An Analysis of Regularized Approaches for Constrained Machine Learning

(Article begins on next page)

19 July 2024

An Analysis of Regularized Approaches for Constrained Machine Learning

Michele Lombardi, Federico Baldo, Andrea Borghesi, and Michela Milano

University of Bologna, Italy
{michele.lombardi2,federico.baldo2}@unibo.it
{andrea.borghesi3,michela.milano}@unibo.it

1 Context

Regularization-based approaches for injecting constraints in Machine Learning (ML) were introduced (see e.g. [4]) to improve a predictive model via expert knowledge. Given the recent interest in ethical and trustworthy AI, however, several works are resorting to these approaches for enforcing desired properties over a ML model (e.g. fairness [1,6]). Regularized approaches for constraint injection solve, in an exact or approximate fashion, a problem in the form:

$$\arg\min_{w \in W}\{L(y) + \lambda^\top C(y)\} \quad \text{with: } y = f(\mathbf{x}; w) \tag{1}$$

where L is a loss function and f is the model to be trained, with parameter vector w from space W. $f(\mathbf{x}; w)$ refers to the model output for the whole training set $\mathbf{x}$. The regularization function C denotes a vector of (non-negative) constraint violation indices for m constraints, while $\lambda \geq 0$ is a vector of weights (or *multipliers*). For instance, in a regression problem we may desire a specific output ordering for two input vectors in the training set. A viable regularizer may be:

$$C(y) \equiv \max(0, y_i - y_j) \tag{2}$$

the term is zero iff the constraint $y_i \leq y_j$ is satisfied. For obtaining balanced predictions in a binary classification problem, we may use instead:

$$C(y) \equiv \left| \sum_{i=1}^{n} y_i - \frac{n}{2} \right| \tag{3}$$

where y_i is the binary output associated to one of the two classes. If n is even, the term is 0 for perfectly balanced classifications.

When regularized methods are used to enforce constraints, *a typical approach consists in adjusting the λ vector until a suitable compromise between accuracy and constraint satisfaction is reached* (e.g. a discrimination index becomes sufficiently low). This approach enables the use of traditional training algorithms, at the cost of having to search over the space of possible multipliers. Though the method is known to work well in many practical cases, the process has been subject to little general analysis. However, some hints of its limitations are shown

---

Algorithm 1 PR4PC(θ)

1: **for** $\lambda \in (\mathbb{R}^+)^m$ **do**
2:     Optimize PR to find w^*
3:     **if** $C(w^*) \leq \theta$ **then**
4:         Store $w^*, L(w^*)$
5: Pick the stored solution with the smallest $L(w^*)$

---

by Cotter et al.[3] that demonstrate how regularized-based approaches risk to be structurally sub-optimal. With this note, we aim to make a preliminary step in this direction, providing a more systematic overview of the strengths and (in particular) potential weaknesses of this class of approaches.

2   Analysis

Regularized approaches for constraint injection are strongly related to duality in optimization. Despite this, we present an analysis based on first principles , as it provides additional insights. It will be convenient to reformulate Equation (1) by embedding the ML model structure in the L and C functions:

$$\mathbf{PR}(\theta): \quad \underset{w \in W}{\arg\min}\{L(w) + \lambda^\top C(w)\} \tag{4}$$

With some abuse of notation $L(w)$ refers to $L(f(\mathbf{x}; w))$, and the same for $C(w)$. This approach enables a uniform treatment of convex and non-convex models or functions. We are interested in the relation between the unconstrained PR formulation and the following constrained training problem:

$$\mathbf{PC}(\lambda): \quad \underset{w \in W}{\arg\min}\{L(w) \mid C(w) \leq \theta\} \tag{5}$$

where θ is a vector of thresholds for the constraint violation indices. In ethical or trustworthy AI applications, PC will be the most natural problem formulation.

We wish to understand *the viability of solving PC indirectly, by adjusting the λ vector and solving the unconstrained problem PR*, as depicted in Algorithm 1: line 2 refers to some kind of search over the multiplier space. Ideally, the algorithm should be equivalent to solving the PC formulation directly. For this to be true, solving PR(λ) should have a chance to yield assignments that are optimal for the constrained problem. Moreover, an optimum of PC(θ) should always be attainable in this fashion. Additional properties may enable more efficient search. In the note, we will characterize Algorithm 1.

Regularized and Constrained Optima The relation between the PR and PC formulations are tied to the properties of their optimal solutions. An optimal PC solution w_c^* satisfies:

$$\mathbf{opt_c}(\mathbf{w}^*, \theta): \quad L(w) \geq L(w^*) \quad \forall w \in W \mid C(w) \leq \theta \tag{6}$$

while for an optimal solution w_r^* of PR with multipliers λ we have:

$$\mathbf{opt_r(w^*, \lambda)}: \quad L(w) + \lambda^\top C(w) \geq L(w^*) + \lambda^\top C(w^*) \quad \forall w \in W \qquad (7)$$

The definitions apply also to local optima, by swapping W with some neighborhood of w_c^* and w_r^*. We can now provide the following result:

Theorem 1. *an optimal solution w^* for PR is also optimal for PC, for a threshold equal to $C(w^*)$:*

$$opt_r(w^*, \lambda) \Rightarrow opt_c(w^*, C(w^*)) \qquad (8)$$

Proof (by contradiction). Let us assume that w^* is an optimal solution for PR but not optimal for PC, i.e. that there is a feasible $w' \in W$ such that:

$$L(w') < L(w^*) \qquad (9)$$

Since w^* is optimal for PR, we have that:

$$L(w') \geq L(w^*) + \lambda^\top (C(w^*) - C(w')) \qquad (10)$$

Since w' is feasible for $\theta = C(w^*)$, we have that its violation vector cannot be greater than that of w^*. Formally, we have that $C(w') \leq C(w^*)$, or equivalently $C(w^*) - C(w') \geq 0$. Therefore Equation (10) contradicts Equation (9), thus proving the original point. The same reasoning applies to local optima. $\square$

Theorem 1 shows that solving PR(λ) *always results in an optimum for the constrained formulation*, albeit for threshold $\theta = C(w^*)$ that cannot be a priori chosen. The statement is true for non-convex loss, reguralizer, and model structure. This is a powerful result, which provides a strong motivation for Alg. 1.

Global vs Local Optimality If regularized problems can be solved to global optimality, then increasing a weight in the λ vector cannot have an adverse effect on the satisfaction level of the corresponding constraint. Formally, there is a monotonic relation between λ and $C(w^*)$:

$$opt_r(w', \lambda'), opt_r(w'', \lambda''), \lambda_j' \geq \lambda_j'' \Rightarrow C_j(w') \leq C_j(w'') \qquad (11)$$

The proof is omitted due to lack of space. When monotonicity holds, searching over the multiplier space in Algorithm 1 can be considerably simpler (e.g. binary search for a single multiplier, or sub-gradient descent in general [5]). However, global optimality is attainable only in very specific cases (e.g. convex loss, regularizer, and model) or by solving PR in an exact fashion (which may be computationally expensive). Failing this, monotonicity will not strictly hold, in the worst case requiring exhaustive (or semi-exhaustive) search on the multiplier space. Additionally, relying on local optima will lead to suboptimal solutions (subject to uncertainty if stochastic training algorithm is employed).
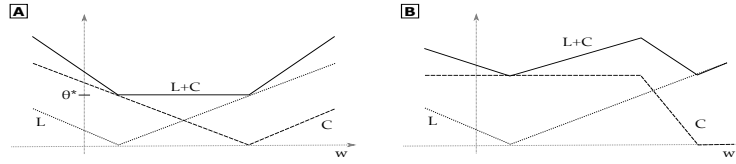
Fig. 1. Multiple Optima in Convex (A) and Non-Convex (B) Regularized Problems

Unique vs Multiple Optima Further issues arise when the regularized problem PR(λ) has multiple equivalent optima. In the fully convex case, this may happen if the multiplier values generate plateaus (see Fig. 1A, where $\lambda = 1$). In the non-convex case, there may be separate optima with the same value for the regularized loss, but different trade-offs between loss and constraint violation: this is depicted in Fig. 1B. Multiple equivalent optima may cause a non-monotonic relation between λ and the constraint satisfaction level. Additionally, it may happen that different constrained optima are associated to the same multiplier, *and to no other multiplier*. In Fig. 1A, for example, the multiplier $\lambda = 1$ is associated to all optimal solutions of PC(θ) with $\theta \leq \theta^*$; no other multiplier is associated to the same solutions. Unless some kind of tie breaking technique is employed, this situation makes specific constrained optima impossible to reach.

Inaccessible Constrained Optima We next proceed to investigate whether an optimum of the constrained formulation may be associated to no multiplier value: any such point would be completely unattainable via Algorithm 1. We have that:

Theorem 2. *An optimal solution w^* for PC is optimal for PR iff there exists a multiplier vector λ that satisfies:*

$$\max_{\substack{w \in W, \\ C_j(w) > C_j(w^*)}} R(w, \lambda) \leq \lambda_j \leq \min_{\substack{w \in W, \\ C_j(w) < C_j(w^*)}} R(w, \lambda) \tag{12}$$

with:

$$R(w, \lambda) = -\frac{\Delta L(w, w^*) + \lambda_{\bar{j}}^{\top} \Delta C_{\bar{j}}(w, w^*)}{\Delta C_j(w, w^*)} \tag{13}$$

In the theorem, we refer with $\Delta C(w, w^*)$ to the difference $C(w) - C(w^*)$ and with $\Delta L(w, w^*)$ to the difference $L(w) - L(w^*)$. Moreover, $\bar{j}$ refers to the set of all multiplier indices, except for j. Intuitively, every assignment for which constraint j has a lower degree of violation than in w^* enforces an upper bound on λ_j; every assignment for which the violation is higher enforces a lower bound.

Proof. Let w^* be a PC optimum for some threshold θ; this implies that w^* is also optimal for a tightened threshold, i.e. for $\theta = C(w^*)$. We therefore have:

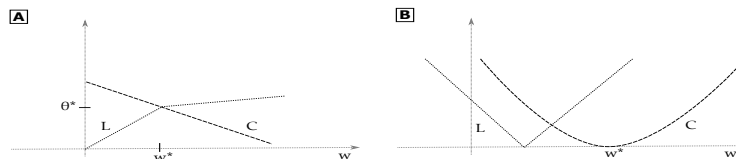$$L(w) \geq L(w^*) \quad \forall w \in W, C(w) \leq C(w^*) \tag{14}$$

Fig. 2. (A) Unattainable Constrained Optimum; (B) Numerical Issues for w^*

We are interested in the conditions for w^* to be optimal for the regularized formulation, for some multiplier vector λ. This is true iff:

$$L(w) + \lambda^\top C(w) \geq L(w^*) + \lambda^\top C(w^*) \quad \forall w \in W \tag{15}$$

which can rewritten as:

$$\lambda^\top \Delta C(w, w^*) + \Delta L(w, w^*) \geq 0 \quad \forall w \in W \tag{16}$$

If $\Delta C(w, w^*) = 0$, then Equation (16) is trivially satisfied for every multiplier vector, due to Equation (14). Otherwise, at least some component in $\Delta C(w, w^*)$ will be non-null, so that we can write:

$$\lambda_j \Delta C_j(w, w^*) + \lambda_{\bar{j}}^\top \Delta C_{\bar{j}}(w, w^*) + \Delta L(w, w^*) \geq 0 \tag{17}$$

If $\Delta C_j(w, w^*) < 0$, we get:

$$\lambda_j \leq -\frac{\Delta L(w, w^*) + \lambda_{\bar{j}}^\top \Delta C_{\bar{j}}(w, w^*)}{\Delta C_j(w, w^*)} \quad \forall w \in W \mid C_j(w) < C_j(w^*) \tag{18}$$

I.e. a series of upper bounds for λ_j. If $\Delta C_j(w, w^*) > 0$, we get:

$$\lambda_j \geq -\frac{\Delta L(w, w^*) + \lambda_{\bar{j}}^\top \Delta C_{\bar{j}}(w, w^*)}{\Delta C_j(w, w^*)} \quad \forall w \in W \mid C_j(w) > C_j(w^*) \tag{19}$$

I.e. a series of lower bounds on λ_j. From these the original result is obtained. □

The main consequence of Theorem 2 is that the reported system of inequalities may actually admit no solution, meaning that *some constrained optima may be unattainable* via Algorithm 1. This is the case for the optimum w^* (for threshold θ^*) in the simple example from Figure 2, since any multiplier value will result in an unbounded regularized problem. This is a potentially serious limitation of regularized methods: the actual severity of the issue will depend on the specific properties of the loss, regularizer, and ML model being considered.

Numerical Issues Theorem 2 highlights another issue of regularized approaches, arising when assignments with constraint violations arbitrarily close to $C(w^*)$ exist. The denominator in Equation (13) becomes vanishingly small: this may

result in arbitrarily high lower bounds or arbitrarily small upper bounds. Reaching a specific optimum for the constrained problem may require *extremely high or extremely low multipliers*, which may cause numerical issues at training time. An example is depicted in Figure 2B, where a regularizer with vanishing gradient and a loss with non-vanishing gradient are combined – the constrained optimum w^* is reached via Algorithm 1 only for $\lambda \to \infty$.

Differentiability Besides the ones reported here, one should be wary of pitfalls that are not immediately related to Algorithm 1. Many regularization based approaches for constraint injection, for example, require differentiability of the C function, which is often obtained by making approximations. For instance, in Equation (3) differentiability does not hold due to the use of binary variables; relaxing the integrally constraint address the issue, but allows to satisfy the constraints by assigning 0.5 to all outputs, i.e. by having completely uncertain, rather than balanced, predictions.

3 Conclusions

Combining the ML and optimization paradigms is a research avenue still under ongoing exploration by the AI community, with advancements towards ethical and trustworthy AI (e.g. by making sub-symbolic models fair and explainable). A possible method to merge these paradigm consists in adding a regularization term to the loss of a learner, to constrain its behaviour. In this note we tackle the issue of finding the right balance between the loss and the regularization term; typically, this search is performed by adjusting a set of multipliers until the desired compromise is reached. The key results of this paper is the formal demonstration that this type of approach *cannot guarantee to find all optimal solutions*. In particular, in the non-convex case there might be optima for the constrained problem that do not correspond to any multiplier value. This result clearly hinders the applicability of regularizer-based methods, at least unless more research effort is devoted to discover new formulations or algorithms.

References

1. Aghaei, S., Azizi, M.J., Vayanos, P.: Learning optimal and fair decision trees for non-discriminative decision-making. In: Proceedings of AAAI. pp. 1418–1426 (2019)
2. Cotter, A., Jiang, H., et al.: Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. Journal of Machine Learning Research **20**(172), 1–59 (2019)
3. Diligenti, M., Gori, M., Saccà, C.: Semantic-based regularization for learning and inference. Artif. Intell. **244**, 143–165 (2017)
4. Fioretto, F., Mak, T.W., et al.: A lagrangian dual framework for deep neural networks with constraints. arXiv preprint arXiv:2001.09394 (2020)
5. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)