed, ϵ_0 is independent of the sample size and depends only on the learning problem (M and dual variables) and the richness of the parametrization (ν). Although the parametrization error is treated separately in unconstrained learning, it is not possible to untangle it for (dual) PACC learning (see proof of Thm. 2). In terms of Alg. 1, its input is a sample set and its output is a parameter vector $\boldsymbol{\theta}^{(T)}$ that describes the classifier. Dual variables can also be extracted for analysis (see Section D). In these analyses, we leverage the sensitivity interpretation of the dual variables, which is formalized by the fact that the optimal dual variables are *subgradients* of the empirical Lagrangian at the optimal primal solution. In fact, we show that this is (approximately) true even at the approximate oracle from Assumption 4 (see Lemma 1 in Section C). To be more specific, it means that the larger the dual variable, the more the optimal value would increase if we were to tighten that constraint, indicating that the constraint is "harder" to satisfy. We use a mini-batch SGD algorithm for step 3 of Alg. 1 (details can be found in the appendices). Note that the results of Thm. 3 do not depend on this method and there may be better suited approaches in certain problems, but this was a natural choice for the logistic and CNN classifiers used in the experiments. While we will certainly expand Section 2, we note that, to the best of our knowledge, this is the first paper to explicitly analyze constraints in the context of (PAC) learning theory. Previous work focus on specific constraints, e.g., rates, and provide guarantees for specific algorithms that typically yield randomized solutions. In contrast, this paper provides algorithm-independent results (Thm. 1 and 2) for deterministic learners. It also considers pointwise constraints that are important in domains such as fairness.