STOCHASTIC PENALTY-BARRIER METHOD FOR CONSTRAINED MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Constrained Machine Learning (CML) enables fairness-aware training, physics-informed neural networks, and integration of symbolic domain knowledge into statistical models. In this work, we introduce the Stochastic Penalty-Barrier Method (SPBM) for CML problems. SPBM extends classical penalty and barrier methods by incorporating an exponential averaging of the dual variables, a stabilized penalty schedule, and the Moreau envelope to handle non-smoothness. We analyze the bias that mini-batching introduces in the barrier function and show that the feasible set of the resulting transformed problem is contained within the original one. We compare SPBM with CML baselines across multiple fairness and physics informed neural networks experiments. We find that SPBM is competitive with state-of-the-art methods. We also observe, on our fairness-based computational benchmark, that the per-epoch runtime of CML methods is largely independent of the number of constraints, and within $1.3\times$ of the per-epoch runtime of regularized Adam, for a number of constraints ranging from 90 to 9900.

1 INTRODUCTION

In recent years, artificial intelligence (AI) has become an indispensable part of many fields such as healthcare, credit scoring, criminal justice or the car industry. In such high-risk settings, AI predictions should be as accurate as possible, but also meet critical requirements, whether normative, such as fairness or safety, or structural, such as consistency with physical laws. There, constrained machine learning (CML) becomes increasingly important (Ramirez et al., 2025) enabling the enforcement of fairness constraints (Li et al., 2022; Buyl et al., 2024), safety constraints (Achiam et al., 2017), or ensuring consistency with physical laws (Lu et al., 2021; Son et al., 2023; Song et al., 2024).

From a mathematical point of view, general CML is cast as an optimization problem with a stochastic non-convex non-smooth objective subject to stochastic non-convex non-smooth constraints:

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \quad g(x) \leq 0, \tag{1}$$

where $f(x) = \mathbb{E}_{\xi\sim\Xi}[f(x,\xi)]$ and $g(x) = \mathbb{E}_{\xi\sim\Xi}[g(x,\xi)]$. On the one hand, there is a substantial and recent body of work on CML with simplifying assumptions: Curtis et al. (2026); Fang et al. (2026) focus on deterministic constraints; Curtis et al. (2025) consider deterministic linear constraints; Necoara & Nedić (2021); Necoara & Singh (2022); Singh et al. (2024); Singh & Necoara (2024); Boob et al. (2023) address stochastic constrained convex composite problems; finally, Huang & Lin (2023) assume stochastic non-convex and weakly convex constraints in their Stochastic Switching Subgradient (SSw). Interestingly, SSL-ALM, a method of Huang et al. (2025), was shown by Kliachkin et al. (2026) to be the state of the art for general CML, despite being proposed for stochastic weakly-convex objective with linear constraints. On the other hand, Kliachkin et al. (2026) identify that the above-mentioned settings do not capture realistic CML problems. Therefore;

> *It remains unclear how to implement CML efficiently, and little can be said about the theoretical guarantees of methods for these problems.*

In this paper, we aim to partially bridge this gap. Firstly, we propose a new method, called Stochastic Penalty-Barrier Method (SPBM), and give a theoretical analysis of the transformed problem SPBM solves and of a mild convergence result. Secondly, we analyse the computational demand of CML

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

methods to answer whether it is feasible to train deep neural networks with constraints. Lastly, we show why the general CML setting is difficult to solve. Specifically, we exhibit an empirical risk minimization (ERM) problem built from a two-parameter single-ReLU network that neither SPBM nor state-of-the-art methods can solve. This raises the question: *What are natural assumptions that would include a general CML problem while remaining strong enough to prove convergence results for **some** practical algorithm?* In this paper, we assume the mini-batch estimates $\bar{f}, \bar{g}$ to be integrable in ξ, and locally Lipschitz for every realisation. In addition, we assume a tame setting, in which both the objective and the constraints are definable in a common o-minimal structure, enabling the mild convergence result of SPBM. In Appendix D we give background on o-minimality and argue that tameness is a natural assumption in deep learning. However, whether the o-minimal assumption is an answer to our question remains unresolved. The contributions of this paper are as follows:

- *New method:* We introduce the Stochastic Penalty-Barrier Method (SPBM) for CML problems, which extends deterministic penalty-barrier methods (Ben-Tal & Zibulevsky, 1997) by exponential averaging of the dual variables, a stabilized penalty schedule, and the Moreau envelope to handle non-smoothness. Interestingly, for a specific choice of parameters, SPBM reduces to a SSL-ALM-like primal-dual update (see Appendix E.2). We compare SPBM with CML baselines across multiple fairness and physics-informed neural networks experiments. Our experiments show that SPBM matches or outperforms SSL-ALM.
- *Analysis:* We analyze the bias that mini-batching introduces in the barrier function and show that the feasible set of the resulting transformed problem is contained within the original one. We also prove almost sure convergence of expected values of the Lagrangian in the o-minimal setting, although *we provide no complementarity or feasibility* guarantees.
- *Why is CML difficult:* we provide a simple CML instance on which SPBM, SSL-ALM, SSw, and methods that assume simplified weakly-convex settings fail to achieve feasibility, and can remain arbitrarily far from the feasible set; complementing known hardness results for unconstrained nonsmooth non-convex optimization (Zhang et al., 2020; Arjevani et al., 2023).
- *Open-source implementation:* we release our implementation, in the form of a drop-in addition to PyTorch unconstrained neural network training loops requiring only a few additional lines of code.

Related work. Several works offer different routes to CML. Cotter et al. (2019) tackle indicator constraints via a two-player game whose solution, although approximately feasible, is a mixture of up to $m + 1$ models, which is impractical when m is large. Chamon et al. (2022) reduce the original stochastic constrained problem to a deterministic unconstrained one, and bound the resulting duality gap. The bound requires rich parametrization of the network, which may fail for smaller networks. Hounie et al. (2023) build upon this theory and provide an adaptive relaxation method that loosens constraints that are harder to satisfy. Elenter et al. (2024) bound the infeasibility of Lagrangian minimizers associated with optimal dual variables by reparametrizing the non-convex CML problem over a function space. Their final result, however, applies to exact minimizers of the Lagrangian rather than to the iterates produced in practice, and relies on the Fréchet derivative of the (unparametrized) constraints having full row rank at the optimum – a LICQ condition that they acknowledge is not straightforward to satisfy. Libraries for CML include: Cooper, the PyTorch framework for constrained deep learning (Gallego-Posada et al., 2025), implementing a single-loop ALM without smoothing; and humancompatible.train (Kliachkin et al., 2025) with smoothed ALM.

Recent works in optimization by Necoara & Nedić (2021); Necoara & Singh (2022); Singh et al. (2024); Singh & Necoara (2024) and Boob et al. (2023) address functionally constrained convex composite problems. This allows guarantees, but does not extend to the non-convex setting of problem (1). There also exist SQP-based methods; their assumptions include stochastic objectives subject to deterministic constraints (Fang et al., 2026; Curtis et al., 2024b), stochastic non-convex objectives subject to deterministic constraints (Berahas et al., 2023; Curtis et al., 2024a; Fang et al., 2024; Na et al., 2023a), stochastic non-convex objective subject to deterministic non-convex constraints (Na et al., 2023b), and finally stochastic objectives subject to stochastic constraints (Oztoprak et al., 2023). These SQP methods, however, require the computation and storage of a Hessian, which makes them impractical in our case where f is a neural network. An approach similar to SQP was proposed by Ma et al. (2020) for weakly-convex objectives and constraints, which solves a sequence of strongly convex subproblems formed with a quadratic regularization term; and by Facchinei & Kungurtsev (2023) for stochastic objective and constraints, which computes the update direction by solving a number of quadratic subproblems in each iteration. Methods incorporating an augmented Lagrangian include (Shi et al., 2022) with stochastic non-convex objective under deterministic non-convex inequality

and/or equality constraints and (Bollapragada et al., 2023) with stochastic objectives subject to deterministic linear constraints. Finally, some interior point methods have been proposed to solve non-convex constrained optimization (Curtis et al., 2025) (bound constraints) and (Curtis et al., 2026) (nonlinear inequality constraints). However, they assume a strictly feasible initial point or deterministic constraints, making them unsuitable for our nonconvex stochastic constrained setting. Yang et al. (2025); Liu & Xu (2026) present single-loop algorithms designed for non-smooth weakly-convex problems. The penalty-barrier multiplier method (PBM) was generalized to solve deterministic non-convex non-linear problems (Kočvara & Stingl, 2003a;b; 2012). Recently, De Marchi & Themelis (2025) defined a deterministic non-convex variant of problem (1) by combining a barrier term with an L_1 penalty into a new objective function. Our method is also closely related to the proximal augmented Lagrangian method, which has been studied since the 1970s (Rockafellar, 1976; 1974; 1973b;a; Powell, 1969; Hestenes, 1969; Kong et al., 2019; 2023; Pu et al., 2024; Lin et al., 2022; Zhang & Luo, 2020; Zhang et al., 2022).

2 Review of Penalty/Barrier Multiplier Method (PBM)

The PBM method (Ben-Tal & Zibulevsky, 1997) was proposed for convex programming problems:

$$f^* = \inf\{f(x) : g_i(x) \le 0,\ i = 1, \dots, m,\ x \in \mathbb{R}^n\}, \tag{2}$$

where $f, g_1, \dots, g_m$ are closed proper convex functions. Ben-Tal and Zibulevsky showed that PBM is closely related to the proximal point algorithm in the sense of the sequence of generated dual variables. One of the main ideas behind the method is the transformation of the constraints onto a uniform space using a strictly increasing twice differentiable strictly convex function φ with some additional properties; see (Ben-Tal & Zibulevsky, 1997, Section 3) for details. These properties imply that, given a positive number p, the function $t \mapsto \varphi(t/p)$ is convex increasing and

$$p\varphi(t/p) \le 0 \quad \text{if and only if} \quad t \le 0. \tag{3}$$

Using Eq. (3), the original problem (2) can be rewritten as

$$f^* = \inf\{f(x) : p_i\varphi(g_i(x)/p_i) \le 0,\ i = 1, \dots, m,\ x \in \mathbb{R}^n, p \in \mathbb{R}^m_{++}\}, \tag{4}$$

where p_i is referred to as the *penalty parameter* associated with the i-th constraint, and φ is referred to as the *penalty-barrier function*. The function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^m_{++} \to \mathbb{R}$, referred to as the *Lagrangian* for problem (4), has the form $\mathcal{L}_p(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i p_i \varphi(g_i(x)/p_i)$ and it is used in the following iterative scheme for solving (4):

$$x^{k+1} = \arg\min_x \mathcal{L}_{p^k}(x, \lambda^k) \tag{5}$$

$$\lambda^{k+1} = \lambda^k \odot \varphi'\left(g(x^{k+1}) \oslash p^k\right) \tag{6}$$

$$p^{k+1} = \pi^k(\lambda^{k+1}) \tag{7}$$

where $\odot$ and $\oslash$ denote the element-wise product and division, and φ' and π^k are applied coordinate-wise. The initial *dual variable* λ^0 is required to be positive and $\pi^k : \mathbb{R}_{++} \to \mathbb{R}_{++}$ is referred to as the *penalty updating function*. We recall two glued penalty/barrier functions: the *Quadratic-Logarithmic penalty/barrier function* φ_{QL} and the *Quadratic-Reciprocal penalty/barrier function* φ_{QR}:

$$\varphi_{\text{QL}}(t) = \begin{cases} t + \frac{1}{2}t^2 & \text{if } t \ge -\frac{1}{2}, \\ -\frac{1}{4}\log(-2t) - \frac{3}{8} & \text{if } t < -\frac{1}{2}, \end{cases} \quad \varphi_{\text{QR}}(t) = \begin{cases} t + \frac{1}{2}t^2 & \text{if } t \ge -\frac{1}{3}, \\ \frac{32}{27}(\frac{1}{1-t}) - \frac{7}{6} & \text{if } t < -\frac{1}{3}. \end{cases} \tag{8}$$

The penalty updating functions in Ben-Tal & Zibulevsky (1997) are constructed so that in the limit $p \to 0$. The overall idea of the algorithm is as follows. As in Lagrangian methods, larger dual variables are enforced for violated constraints and smaller ones for satisfied constraints which is achieved by (6). Since φ' is strictly increasing with $\varphi'(0) = 1$ and $p^k > 0$, we have

$$\varphi'\left(g_i(x^{k+1})/p^k\right) \begin{cases} > 1 & \text{if } g_i(x^{k+1}) > 0, \\ = 1 & \text{if } g_i(x^{k+1}) = 0, \\ \in (0, 1) & \text{if } g_i(x^{k+1}) < 0. \end{cases}$$

The penalty parameters also govern the speed of change in the dual variables (therefore also the speed of convergence towards the feasible region): the smaller the value of the ith coordinate of p^k, the larger the update of the ith dual variable. This dual update direction is larger either towards 0 (for feasible iterate $g_i(x^{k+1}) < 0$) or towards unbounded values (for infeasible iterate $g_i(x^{k+1}) > 0$).

Algorithm 1 Stochastic Penalty-Barrier Method (SPBM). `PrimalUpdate` can be one Adam / SGD step; state carries its internal metadata across iterations as well as learning rate scheduling.

1: **Input:** $x^0 \in \mathbb{R}^n$; $\lambda^0 \in \mathbb{R}^m_{++}$; $p^0 = \mathbf{1}$(or $p^0 = \rho\,\lambda^0$ if $\hat{\pi} = \hat{\pi}^{\mathrm{ALM}}_\rho$); $s^0 = x^0$; $\kappa \in [0,1)$(alternatively $\rho \in \mathbb{R}_{++}$, if $\hat{\pi} = \hat{\pi}^{\mathrm{ALM}}_\rho$); Primal optimizer state state^0; batch size B; $\hat{\pi} \in \{\hat{\pi}^{\mathrm{ALM}}_\rho, \hat{\pi}^{\mathrm{ADAPT}}_\kappa\}$, $\gamma \in [0,1)$, $\mu \geq 0$, and $\delta \in [0,1)$; $\Lambda \in \mathbb{R}_{++}$.
2: **for** $k = 0$ **to** K **do**
3: Sample $\xi_1, \ldots, \xi_B \overset{\text{iid}}{\sim} \Xi$
4: $\bar{y}^k = -\overline{\nabla}\mathcal{L}_{p^k}(x^k, \lambda^k) - \mu(x^k - s^k)$ {Moreau-augmented gradient, (10)}
5: $x^{k+1}, \mathrm{state}^{k+1} = \mathtt{PrimalUpdate}(x^k, \bar{y}^k, \mathrm{state}^k)$ {Primal variable step, (5)}
6: $\lambda^{k+1} = \mathrm{Proj}_{[0,\Lambda]^m}\left[\gamma\lambda^k + (1-\gamma)\,\lambda^k \odot \varphi'(\overline{g}(x^k) \oslash p^k)\right]$ {Dual variable step, (6)}
7: $p^{k+1} = \hat{\pi}(\overline{g}(x^k), p^k, \lambda^{k+1})$ {Penalty parameter step, (7);(11)/(12)}
8: $s^{k+1} = \delta s^k + (1-\delta)x^{k+1}$ {Prox-center update}
9: **end for**

3 Stochastic Penalty/Barrier Method (SPBM)

Our method, Stochastic Penalty-Barrier Method (SPBM), extends the penalty and barrier framework of Ben-Tal & Zibulevsky (1997) to the stochastic setting. While the general structure follows their iterative update rules, we propose three key modifications. Firstly, to deal with the non-smoothness of the resulting objective, we perform an inexact proximal point update to recover a tractable minimization step. In theory, we assume our oracle to be a Clarke subgradient (denoted as ∂ operator); however, for the maximum compatibility and impact, our implementation uses PyTorch (which might not output an element of Clarke (Kakade & Lee, 2018)). Secondly, the application of the barrier function in the stochastic setting transforms the original problem (1) into a stricter version (proven in Section 3.1), which motivates different penalty-updating functions compared to the original deterministic one. Lastly, we utilize exponential averaging. In the following paragraphs, we address the main challenges in detail. The full algorithm is presented in Algorithm 1.

Mini-batching notation. We use mini-batching with batch size $B \in \mathbb{N}$ to sample both the objective and the constraint distributions. A mini-batch is a tuple $\{\xi_b\}_{b=1}^B$ of i.i.d. random variables $\xi_1, \ldots, \xi_B \overset{\text{iid}}{\sim} \Xi$. Stochastic estimates are then computed from the mini-batch as follows:

$$\overline{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f(x, \xi_b), \quad \overline{g}(x) = \frac{1}{B}\sum_{b=1}^{B} g(x, \xi_b). \tag{9}$$

Similarly, we denote a stochastic estimate of the Lagrangian $\mathcal{L}_p(x, \lambda)$ and its subgradient:

$$\overline{\mathcal{L}}_p(x, \lambda) = \overline{f}(x) + \sum_{i=1}^{m} \lambda_i p_i\, \varphi(\overline{g}_i(x)/p_i), \qquad \overline{\nabla}\mathcal{L}_p(x, \lambda) \in \partial_x \overline{\mathcal{L}}_p(x, \lambda).$$

Challenge 1: Non-smoothness. We treat SPBM as a subgradient method. The intuition is to optimize over the approximate solution of the Moreau envelope of the Lagrangian, with given p^k, s^k, and λ^k:

$$\min_x \left(\overline{\mathcal{L}}_{p^k}(x, \lambda^k) + \frac{\mu}{2}\|x - s^k\|^2\right). \tag{10}$$

In theory, if $\overline{\mathcal{L}}_{p^k}(\cdot, \lambda^k)$ is ρ-weakly convex and $\mu > \rho$, its Moreau envelope is a smooth approximation that shares the same minimizers (Beck, 2017); we use this as motivation, noting that weak convexity is not established in our setting. In practice, we aim to find an approximate minimizer (with no proven guarantees) of the Moreau envelope by performing a single step of a gradient-based primal update with respect to x, e.g. Adam or SGD (`PrimalUpdate` at line 5). This is the same as performing an inexact proximal-point step on the function $\frac{1}{\mu}\overline{\mathcal{L}}_{p^k}(\cdot, \lambda^k)$ (Beck, 2017). We then utilize exponential averaging on the argument s of the proximal operator (line 8, inspired by Zhang & Luo (2020), Huang et al. (2025)).

Challenge 2: Penalty updating functions. The original penalty updating function in the deterministic iterative scheme Eq. (7) implies that eventually $p \to 0$. In the stochastic setting, however, such

procedure might tighten the original feasible set at a cost of a worse objective value; on the other hand, it introduces a controllable bias for constraints that are difficult to satisfy (details in Section 3.1). Therefore, we propose an adaptive penalty-updating formula $\hat{\pi} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$, which adaptively decreases penalties only for unsatisfied constraints, while maintaining larger penalties for satisfied constraints. Otherwise, the motivation behind penalty parameters described in PBM Section 2 consists. The adaptive formula reads as follows:

$$\hat{\pi}_\kappa^{\mathrm{ADAPT}}(\bar{g}, p, \lambda) = \mathrm{Proj}_{[\ell_c, u_c]^m}\left[\left(\kappa + \frac{(1-\kappa)}{\varphi'(\bar{g}) + \varepsilon}\right) p\right], \quad \varepsilon = 10^{-8}, \tag{11}$$

where all operations (sum, product, division) are performed coordinate-wise, and $\mathrm{Proj}_{[\ell_c, u_c]^m}$ denotes the coordinate-wise clipping onto a box $[\ell_c, u_c]^m$ for $0 < \ell_c \leq u_c$. As a second choice of the penalty-updating formula, we suggest

$$\hat{\pi}_\rho^{\mathrm{ALM}}(\bar{g}, p, \lambda) = \mathrm{Proj}_{[\ell_c, u_c]^m}\left[\lambda \cdot \rho\right], \quad \text{where } \rho > 0. \tag{12}$$

Interestingly, as shown in Appendix E.2, for Eq. (12), SPBM reduces to SSL-ALM-like primal-dual update. Note *(i)* by choosing *Eq.* (12), the hyperparameter κ becomes redundant, same applies to ρ in case of Eq. (11); *(ii)* clipping safeguards are proposed for purposes of a controllable bias (Section 3.1); *(iii)* hyperparameters κ, ρ can be thought of as step sizes of the penalty parameters.

3.1 Theoretical Analysis of SPBM

In the deterministic setting, it is immediate that the original formulation (2) and its transformed counterpart (4) define the same problem: the reformulation (4) is exact. In the stochastic setting, however, SPBM evaluates the penalty/barrier function on mini-batch estimates $\bar{g}_i$ rather than on the exact expectations, and therefore implicitly targets the problem

$$\min_{x \in \mathbb{R}^n} \ \mathbb{E}_\xi[f(x,\xi)] \quad \text{s.t.} \quad \mathbb{E}_{\xi_{1:B}}\left[p_i\, \varphi\big(\bar{g}_i(x)/p_i\big)\right] \leq 0, \quad i = 1, \ldots, m, \tag{13}$$

instead of Eq. (1). Here $\bar{g}_i$ is the mini-batch estimator defined in Eq. (9) with $\xi_1, \ldots, \xi_B \overset{\text{iid}}{\sim} \Xi$, and $\mathbb{E}_{\xi_{1:B}}$ denotes the expectation over the mini-batch; the objective is unaffected, since $\bar{f}$ is an unbiased estimator of $\mathbb{E}_\xi[f(x,\xi)]$ and enters SPBM linearly. The constraints, in contrast, are distorted: since φ is convex and non-linear by definition, Jensen's inequality (Jensen, 1906) gives, for any fixed $p_i > 0$,

$$p_i\, \varphi\big(\mathbb{E}_{\xi_{1:B}}[\bar{g}_i(x)]/p_i\big) \ \leq \ \mathbb{E}_{\xi_{1:B}}\left[p_i\, \varphi\big(\bar{g}_i(x)/p_i\big)\right]. \tag{14}$$

The difference between the right- and left-hand sides of Eq. (14) is known as the Jensen gap (Strukov & Timan, 1977; Gao et al., 2017; Abramovich & Persson, 2016). Its non-negativity means that the constraints in Eq. (13) are at least as strict as those of the original problem (1): any point feasible for the transformed problem is feasible for the original one. This is formalized in Proposition 1:

Proposition 1 (Transformed-feasibility gives original-feasibility)**.** *Consider a penalty/barrier function $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$ (8), a positive integer $B \in \mathbb{N}$, and random variables $\xi_1, \ldots, \xi_B \overset{\text{iid}}{\sim} \Xi$. Consider the feasible set of the original problem (1) and of the transformed problem (13), defined as*

$$\mathcal{S} := \bigcap_{i=1}^m \big\{x \in \mathbb{R}^n : \mathbb{E}_\xi\left[g_i(x,\xi)\right] \leq 0\big\}, \quad \mathcal{S}^B_{\mathrm{transform}} := \bigcap_{i=1}^m \big\{x \in \mathbb{R}^n : \mathbb{E}_{\xi_{1:B}}\left[p_i \varphi\big(\bar{g}_i(x)/p_i\big)\right] \leq 0\big\}.$$

Then, $\mathcal{S}^B_{\mathrm{transform}} \subseteq \mathcal{S}$. *(For proof see Appendix E.1.)*

The next proposition bounds the corresponding Jensen gap (see also Appendix E.1 for proof).

Proposition 2 (Jensen gap bounds)**.** *Let $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$ be a penalty/barrier function as defined in Eq. (8), let $B \in \mathbb{N}$ be a positive integer, let $\xi_1, \ldots, \xi_B \overset{\text{iid}}{\sim} \Xi$, and assume $\mathrm{Var}(g_i(x,\xi)) < \infty$. Then, for every $x \in \mathbb{R}^n$, $i \in \{1, \ldots, m\}$, and $p_i > 0$,*

$$\mathbb{E}_{\xi_{1:B}}\left[p_i \varphi\big(\bar{g}_i(x)/p_i\big)\right] - p_i \varphi\big(\mathbb{E}[g_i(x,\xi)]/p_i\big) \ \leq \ \frac{\mathrm{Var}\big(g_i(x,\xi)\big)}{2Bp_i}.$$

Two consequences follow. Firstly, as $B \to \infty$, the gap vanishes and the transformed constraints recover the original ones; the bias is thus a finite-batch effect, decaying at rate $1/B$. Secondly, as

$p_i \to 0$, the gap grows without bound. This is in direct tension with the penalty schedule (7) of PBM, which drives $p \to 0$, and justifies our modified update (11), whose clipping to $[\ell_c, u_c]$ keeps the bias bounded by $\mathrm{Var}(g_i(x,\xi))/(2B\ell_c)$ uniformly over iterations. The bound also guides the choice of the lower clipping bound ℓ_c: when SPBM struggles to reach feasibility, decreasing ℓ_c tightens the transformed constraints (introduces a controllable bias) and enlarges the dual updates of violated constraints, driving the iterates toward $\mathcal{S}^B_{\text{transform}}$ more aggressively - and, by Proposition 1, any point of $\mathcal{S}^B_{\text{transform}}$ is feasible for the original problem. There is a price for this: a smaller ℓ_c shrinks the surrogate feasible set and may exclude solutions of Eq. (1) with lower objective value.

3.2 Convergence Analysis

Consider the two-parameter stochastic fitting problem of Appendix E.4: a single ReLU unit under a fairness constraint. We show it is an instance of Eq. (1) (Proposition 6) but not of the weakly convex setting (Proposition 5); so related work assuming weak convexity excludes even a single-ReLU ERM task. We then prove that both state-of-the-art CML methods and SPBM make no progress from an initialisation that can be arbitrarily far from the feasible region (proof can be found in Appendix E.4).

Proposition 3 (No progress on the instance of Appendix E.4)**.** *Assume the problem described in Appendix E.4, with $x^k = w^k + b^k$ and $\theta^k = (w^k, b^k)$. Then SPBM, SSw, and SSL-ALM, and any method with an update scheme of type $\theta^{k+1} = \theta^k + h(\theta^k)$ for $h(\theta^k) = 0$ whenever $\partial \bar{f}(\theta^k) = \partial \bar{g}(\theta^k) = \{0\}$, fail to reach the feasible region for preactivation $x^0 < 0$ for any realisation. Moreover, it holds for any realisation and any $k \geq 0$*

$$\mathrm{dist}(x^k, \{x : g(x) \leq 0\}) = \mathrm{dist}(x^0, \{x : g(x) \leq 0\}).$$

We therefore aim at a weaker, but non-empty, guarantee. For this, we require additional assumptions which are described in detail in Algorithm 2 and informally here. Let the dataset have size M, and the batch size be B, and define the function $\mathcal{L}^{\text{emp}}_p(\cdot, \lambda) : \mathbb{R}^n \to \mathbb{R}$, and a conservative field of $-\mathcal{L}^{\text{emp}}_p(\cdot, \lambda)$ denoted as $\mathcal{D}_p(\cdot, \lambda) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ such that

$$\mathcal{L}^{\text{emp}}_p(\cdot, \lambda) = \mathbb{E}_{\xi_{1:B}}\left[\bar{\mathcal{L}}_p(\cdot, \lambda)\right], \quad \mathcal{D}_p(\cdot, \lambda) = -\frac{1}{M^B} \sum_{\mathbf{j} \in \{1,\dots,M\}^B} \partial_x \overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j}), \tag{15}$$

where $\overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j})$ is the estimate dependent on batch draw $\mathbf{j}$. We refer the reader to Appendix D.4 for background on conservative fields. Informally, the mild convergence result reads as follows:

Theorem 1 (Primal criticality – informal statement)**.** *Assume definable and locally Lipschitz $\bar{g}, \bar{f}$, almost sure boundedness of the primal variable. Run SPBM with $\gamma^k, \kappa^k \to 1$ with summable deviations, square-summable but not summable primal schedule, dual clipping onto $[0, \Lambda]^m$, $\bar{y}^k$ being an element of Clarke, SGD dynamics (Algorithm 2). Then almost surely there exist random $\hat{\lambda} \in [0, \Lambda]^m, \hat{p} \in [\ell_c, u_c]^m$ such that* (1) *$\lambda^k \to \hat{\lambda}$, $p^k \to \hat{p}$;* (2) *$\|x^k - s^k\| \to 0$;* (3) *the function values $\{\mathcal{L}^{\text{emp}}_{\hat{p}}(x^k, \hat{\lambda})\}_{k \geq 0}$ converge;* (4) *every limit point $\hat{x}$ of $\{x^k\}_{k \geq 0}$ satisfies $0 \in \mathcal{D}_{\hat{p}}(\hat{x}, \hat{\lambda})$.*

Proof sketch (full proof in Appendix E.3). SPBM is a primal-dual algorithm. As such, it performs descent on the primal variables, and ascent on the dual variables of the penalised Lagrangian $\mathcal{L}^{\text{emp}}_{p^k}(\cdot, \lambda^k)$. This is the main difficulty in obtaining results for SPBM, and for primal-dual methods more generally. We consider the general setting where the objective function and constraints are neither differentiable nor (weakly)-convex. We build on the results obtained by Davis et al. (2018); Bolte & Pauwels (2021) for SGD in such a general setting. Their result would provide primal variable convergence were the dual variables and penalty parameters fixed.

We fill this gap by assuming summability of $(1 - \gamma^k)$ and $(1 - \kappa^k)$. This gives the dual and penalty paths finite length, so they admit limit points $\hat{\lambda}$ and $\hat{p}$ almost surely. A central observation is that fixing the dual and penalty variables reduces SPBM to SGD, which allows applying the above mentioned convergence guarantees. Therefore, SPBM with dual and penalty variables fixed to their limiting values converges to critical points of $\mathcal{L}^{\text{emp}}_{\hat{p}}(\cdot, \hat{\lambda})$. The rest of the proof consists in bounding the distance between the dual and penalty iterates of the actual SPBM scheme and the SPBM scheme with fixed dual and penalty variables. Along the way, we justify that the conservative field associated with the penalty barrier, defined in Eq. (15), is definable in the same structure as the objective and constraint functions, which allows applying the standard SGD guarantees. □

The function $\mathcal{L}_p^{\text{emp}}(\cdot, \lambda)$ is the Lagrangian of the transformed problem (13), so item *(4)* says that $\hat{x}$ is stationary for (13) rather than for the original problem. This is informative in two ways: The feasible set of (13) lies inside $\mathcal{S}$ by Proposition 1, so a point that is feasible for the transformed problem is feasible for the original one, and by Proposition 2 the gap is at most $\text{Var}(g_i(x,\xi))/(2B\ell_c)$. By the Clarke sum rule $-\partial\mathcal{L}_{\hat{p}}(\hat{x}, \hat{\lambda}) \subseteq \mathcal{D}_{\hat{p}}(\hat{x}, \hat{\lambda})$, hence *(4)* is implied by, but does not imply Clarke criticality. As already discussed, the theorem *does not* provide any primal feasibility or complementarity guarantees. Full analysis is provided in Appendix E.3. Algorithm 2 is used in the analysis, whereas Algorithm 1 is used for benchmarking.

4 Numerical Experiments

Overview. We create a benchmark comparing SPBM to three other optimization methods: Adam (Kingma & Ba, 2014) (baseline with no constraints[1]), SSw (Huang & Lin, 2023), and SSL-ALM (Huang et al., 2025). Our benchmark contains 5 classification problems based on the ACSIncome dataset (`folktables`; binary classification), the Dutch Census dataset (binary classification), and the CIFAR-10 and CIFAR-100 datasets (multiclass classification), and three experiments involving Physics-Informed Neural Networks (PINNs). We benchmark the performance of each method on the training data of the datasets. We present the constraints of each problem below.

Fairness and Regularization as constrained optimization. In classification experiments, we implement two types of constraints. The first type is deterministic and it enforces weights of a neural network to lie in an ε ball:

$$\|W_i\|_F \leq \varepsilon, \quad \text{for each layer } i, \tag{16}$$

where W_i is a weight matrix of the fully connected neural network and $\|\cdot\|_F$ is the Frobenius norm. The other constraint is stochastic and its goal is to enforce a similar prediction accuracy across different groups – either demographic groups (*e.g.* sex), or classes in the CIFAR datasets. We model this by constraining differences between all pairs of groups:

$$-\varepsilon \leq \mathbb{E}_{\xi\sim\text{group } R_i}[f(x,\xi)] - \mathbb{E}_{\xi\sim\text{group } R_j}[f(x,\xi)] \leq \varepsilon, \quad \text{for every } i \neq j, \tag{17}$$

which results, for a set of all demographic groups $\mathcal{R}$, in $m = |\mathcal{R}| \cdot (|\mathcal{R}| - 1)$ constraints.

Solving PDEs as constrained optimization. PINNs are neural networks trained to satisfy physical laws governed by nonlinear partial differential equations (PDEs). An example of such a physical law can be the Helmholtz equation:

$$\Delta u + u = q(z_1, z_2), \qquad \text{for every } (z_1, z_2) \in \mathcal{Z}, \tag{18}$$

$$u(z_1, z_2) = 0, \qquad \text{for every } (z_1, z_2) \in \text{Bd}\,\mathcal{Z}, \tag{19}$$

where $\mathcal{Z} = [-1, 1]^2 \subseteq \mathbb{R}^2$ and $\text{Bd}\,\mathcal{Z}$ is its boundary, and the function q is fixed (see Appendix B.1.1). Since boundary conditions are known to cause instability in PINN training (Wang et al., 2022; Son et al., 2023; Wang et al., 2021), it is natural to enforce them as constraints, yielding the constrained optimization problem:

$$\min_x \quad \mathbb{E}_{\xi\sim\Xi_{\mathcal{Z}}}\left[|Nu_{\text{nn}}(x,\xi) - h(\xi)|^2\right] \tag{20}$$

$$\text{s.t.} \quad \mathbb{E}_{\xi\sim\Xi_{\Gamma(i)}}\left[|N_i u_{\text{nn}}(x,\xi) - h_i(\xi)|^2\right] \leq \varepsilon, \quad \text{for each } i \in \{1,\ldots,m\}, \tag{21}$$

where $\varepsilon > 0$ is small, N and N_i are differential operators, u_{nn} is a neural network representing the solution with weights x, and h, h_i are prescribed. Each $\Gamma(i) \subseteq \text{Bd}\,\mathcal{Z}$ is the set on which the ith condition is imposed; $\Xi_{\mathcal{Z}}$ and $\Xi_{\Gamma(i)}$ denote the corresponding distributions of collocation points.

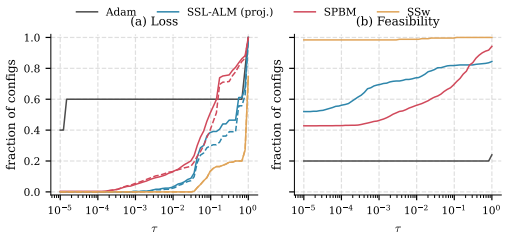
Experiments. We denote our experiments by **($\mathbf{E}_1$)–($\mathbf{E}_8$)** and summarize them in Table 1 together with the demographic attributes used in the fairness experiments. **($\mathbf{E}_6$)**, **($\mathbf{E}_7$)**, and **($\mathbf{E}_8$)** are based on the Helmholtz equation, viscous Burgers equation and Klein-Gordon equation, respectively; see Appendix B for details on the experimental setup such as the datasets, data processing, implementation details of methods, network architectures, choice of hyperparameters, number of seeds, and hardware.

[1]For PINNs, we use Adam to optimize the network as suggested in the original paper (Raissi et al., 2017) with a fixed regularization term (a hyperparameter) of the initial and boundary constraints.

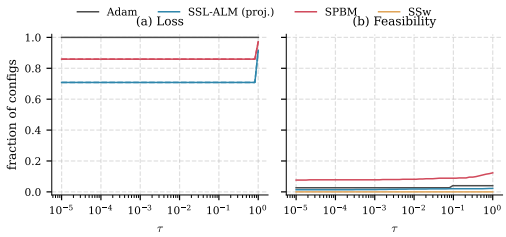
Table 1: Summary of experiments; MSE stands for Mean Squared Error.

Exp.	Dataset	Objective function f	Groups $\mathcal{R}$	Constraint g	# Constraints (m)	Bounds ε
$(\mathbf{E}_1)$	ACSIncome	binary cross-entropy	–	Bound constraint Eq. (16)	6	2
$(\mathbf{E}_2)$	ACSIncome	binary cross-entropy	6 (sex × marital)	Eq. (17)	30	10^{-1}
$(\mathbf{E}_3)$	Dutch Census	binary cross-entropy	18 (sex × age)	Eq. (17)	306	10^{-1}
$(\mathbf{E}_4)$	CIFAR-10	multiclass cross-entropy	10 (classes)	Eq. (17)	90	10^{-1}
$(\mathbf{E}_5)$	CIFAR-100	multiclass cross-entropy	100 (classes)	Eq. (17)	9900	10^{-1}
$(\mathbf{E}_6)$	Sampled $[-1,1]^2$	MSE Eq. (20) with Eq. (18)	–	Eq. (21) with Eq. (19)	1	10^{-4}
$(\mathbf{E}_7)$	Sampled $[0,1]\times[-1,1]$	MSE Eq. (20) with Eq. (22)	–	Eq. (21) with Eq. (23), Eq. (24)	2	10^{-4}
$(\mathbf{E}_8)$	Sampled $[0,1]^2$	MSE Eq. (20) with Eq. (25)	–	Eq. (21) with Eq. (26), Eq. (27), Eq. (28)	3	10^{-4}

Figure 1: Moré-Wild profiles over the hyperparameter grid. For each method, the curve gives the fraction of hyperparameters whose relative score is at most τ. A score of 0 means the run matched the best value reached by any method, and a score of 1 means it made no progress from initialisation. Earlier and higher curve indicates a method that is more robust to the hyper-parameter choice. See Section 4.1 for the definition. Loss metric is plotted only for feasible runs; dashed line indicates feasibility within 1% tolerance, solid line indicates feasibility within 10% tolerance. Feasibility is defined as $[\max_{i\in\{1,\dots,m\}} \bar{g}_i - \varepsilon]_+$, where $\bar{g}_i$ here is averaged over all seeds.



(a) $(\mathbf{E}_{2,3,4,5})$: Moré–Wild profile.



(b) $(\mathbf{E}_{6,7,8})$: Moré–Wild profile.

4.1 Methodology

Moré–Wild profiles. Inspired by Moré & Wild (2009), we measure performance across hyperparameter configurations rather than across problems. Let $\mathcal{E}$ be the set of experiments, $\mathcal{M}$ the set of methods, and $\mathcal{H}(\mathrm{mt})$ the hyperparameter grid of method mt. Let $\mathcal{Q}(\mathrm{ex},\mathrm{mt},\mathrm{hp})\in\mathbb{R}$ denote a metric; in our case training loss or constraint violation. Write $\mathcal{Q}^\star(\mathrm{ex}) := \min_{\mathrm{mt}\in\mathcal{M},\,\mathrm{hp}\in\mathcal{H}(\mathrm{mt})} \mathcal{Q}(\mathrm{ex},\mathrm{mt},\mathrm{hp})$, and denote the metric at initialization $\mathcal{Q}^0(\mathrm{ex})$, which is the same for all methods. For each triple we define the score:

$$\mathrm{score}(\mathrm{ex},\mathrm{mt},\mathrm{hp}) := \max\left(\frac{\mathcal{Q}(\mathrm{ex},\mathrm{mt},\mathrm{hp}) - \mathcal{Q}^\star(\mathrm{ex})}{\mathcal{Q}^0(\mathrm{ex}) - \mathcal{Q}^\star(\mathrm{ex})},\ 0\right).$$

For each method $\mathrm{mt}\in\mathcal{M}$ we plot the empirical CDF:

$$\mathrm{cdf}_{\mathrm{mt}}(\tau) := \frac{1}{|\mathcal{E}|\cdot|\mathcal{H}(\mathrm{mt})|}\Big|\big\{(\mathrm{ex},\mathrm{hp})\in\mathcal{E}\times\mathcal{H}(\mathrm{mt})\ :\ \mathrm{score}(\mathrm{ex},\mathrm{mt},\mathrm{hp})\le\tau\big\}\Big|,\qquad \tau\in\mathbb{R}_+.$$

In other words, a configuration passes the τ-test at $\tau=0$ only if it matches the best value reached by any method on that experiment; all non-divergent configurations pass by $\tau=1$. Therefore, a CDF that rises earlier and higher indicates a method that is more robust to the hyper-parameter choice.

Best configurations. We report a table of the best run per method over the hyperparameter grid. Selection is based on the tail-mean over the last N epochs ($N=5$ for fairness; $N=50$ for PINN experiments) in the following way: we take a configuration feasible within a 10% tolerance with the lowest training loss; if no feasible configuration is available, we choose the least infeasible one. The full hyperparameter grids and the final hyperparameter choices are listed in Appendix B. We also provide a computational benchmark based on $(\mathbf{E}_5)$ with a varying number of constraints. Finally, we provide convergence plots in Appendix C.2 with an ablation study of SPBM in Appendix C.1.

Table 2: Final results on our benchmark for SPBM, SSL-ALM, SSw, and Adam (unconstrained baseline). For each method and experiment we report the best-performing hyperparameter configuration from our grid, selected as the feasible run with the lowest training loss; see Section 4.1 for the selection rule. Values are means ± standard deviation over 5 seeds. **Bold** marks the best value and underline the second best within each experiment. We omit marking the unconstrained Adam. Feasibility is defined as $[\max_{i\in\{1,\dots,m\}} \bar{g}_i - \varepsilon]_+$, where $\bar{g}_i$ here is averaged over all seeds.

Exp.	Method	Best loss	Feasibility	Exp.	Method	Best loss	Feasibility
($\mathbf{E}_1$)	Adam	0.186 ± 0.004	52.39 ± 1.1	($\mathbf{E}_5$)	Adam	0.000 ± 0.000	0.000 ± 0.002
	SPBM	**0.373** ± 0.001	**0.000** ± 0.003		SPBM	**0.001** ± 0.000	**0.000** ± 0.002
	SSL-ALM	**0.373** ± 0.000	**0.000** ± 0.000		SSL-ALM	4.577 ± 0.004	**0.000** ± 0.009
	SSw	0.383 ± 0.007	**0.000** ± 0.035		SSw	4.587 ± 0.002	**0.000** ± 0.008
($\mathbf{E}_2$)	Adam	0.305 ± 0.002	0.210 ± 0.002	($\mathbf{E}_6$)	Adam (reg.)	0.153 ± 0.032	0.0085 ± 0.0064
	SPBM	**0.350** ± 0.003	0.002 ± 0.007		SPBM	**4010.463** ± 3274.490	0.0001 ± 0.0002
	SSL-ALM	0.351 ± 0.003	**0.000** ± 0.006		SSL-ALM	5347.374 ± 2673.665	**0.0000** ± 0.0002
	SSw	0.623 ± 0.013	**0.000** ± 0.006		SSw	6460.98 ± 446.31	0.0025 ± 0.0047
($\mathbf{E}_3$)	Adam	0.394 ± 0.000	0.497 ± 0.001	($\mathbf{E}_7$)	Adam (reg.)	0.000 ± 0.000	0.4901 ± 0.0000
	SPBM	**0.470** ± 0.002	0.006 ± 0.005		SPBM	**0.000** ± 0.000	**0.0000** ± 0.0000
	SSL-ALM	0.478 ± 0.004	0.003 ± 0.012		SSL-ALM	**0.000** ± 0.000	**0.0000** ± 0.0001
	SSw	0.669 ± 0.011	**0.000** ± 0.009		SSw	**0.000** ± 0.000	0.4945 ± 0.0037
($\mathbf{E}_4$)	Adam	0.161 ± 0.036	0.167 ± 0.054	($\mathbf{E}_8$)	Adam (reg.)	14.4 ± 24.6	79.907 ± 158.29
	SPBM	**0.016** ± 0.004	**0.000** ± 0.013		SPBM	0.252 ± 0.109	**0.0002** ± 0.0004
	SSL-ALM	1.380 ± 0.025	**0.000** ± 0.006		SSL-ALM	**0.243** ± 0.141	0.0014 ± 0.0013
	SSw	2.203 ± 0.014	**0.000** ± 0.020		SSw	10196.38 ± 389.67	0.3459 ± 0.0309

4.2 Results

Two observations emerge from Table 2. First, SPBM clearly outperforms SSL-ALM on ($\mathbf{E}_{4,5}$), the two multiclass experiments, which include the largest task ($m = 9900$); on ($\mathbf{E}_8$) it attains a similar loss with better constraint satisfaction; and it matches SSL-ALM on the remaining experiments. Second, SPBM also improves on the unconstrained Adam baseline in ($\mathbf{E}_4$). We conjecture that the constraints prevent the optimizer from concentrating on the easier classes, which is supported by the accuracy curves in Fig. 5 (Appendix C.2).

Fig. 1 gives more detail on robustness. A larger fraction of SPBM's configurations attains a low loss than SSL-ALM's on both fairness and PINNs tasks; SSL-ALM compensates with more feasible configurations on the fairness tasks, while SPBM leads on both counts for PINNs. SPBM has more hyperparameters than SSL-ALM, so its grid is larger: 360 vs. 100 configurations on the fairness problems and 384 vs. 360 on the PINN problems. Lastly, we observe from Table 3 that the per-epoch runtime of CML methods is within 1.3× of the per-epoch runtime of regularized Adam.

Method	m=100	m=5000	m=9900
Adam	1.74 ± 0.10	1.71 ± 0.00	1.71 ± 0.00
Adam + Reg.	1.79 ± 0.09	1.82 ± 0.03	1.84 ± 0.02
SPBM	2.20 ± 0.01	2.26 ± 0.01	2.26 ± 0.01
SSL-ALM	1.98 ± 0.01	1.99 ± 0.01	1.98 ± 0.01
SSw	1.98 ± 0.09	1.96 ± 0.03	1.96 ± 0.02

Table 3: Computational benchmark on ($\mathbf{E}_5$): per-epoch runtime (s), mean ± std over epochs and seeds. Runtime is essentially constant in the number of constraints m.

Conclusion. We have aimed at partially bridging the gap between theory and practice of constrained machine learning. We proposed and analysed SPBM, which matches or outperforms the state of the art across fairness and PINN benchmarks. We also exhibited an ERM instance on which SPBM and state-of-the-art methods fail: evidence that current assumptions are too strong for realistic CML.

Limitations. SPBM has more hyperparameters than SSL-ALM, though a comparable or higher fraction of configurations performs well. Our convergence result is partial: it gives criticality of the transformed Lagrangian but no feasibility or complementarity. What assumptions are weak enough to cover general CML yet strong enough for a KKT guarantee remains open.

5 AI USE, ETHICS AND REPRODUCIBILITY

AI USE STATEMENT

In this work, we used generative AI tools:

1. to aid or polish writing; specifically to correct grammar, to improve readability of plain text, to format tables.
2. for retrieval and discovery: specifically, to supplement our list of existing literature by finding and summarizing papers on the topic that the authors might have missed. Each paper was then read and analyzed by us.
3. to create and edit software code: parts of the experimental scripts and plotting code.
4. for research ideation and execution: in Appendix E.4: given the counterexample, AI had to come up with numbers that would be easy to calculate by hand.
5. for proving mathematical claims; specifically, given a proof, an LLM was tasked with improving readability of the proof, and/or to simplify notation, and/or format theorems and equations. AI was also used for proof brainstorming: acting as an independent reviewer / collaborator. AI gave us suggestions and feedback on correctness of our proof approaches, pointing to wrong steps, which helped us identify false claims faster.
6. for drafting sections of the paper: In Appendix E.3, AI suggested ways to merge lemmas or simplify propositions to make the text easier to follow. AI also suggested additional remarks in the convergence analysis to facilitate understanding of the larger picture of the proof. We used AI to generate our list of notation.

We did not use generative AI tools for generating synthetic datasets.

Additionally, we used generative AI tools for proposing keywords for our paper. We also used AI to identify gaps in the paper, such as incorrect grammar, ambiguous/incorrect claims.

We have reviewed all AI-assisted work; the AI generated code was tested and verified by two authors of the paper. The AI-polished lemmas / theorems / proofs were carefully reviewed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Some of the experiments presented in this paper concern the fair ML use-case. We caution the readers that no algorithmic fairness method tackles all ethical issues that arise in training machine learning models. Care must be taken that fair ML is part of an interdisciplinary pipeline that integrates the specifics of the use-case, and that it does not serve as an excuse for pursuing Business-As-Usual policies that fail to tackle ethical issues (Balayn et al., 2023; Wachter et al., 2021).

REPRODUCIBILITY STATEMENT

The datasets used in this work (FOLKTABLES, Dutch Census, CIFAR-10, CIFAR-100) are well-known and publicly available as Python packages. The source code needed to reproduce the experiments is attached as supplementary materials, including a readme file with instructions. The details of the implementation and experiments are in Appendix B. All theoretical results obtained in this paper are proved in Appendix E. We include in Appendix D necessary preliminaries to facilitate understanding of the proofs.

REFERENCES

Shoshana Abramovich and Lars-Erik Persson. Some new estimates of the 'jensen gap'. *Journal of Inequalities and Applications*, 2016(1):39, 2016. Cited on page 5.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/achiam17a.html. Cited on page 1.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization: Y. arjevani et al. *Mathematical Programming*, 199(1):165–214, 2023. Cited on page 2.

Matthias Aschenbrenner, Lou van den Dries, and Joris van der Hoeven. Asymptotic differential algebra and model theory of transseries. *arXiv preprint arXiv:1509.02588*, 2015. Cited on page 30.

Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. "✓ Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 482–495, New York, NY, USA, August 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604674. Cited on page 10.

Gilles Bareilles, Allen Gehret, Johannes Aspman, Jana Lepšová, and Jakub Mareček. Deep learning as the disciplined construction of tame objects. *arXiv preprint arXiv:2509.18025*, 2025. Cited on pages 29, 31, 32, and 41.

Amir Beck. *First-Order Methods in Optimization*. SIAM, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997. Cited on page 4.

Aharon Ben-Tal and Michael Zibulevsky. Penalty/barrier multiplier methods for convex programming problems. *SIAM Journal on Optimization*, 7(2):347–366, 1997. Cited on pages 2, 3, 4, and 38.

Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians, 2023. URL https://arxiv.org/abs/2106.13015. Cited on page 2.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30 (3):1117–1147, 2022. Cited on page 47.

Raghu Bollapragada, Cem Karamanli, Brendan Keith, Boyan Lazarov, Socratis Petrides, and Jingyi Wang. An adaptive sampling augmented lagrangian method for stochastic optimization with deterministic constraints. *Computers and Mathematics with Applications*, 149:239–258, 2023. ISSN 0898-1221. doi: https://doi.org/10.1016/j.camwa.2023.09.014. URL https://www.sciencedirect.com/science/article/pii/S0898122123003991. Cited on page 3.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021. Cited on pages 6, 29, 32, 34, 35, 43, 45, 47, 53, and 54.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007. Cited on pages 32, 42, and 47.

Jérôme Bolte, Tam Le, and Edouard Pauwels. Subgradient sampling for nonsmooth nonconvex minimization. *SIAM Journal on Optimization*, 33(4):2542–2569, 2023. Cited on page 47.

Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023. Cited on pages 1 and 2.

Jonathan Borwein and Xianfu Wang. Lipschitz functions with maximal clarke subdifferentials are generic. *Proceedings of the American Mathematical Society*, 128(11):3221–3229, 2000. Cited on page 29.

Maarten Buyl, Marybeth Defrance, and Tijl De Bie. fairret: a framework for differentiable fairness regularization terms. In *International Conference on Learning Representations*, 2024. Cited on page 1.

Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2022. Cited on page 2.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990. Cited on pages 20, 33, 34, and 46.

Christian Clason and Tuomo Valkonen. *Introduction to nonsmooth analysis and optimization*. SIAM, 2026. Cited on pages 42, 43, 50, and 53.

Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, and Maya R. Gupta. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. Cited on page 2.

Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, 205(1):431–483, May 2024a. ISSN 1436-4646. doi: 10.1007/s10107-023-01981-1. URL https://doi.org/10.1007/s10107-023-01981-1. Cited on page 2.

Frank E. Curtis, Daniel P. Robinson, and Baoyu Zhou. Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. *SIAM Journal on Optimization*, 34(4):3592–3622, 2024b. doi: 10.1137/23M1556149. URL https://doi.org/10.1137/23M1556149. Cited on page 2.

Frank E Curtis, Vyacheslav Kungurtsev, Daniel P Robinson, and Qi Wang. A stochastic-gradient-based interior-point algorithm for solving smooth bound-constrained optimization problems. *SIAM Journal on Optimization*, 35(2):1030–1059, 2025. Cited on pages 1 and 3.

Frank E Curtis, Xin Jiang, and Qi Wang. A single-loop stochastic feasible interior-point algorithm for nonlinear inequality-constrained optimization: F. curtis et al. *Mathematical Programming*, pp. 1–38, 2026. Cited on pages 1 and 3.

Aris Daniilidis and Dmitriy Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020. Cited on page 29.

Aris Daniilidis, Robert Deville, and Sebastián Tapia-García. All convex bodies are in the subdifferential of some everywhere differentiable locally lipschitz function. *Proceedings of the London Mathematical Society*, 129(5):e70007, 2024. Cited on page 29.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions, 2018. URL https://arxiv.org/abs/1804.07795. Cited on pages 6, 29, 32, 34, 35, 40, 42, 44, 45, 46, 47, 49, 51, 53, and 54.

Alberto De Marchi and Andreas Themelis. A penalty barrier framework for nonconvex constrained optimization. *Journal of Nonsmooth Analysis and Optimization*, 5(Original research articles), 2025. Cited on page 3.

A Dembo. Lecture notes for probability theory: Stat310/math230, 2019. URL https://web.stanford.edu/class/stats310a/lnotes.pdf. Cited on pages 42 and 44.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021. Cited on page 22.

Lou van den Dries, Angus Macintyre, and David Marker. The elementary theory of restricted analytic fields with exponentiation. *Ann. of Math. (2)*, 140(1):183–205, 1994. ISSN 0003-486X,1939-8980. doi: 10.2307/2118545. Cited on page 31.

Juan Elenter, Luiz Chamon, and Alejandro Ribeiro. Near-Optimal Solutions of Constrained Learning Problems. *International Conference on Learning Representations*, 2024:17426–17450, May 2024. Cited on page 2.

Herbert B Enderton. *A mathematical introduction to logic*. Elsevier, 2001. Cited on page 30.

Francisco Facchinei and Vyacheslav Kungurtsev. Stochastic approximation for expectation objective and expectation inequality-constrained nonconvex optimization, 2023. URL https://arxiv.org/abs/2307.02943. Cited on page 2.

Yuchen Fang, Sen Na, Michael W. Mahoney, and Mladen Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization*, 34(2):2007–2037, 2024. doi: 10.1137/22M1537862. URL https://doi.org/10.1137/22M1537862. Cited on page 2.

Yuchen Fang, Jihun Kim, Sen Na, James Demmel, and Javad Lavaei. A trust-region interior-point stochastic sequential quadratic programming method. *arXiv preprint arXiv:2603.10230*, 2026. Cited on pages 1 and 2.

Peter Fusek, Diethard Klatte, and Bernd Kummer. Examples and counterexamples in lipschitz analysis. *Control and Cybernetics*, 31(3):471–492, 2002. Cited on page 29.

Jose Gallego-Posada, Juan Ramirez, Meraj Hashemizadeh, and Simon Lacoste-Julien. Cooper: A Library for Constrained Optimization in Deep Learning, April 2025. Cited on page 2.

Xiang Gao, Meera Sitharam, and Adrian E Roitberg. Bounds on the jensen gap, and implications for mean-concentrated distributions. *arXiv preprint arXiv:1712.05267*, 2017. Cited on page 5.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. Cited on page 24.

Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969. Cited on pages 3 and 38.

Martin Hils and François Loeser. *A first journey through logic*, volume 89. American Mathematical Soc., 2019. Cited on page 30.

Ignacio Hounie, Alejandro Ribeiro, and Luiz F. O. Chamon. Resilient Constrained Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71767–71798. Curran Associates, Inc., 2023. doi: 10.52202/075280-3142. Cited on page 2.

Ruichuan Huang, Jiawei Zhang, and Ahmet Alacaoglu. Stochastic smoothed primal-dual algorithms for nonconvex optimization with linear inequality constraints. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26106–26142. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/huang25au.html. Cited on pages 1, 4, 7, and 23.

Yankun Huang and Qihang Lin. Oracle complexity of single-loop switching subgradient methods for non-smooth weakly convex functional constrained optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 61327–61340. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c132c02176577c4319a878f6417a331a-Paper-Conference.pdf. Cited on pages 1 and 7.

A. D. Ioffe. An invitation to tame optimization. *SIAM Journal on Optimization*, 19(4):1894–1917, 2009. doi: 10.1137/080722059. URL https://doi.org/10.1137/080722059. Cited on page 30.

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906. Cited on page 5.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020. Cited on pages 29 and 34.

Sham M Kakade and Jason D Lee. Provably correct automatic sub-differentiation for qualified programs. *Advances in neural information processing systems*, 31, 2018. Cited on pages 4 and 41.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Cited on page 7.

Clemens Kinn and Philipp Petersen. Fast approximation and learning of binary classification tasks in o-minimal structures using relu neural networks. *arXiv preprint arXiv:2607.01266*, 2026. Cited on page 29.

Andrii Kliachkin, Jana Lepšová, Gilles Bareilles, and Jakub Mareček. humancompatible.train: Implementing optimization algorithms for stochastically-constrained stochastic optimization problems. *NeurIPS Workshop on Constrained Optimization; arXiv preprint arXiv:2509.21254*, 2025. Cited on pages 2 and 23.

Andrii Kliachkin, Jana Lepšová, Gilles Bareilles, and Jakub Marecek. Benchmarking stochastic approximation algorithms for fairness-constrained training of deep neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JxmjzC6syB. Cited on page 1.

Michal Kočvara and Michael Stingl. Pennon: A code for convex nonlinear and semidefinite programming. *Optimization methods and software*, 18(3):317–333, 2003a. Cited on page 3.

Michal Kočvara and Michael Stingl. Pennon: a generalized augmented lagrangian method for semidefinite programming. In *High performance algorithms and software for nonlinear optimization*, pp. 303–321. Springer, 2003b. Cited on page 3.

Michal Kočvara and Michael Stingl. Pennon: Software for linear and nonlinear matrix inequalities. In *Handbook on semidefinite, conic and polynomial optimization*, pp. 755–791. Springer, 2012. Cited on page 3.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019. Cited on page 3.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Iteration complexity of a proximal augmented lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. *Mathematics of Operations Research*, 48(2):1066–1094, 2023. Cited on page 3.

Julian Kranz, Davide Gallon, Steffen Dereich, and Arnulf Jentzen. Sad neural networks: Divergent gradient flows and asymptotic optimality via o-minimal structures. *Advances in Neural Information Processing Systems*, 38:3930–3966, 2026. Cited on page 29.

Anastasis Kratsios, Gregory Cousins, Haitz Sáez de Ocáriz Borde, Bum Jun Kim, and Simone Brugiapaglia. Every feedforward neural network definable in an o-minimal structure has finite sample complexity. *arXiv preprint arXiv:2605.07097*, 2026. Cited on page 29.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. Cited on page 22.

Vladimír Kunc and Jiří Kléma. Three decades of activations: A comprehensive survey of 400 activation functions for neural networks, 2024. URL https://arxiv.org/abs/2402.09092. Cited on page 31.

Zhu Li, Adrián Pérez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. Kernel dependence regularizers and Gaussian processes with applications to algorithmic fairness. *Pattern Recognition*, 132: 108922, December 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.108922. Cited on page 1.

Qihang Lin, Runchao Ma, and Yangyang Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational optimization and applications*, 82(1):175–224, 2022. Cited on page 3.

Wei Liu and Yangyang Xu. A spider-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. *SIAM Journal on Optimization*, 36(2):1125–1153, 2026. Cited on page 3.

Ta Lê Loi. Verdier and strict Thom stratifications in o-minimal structures. *Illinois J. Math.*, 42(2): 347–356, 1998. ISSN 0019-2082,1945-6581. Cited on page 32.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. Cited on page 23.

Lu Lu, Raphael Pestourie, Wenjie Yao, Zhicheng Wang, Francesc Verdugo, and Steven G Johnson. Physics-informed neural networks with hard constraints for inverse design. *SIAM Journal on Scientific Computing*, 43(6):B1105–B1132, 2021. Cited on page 1.

Runchao Ma, Qihang Lin, and Tianbao Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pp. 6554–6564. PMLR, 2020. Cited on page 2.

David Marker. *Model theory: an introduction*. Springer, 2002. Cited on page 30.

Levi D. McClenny and Ulisses M. Braga-Neto. Self-adaptive physics-informed neural networks. *Journal of Computational Physics*, 474:111722, February 2023. ISSN 0021-9991. doi: 10.1016/j.jcp.2022.111722. URL http://dx.doi.org/10.1016/j.jcp.2022.111722. Cited on page 22.

Jorge J Moré and Stefan M Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009. Cited on page 8.

Sen Na, Mihai Anitescu, and Mladen Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, 199(1): 721–791, May 2023a. doi: 10.1007/s10107-022-01846-z. URL https://doi.org/10.1007/s10107-022-01846-z. Cited on page 2.

Sen Na, Mihai Anitescu, and Mladen Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming, 2023b. URL https://arxiv.org/abs/2109.11502. Cited on page 2.

Ion Necoara and Angelia Nedić. Minibatch stochastic subgradient-based projection algorithms for feasibility problems with convex inequalities. *Computational Optimization and Applications*, 80 (1):121–152, 2021. Cited on pages 1 and 2.

Ion Necoara and Nitesh Kumar Singh. Stochastic subgradient for composite convex optimization with functional constraints. *Journal of Machine Learning Research*, 23(265):1–35, 2022. Cited on pages 1 and 2.

Aleš Nekvinda and Luděk Zajíček. A simple proof of the Rademacher theorem. *Časopis pro pěstování matematiky*, 113(4):337–341, 1988. ISSN 0528-2195. Cited on page 33.

Figen Oztoprak, Richard Byrd, and Jorge Nocedal. Constrained optimization in the presence of noise. *SIAM Journal on Optimization*, 33(3):2118–2136, 2023. doi: 10.1137/21M1450999. URL https://doi.org/10.1137/21M1450999. Cited on page 2.

Anand Pillay and Charles Steinhorn. Definable sets in ordered structures. 1984. Cited on page 29.

Michael JD Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pp. 283–298, 1969. Cited on pages 3 and 38.

Wenqiang Pu, Kaizhao Sun, and Jiawei Zhang. Smoothed proximal lagrangian method for nonlinear constrained programs. *arXiv preprint arXiv:2408.15047*, 2024. Cited on page 3.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017. Cited on page 7.

Juan Ramirez, Meraj Hashemizadeh, and Simon Lacoste-Julien. Position: Adopt Constraints Over Penalties in Deep Learning, July 2025. Cited on page 1.

Rodolfo Rios-Zertuche. Examples of pathological dynamics of the subgradient method for lipschitz path-differentiable functions. *Mathematics of Operations Research*, 47(4):3184–3206, 2022. Cited on page 29.

R Tyrrell Rockafellar. A dual approach to solving nonlinear programming problems by unconstrained optimization. *Mathematical programming*, 5(1):354–373, 1973a. Cited on page 3.

R Tyrrell Rockafellar. The multiplier method of hestenes and powell applied to convex programming. *Journal of Optimization Theory and applications*, 12(6):555–562, 1973b. Cited on page 3.

R Tyrrell Rockafellar. Augmented lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, 12(2):268–285, 1974. Cited on page 3.

R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976. Cited on page 3.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. Cited on page 33.

Qiankun Shi, Xiao Wang, and Hao Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Optimization Online*, 2022. URL https://optimization-online.org/?p=19870. Cited on page 2.

Nitesh Kumar Singh and Ion Necoara. Stochastic halfspace approximation method for convex optimization with nonsmooth functional constraints. *IEEE Transactions on Automatic Control*, 2024. Cited on pages 1 and 2.

Nitesh Kumar Singh, Ion Necoara, and Vyacheslav Kungurtsev. Mini-batch stochastic subgradient for functional constrained optimization. *Optimization*, 73(7):2159–2185, 2024. Cited on pages 1 and 2.

Hwijae Son, Sung Woong Cho, and Hyung Ju Hwang. Enhanced physics-informed neural networks with augmented lagrangian relaxation method (al-pinns). *Neurocomputing*, 548:126424, 2023. Cited on pages 1, 7, and 23.

Yongcun Song, Xiaoming Yuan, and Hangrui Yue. The admm-pinns algorithmic framework for nonsmooth pde-constrained optimization: a deep learning approach. *SIAM Journal on Scientific Computing*, 46(6):C659–C687, 2024. Cited on page 1.

Patrick Speissegger. The Pfaffian closure of an o-minimal structure. *J. Reine Angew. Math.*, 508: 189–211, 1999. ISSN 0075-4102,1435-5345. doi: 10.1515/crll.1999.026. URL https://doi.org/10.1515/crll.1999.026. Cited on page 31.

LI Strukov and Aleksandr Filippovich Timan. Mathematical expectation of continuous functions of random variables. smoothness and variance. *Siberian Mathematical Journal*, 18(3):469–474, 1977. Cited on page 5.

Alfred Tarski. *A Decision Method for Elementary Algebra and Geometry*. The Rand Corporation, Santa Monica, CA, 1948. Cited on page 31.

Lou van den Dries. Remarks on tarski's problem concerning (r,+,*, exp). In *Studies in Logic and the Foundations of Mathematics*, volume 112, pp. 97–121. Elsevier, 1984. Cited on page 29.

Lou van den Dries. *Tame Topology and O-minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1998. ISBN 9780521598385. Cited on pages 30, 32, 41, and 43.

Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. 1996. Cited on page 35.

Paul Van der Laan. *The 2001 Census in the Netherlands: Integration of Registers and Surveys*, pp. 39–52. 12 2001. Cited on page 22.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, January 2021. Cited on page 10.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021. Cited on pages 7 and 22.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022. Cited on page 7.

A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.*, 9(4):1051–1094, 1996. ISSN 0894-0347,1088-6834. doi: 10.1090/S0894-0347-96-00216-0. Cited on page 31.

Hong-Kun Xu. Viscosity approximation methods for nonexpansive mappings. *Journal of Mathematical Analysis and Applications*, 298(1):279–291, 2004. Cited on pages 36 and 49.

Ming Yang, Gang Li, Quanqi Hu, Qihang Lin, and Tianbao Yang. Single-loop algorithms for stochastic non-convex optimization with weakly-convex constraints. *arXiv preprint arXiv:2504.15243*, 2025. Cited on page 3.

Jiawei Zhang and Zhi-Quan Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020. Cited on pages 3 and 4.

Jiawei Zhang, Wenqiang Pu, and Zhi-Quan Luo. On the iteration complexity of smoothed proximal alm for nonconvex optimization problem with convex constraints. *arXiv preprint arXiv:2207.06304*, 2022. Cited on page 3.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020. Cited on page 2.

Supplementary material

OUTLINE OF THE APPENDIX

A List of Notation **19**

B Experimental Setup and Implementation details **22**

B.1 Data and Experiments 22

B.2 Implementation 23

B.3 Hyperparameter grid search 24

B.4 Hyperparameter selection 25

C Remaining Results **26**

C.1 Ablation Study - SPBM 26

C.2 Convergence Plots 27

C.3 Computational Benchmark 28

D Preliminaries **29**

D.1 Tame Theory 29

D.2 Minkowski sums 32

D.3 The Clarke subgradient 33

D.4 Conservative Fields 34

D.5 Weak convexity and an auxiliary convergence lemma. 36

E Theoretical results **37**

E.1 Bias Introduced by Penalty/Barrier Functions 37

E.2 SPBM Reduces to SSL-ALM-like Primal-Dual Update 38

E.3 Convergence Analysis 38

E.4 Counter Example - No subgradient-type method escapes the flat region. 55

A List of Notation

Probability, data and sampling

$(\Omega, \mathcal{F}, P)$	Probability space carrying the mini-batch draws
Ω	Sample space; realised canonically as $(\{1, \dots, M\}^B)^{\mathbb{N}_0}$
P	Probability measure on Ω; the product measure $\bigotimes_{k\geq 0} \mathrm{Unif}(\{1, \dots, M\}^B)$, i.e. its factors are uniform
$\mathcal{F}^k$	σ-algebra generated by the batches drawn before iteration k, $\sigma(\mathbf{j}^0, \dots, \mathbf{j}^{k-1})$
Ω_0	Event of full measure on which the convergence analysis is carried out
$\mathbb{E}$	Expectation operator; $\mathbb{E}[\cdot \mid \mathcal{F}^k]$ the conditional expectation given $\mathcal{F}^k$
ξ	Random variable; a single data sample
Ξ	Distribution of ξ, i.e. a probability measure on the data space;
$\operatorname{supp} \Xi$	Support of Ξ; the dataset $\{\xi^{(1)}, \dots, \xi^{(M)}\}$ in case of no repetitions
$\xi^{(l)}$	l-th element of the dataset, $l \in \{1, \dots, M\}$
M	Number of samples in the dataset
B	Mini-batch size
ξ_b	b-th sample of a mini-batch, $b \in \{1, \dots, B\}$; the batch is written $\xi_{1:B}$
$\mathbf{j}$	Multi-index encoding a mini-batch, $\mathbf{j} = (j_1, \dots, j_B) \in \{1, \dots, M\}^B$; the b-th sample of the batch is $\xi^{(j_b)}$
$\mathbf{j}^k$	Mini-batch index drawn at iteration k; $P(\mathbf{j}^k = \mathbf{j}) = M^{-B}$

Problem, algorithm and iterates

n	Dimension of the primal variable x
m	Number of constraints
$f : \mathbb{R}^n \to \mathbb{R}$	Objective function
$g : \mathbb{R}^n \to \mathbb{R}^m$	Constraint function
$\bar{f}(\cdot\,; \mathbf{j}),\ \bar{g}_i(\cdot\,; \mathbf{j})$	Mini-batch estimators on batch $\mathbf{j}$, e.g. $\bar{g}_i(x; \mathbf{j}) = \frac{1}{B}\sum_{b=1}^{B} g_i(x, \xi^{(j_b)})$
φ	Penalty/barrier function, $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$
$\overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j})$	Mini-batched Lagrangian
$\mathcal{L}_p^{\mathrm{emp}}(\cdot, \lambda)$	Expectation of Lagrangian, $M^{-B}\sum_{\mathbf{j}} \overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j})$
x^k	Primal iterate at iteration k
$\lambda^k,\ \lambda_i^k$	Dual variables; λ_i^k is associated with constraint i at iteration k
p^k	Penalty parameters at iteration k
s^k	Prox-centre at iteration k

$\bar{y}^k,\ y^k,\ \zeta^k$	Stochastic search direction, its conditional mean $\mathbb{E}[\bar{y}^k \mid \mathcal{F}^k]$, and the noise $\bar{y}^k - y^k$
α^k	Primal step size; square-summable but not summable
$\gamma^k,\ \kappa^k$	Dual and penalty averaging coefficients; $\gamma^k, \kappa^k \to 1$
μ	Weight of the Moreau/proximal term
δ	Prox-centre averaging coefficient, $\delta \in [0, 1)$
ρ	Penalty parameter of the ALM updating rule
Λ	Upper bound of the dual box, $\lambda^k \in [0, \Lambda]^m$
$\ell_c,\ u_c$	Lower and upper penalty clipping bounds, $p^k \in [\ell_c, u_c]^m$
$\hat{\pi}$	Penalty-updating function, $\hat{\pi} \in \{\hat{\pi}_\rho^{\mathrm{ALM}}, \hat{\pi}^{\mathrm{ADAPT}}\}$
$\odot,\ \oslash$	Coordinate-wise (Hadamard) product and division

Convergence analysis

$\mathcal{X}$	Bounded primal set; a closed ball with a bounded radius R
$\hat{\lambda},\ \hat{p}$	Almost-sure limits of λ^k and p^k
$\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$	Conservative field of the potential $-\mathcal{L}_{\hat{p}}^{\mathrm{emp}}(\cdot, \hat{\lambda})$
G	Restriction of $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$ to $\mathcal{X}$
$\Phi_{\hat{p}}(\cdot, \hat{\lambda})$	Lyapunov function $\max\{\mathcal{L}_{\hat{p}}^{\mathrm{emp}}(\cdot, \hat{\lambda}),\ c\}$
$\partial f(x)$	Clarke subdifferential of f at x as in Clarke (1990); see Definition 3
$\mathrm{Lip}_A(f)$	Lipschitz constant of $f\mid_A$
$\mathrm{dist}(x, A)$	Distance from x to the set A
$\mathcal{O}$	Fixed o-minimal structure containing $\mathbb{R}_{\mathrm{exp}}$
$\mathbb{R}_{\mathrm{alg}},\ \mathbb{R}_{\mathrm{exp}}$	Semi-algebraic and exponential o-minimal structures, $\mathbb{R}_{\mathrm{alg}} \subseteq \mathbb{R}_{\mathrm{exp}} \subseteq \mathcal{O}$
$F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$	Set-valued map

General notation

$\mathbb{R}^n$	Euclidean space of dimension n
$\mathbb{R}_+$	Set of non-negative real numbers, i.e., $\{x \in \mathbb{R} \mid x \geq 0\}$
$\mathbb{R}_{++}$	Set of strictly positive real numbers, i.e., $\{x \in \mathbb{R} \mid x > 0\}$
$\mathbb{N},\ \mathbb{N}_0$	Positive integers $\{1, 2, \dots\}$ and non-negative integers $\{0, 1, 2, \dots\}$
$\mathbb{B}$	Closed unit ball
$[a, b]$	Closed interval
(a, b)	Open interval
$f\mid_{\mathcal{X}} \colon \mathcal{X} \to \mathbb{R}$	Restriction of f to a domain $\mathcal{X} \subseteq \mathbb{R}^n$
$\mathrm{Proj}_{\mathcal{X}}$	Projection operator onto the set $\mathcal{X}$
$[t]_+$	Positive part, $[t]_+ := \max\{t, 0\} = \mathrm{Proj}_{\mathbb{R}_+}(t)$, applied coordinate-wise

prox_f	Proximal operator of a function f
$\mathcal{R}$	Set of groups in the fairness tasks; $\mathcal{R}_i$ denotes the i-th group
$\mathbf{1}_A$	Indicator function of a set A
$\nabla f(x)$	Gradient of f at x

B EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

B.1 DATA AND EXPERIMENTS

Datasets and problem tasks. Our fairness-based experiments are based on three datasets. For the ACSIncome dataset Ding et al. (2021), we use a subset corresponding to the state of Florida, which includes 98,925 samples, and consider the binary classification task of predicting whether an individual's income exceeds $50,000. For the Dutch Census dataset Van der Laan (2001), which represents aggregated groups of people in the Netherlands and includes 60,000 samples, the binary classification task is to predict whether an individual has a high-level or low-level profession. The CIFAR-10 dataset Krizhevsky et al. (2009) contains 60,000 32x32 color images in 10 different classes, whereas the CIFAR-100 dataset contains 100 different classes with 600 images per class. For both datasets, the task is to learn an accurate classifier of the classes. PINNs are data efficient, allowing us to sample as many data as needed for our three experiments. We sample 2601 training data points.The task in PINNs experiments is to approximate a solution to a given differential equation.

B.1.1 PDES USED FOR PINNS EXPERIMENTS

The *Helmholtz equation* reads:

$$\begin{aligned} &\Delta u + u = q(z_1, z_2), && \text{for every } (z_1, z_2) \in [-1, 1]^2, \\ &u(z_1, z_2) = 0, && \text{for every } (z_1, z_2) \in \mathrm{Bd}[-1, 1]^2, \end{aligned}$$

where the function q is chosen as

$$\begin{aligned} q(z_1, z_2) = &-\pi^2 \sin(\pi z_1) \sin(4\pi z_2) \\ &- (4\pi)^2 \sin(\pi z_1) \sin(4\pi z_2) + \sin(\pi z_1) \sin(4\pi z_2). \end{aligned}$$

The *viscous Burgers equation* reads:

$$\partial_t u + \partial_z(\tfrac{1}{2}u^2 - c\partial_z u) = 0, \qquad \text{for } (t, z) \in [0, 1] \times [-1, 1], \tag{22}$$

$$u(0, z) = -\sin(\pi z), \qquad \text{for } z \in [-1, 1], \tag{23}$$

$$u(t, -1) = u(t, 1) = 0, \qquad \text{for } t \in [0, 1], \tag{24}$$

for $(t, z) \in [0, 1] \times [-1, 1]$, and $c = \frac{0.01}{\pi}$. This setting is the same as in McClenny & Braga-Neto (2023).

The *Klein–Gordon equation* we consider is:

$$\partial_t^2 u - \partial_z^2 u + u^3 = f(t, z), \qquad \text{for } (t, z) \in [0, 1] \times [0, 1], \tag{25}$$

$$u(0, z) = g_1(z), \qquad \text{for } z \in [0, 1], \tag{26}$$

$$\partial_t u(0, z) = g_2(z), \qquad \text{for } z \in [0, 1], \tag{27}$$

$$u(t, z) = h(t, z), \qquad \text{for } (t, z) \in [0, 1] \times \{0, 1\}, \tag{28}$$

where f, g_1, g_2, h are computed using a fabricated solution, as in Wang et al. (2021):

$$u(t, z) = z \cos(5\pi t) + (tz)^3.$$

B.2 IMPLEMENTATION

Implementation Details

1. *(SSL-ALM, SSw implementation)* The implementation of SSw and SSL-ALM are based on the repository Kliachkin et al. (2025).
2. *(PINN implementation)* We base our implementation of the PINN experiments **($\mathbf{E}_6$)**, **($\mathbf{E}_7$)**, **($\mathbf{E}_8$)** on the code repository of Son et al. (2023).
3. *(Primal update step used)* The SPBM algorithm uses Adam as the `PrimalUpdate` function specified in Algorithm 1.
4. *(SSL-ALM adaptation to our setting)* SSL-ALM Huang et al. (2025) is designed for equality constraints; we adapt it to our inequality-constrained setting by projecting the dual variables onto the nonnegative orthant.
5. *(Loss functions)* Specifically, the loss function of **($\mathbf{E}_1$)–($\mathbf{E}_5$)**, for a neural network u_{nn} with weights x, and one-hot label Target,

$$f(x,\xi_j) = -\sum_{i\in\text{Classes}} \mathbf{1}[\text{Target}(i,j)=1]\cdot\log\left(\frac{\exp\big(u_{\text{nn}}(x,\xi_j)[i]\big)}{\sum_{c\in\text{Classes}}\exp\big(u_{\text{nn}}(x,\xi_j)[c]\big)}\right).$$

   And the loss function of **($\mathbf{E}_6$)–($\mathbf{E}_8$)** is:

$$f(x,\xi_j) = \big(Nu_{\text{nn}}(x,\xi_j) - h(\xi_j)\big)^2.$$

6. *(Schedules)* To schedule primal learning rate (α^k), we use linear warmup over the first 5% of training steps, followed by a cosine-annealing decay of the learning rate for the remainder Loshchilov & Hutter (2016). We do not implement scheduling neither for the dual learning rate in SSL-ALM nor for the dual update parameters γ, κ, ρ in SPBM.
7. *(Note on objective and constraint sampling)* To ensure that we are able to compute constraints at each iteration, we construct minibatches by taking $\frac{B}{|\mathcal{R}|}$ samples from each group independently. This approach produces an unbiased estimate of the constraints as defined in Eq. (17); however, this introduces a bias into the objective, as it is defined as the expectation over the entire population. To obtain an unbiased estimator, one could scale each group's samples according to the group's size with respect to the size of the entire dataset. Assume equally-sized batches per demographic groups b_i such that $|b_i| = B/|\mathcal{R}|$ is the number of R_i-samples per batch, and $P(R_i) = |R_i|/M$, then define the estimator

$$\begin{aligned}\overline{f}(x) &= \sum_{R_i\in\mathcal{R}}\sum_{l\in b_i}\frac{|R_i|}{M}\cdot\frac{1}{|b_i|}f(x,\xi_{i,l})\\ &= \sum_{R_i\in\mathcal{R}} P(R_i)\cdot\frac{1}{|b_i|}\sum_{l\in b_i} f(x,\xi_{i,l}).\end{aligned}$$

   By the law of total expectation,

$$\mathbb{E}\big[\bar{f}(x)\big] = \sum_{R_i\in\mathcal{R}} P(R_i)\mathbb{E}_{\xi\sim\text{Unif}(R_i)}[f(x,\xi)] = \mathbb{E}_\xi[f(x,\xi)],$$

   hence $\bar{f}(x)$ is an unbiased estimator of the population objective. Note that when the size of each subgroup is equally large, the estimator $\bar{f}$ coincides with a plain batch average and reweighting has no contribution. We find the size of each subgroup in each of our experiments roughly equally large – specifically, for CIFAR-10 and CIFAR-100, the size of the groups is exactly the same; for Dutch, the largest group has 3 times the samples of the smallest one (4393 vs. 1488); for Folktables, the largest group has 5.2 more samples than the smallest one (28697 vs. 5488). We choose to omit the reweighting of the samples in experiments **($\mathbf{E}_2$)** and **($\mathbf{E}_3$)** (based on the Folktables and Dutch datasets, respectively) due to the relative simplicity of the task. Therefore, we do *not* implement the reweighting.

Architectures and epochs. To get a reliable estimate of the methods' performance, we conduct five 40-epoch runs with five different seeds per each experiment for **($\mathbf{E}_1$)–($\mathbf{E}_5$)**; for **($\mathbf{E}_6$)–($\mathbf{E}_8$)**, we conduct five 6000-epoch runs with different seeds. For experiment **($\mathbf{E}_1$)**, we use an MLP with two hidden layers of size 64 and 32. For experiment **($\mathbf{E}_2$)** we use an MLP with 2 hidden layers with sizes 64 and

16, and with sizes 128 and 64 for **($\mathbf{E}_3$)**. The experiment **($\mathbf{E}_4$)** is benchmarked using a convolutional neural network with two convolutional layers of respective depths 6 and 16 and a window size of 5, each followed by a 2×2 max-pooling layer, and three fully-connected layers. The experiment **($\mathbf{E}_5$)** is benchmarked using the ResNet-18 architecture He et al. (2016). For experiments **($\mathbf{E}_6$)–($\mathbf{E}_8$)**, we use a deep narrow MLP with 8 hidden layers of size 64.

Data split and processing. In all experiments, we use the entire dataset for training and hyperparameter selection, omitting a train/test/validation split. We believe that this allows us to fairly compare SPBM against the state-of-the-art purely on optimization performance. In experiments **($\mathbf{E}_1$)–($\mathbf{E}_3$)**, we preprocess the data to have zero mean and unit variance; for CIFAR experiments **($\mathbf{E}_{4,5}$)** we rescale the data to $[-1, 1]$ using default PyTorch tools for CIFAR data.

Mini-batching The mini-batch size varied by problem: for experiments **($\mathbf{E}_1$)–($\mathbf{E}_3$)** the mini-batch sizes ranged from 32 to 72 samples, whereas for **($\mathbf{E}_4$)** and **($\mathbf{E}_5$)** the mini-batch sizes were 120 and 400, respectively. For experiments **($\mathbf{E}_6$)–($\mathbf{E}_8$)** the mini-batch sizes were set to 1000 samples.

Numerical setup. We ran experiments **($\mathbf{E}_1$)–($\mathbf{E}_4$)**, **($\mathbf{E}_6$)–($\mathbf{E}_8$)** on a single Tesla V100 GPU; we ran **($\mathbf{E}_5$)** on a single Tesla H200 141GB NVLink (reference will be added for a camera-ready version).

Choice of hyperparameters. To determine the hyperparameters, we performed a grid search for each algorithm averaging the results over five runs to account for random initialization; the process in more detail is described in Section 4. The hyperparameter grids are listed in Appendix B.3, and the final hyperparameters are listed in Appendix B.4.

B.3 Hyperparameter grid search

In all experiments, we use the PyTorch implementation of the Adam optimizer for the minimization of the objective (Adam, SSw) and of the Lagrangian (SSL-ALM, SPBM).

Fairness experiments (($\mathbf{E}_1$)–($\mathbf{E}_5$)). The per-method grids are:

- **Adam:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$.
- **SSw:** objective and constraint learning rate, each in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$, and Moreau regularization multiplier μ in $\{0, 1, 2\}$.
- **SSL-ALM:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$, dual learning rate in $\{0.01, 0.02, 0.05, 0.1, 0.5\}$, penalty multiplier $\rho \in \{0, 1\}$, and Moreau regularization multiplier in $\{1, 2\}$.
- **SPBM:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$, dual EMA parameter γ in $\{0.9, 0.99, 0.999\}$, penalty-barrier function in $\{\varphi_{QL}, \varphi_{QR}\}$, penalty updating function $\hat{\pi}$ in $\{\hat{\pi}_{\mathrm{ALM}}, \hat{\pi}_{\mathrm{ADAPT}}\}$ (11),(12) and corresponding penalty update multipliers $\kappa \in \{0.1, 0.5, 0.9\}$ ($\hat{\pi}_{\mathrm{ADAPT}}$) and ρ in $\{0.5, 1.0, 2.0\}$ ($\hat{\pi}_{\mathrm{ALM}}$), and μ in $\{1, 2\}$.

PINN experiments (($\mathbf{E}_6$)–($\mathbf{E}_8$)). Because PINN training differs substantially from classification in scale and conditioning (long horizons, few minibatches per epoch, stiff residual losses), we use a separate per-regime grid for each method.

The per-method grids are:

- **Adam:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$, and regularization multiplier β on the initial/boundary-condition loss terms in $\{0.5, 1, 2, 5, 10\}$.
- **SSw:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$ and switching (constraint) learning rate in $\{0.0005, 0.001, 0.005, 0.01, 0.05\}$; no Moreau regularization ($\mu = 0$).
- **SSL-ALM:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02, 0.05\}$, dual learning rate in $\{0.0005, 0.001, 0.005, 0.008, 0.01, 0.05\}$, penalty multiplier $\rho \in \{0, 0.1, 1, 2\}$, and Moreau regularization multiplier μ in $\{0, 1, 2\}$.
- **SPBM:** objective learning rate in $\{0.001, 0.005, 0.01, 0.02\}$, dual EMA parameter γ in $\{0, 0.1, 0.9, 0.999\}$, penalty-barrier function in $\{\varphi_{QL}, \varphi_{QR}\}$, penalty updating function $\hat{\pi}$ in $\{\hat{\pi}_{\mathrm{ALM}}, \hat{\pi}_{\mathrm{ADAPT}}\}$ and corresponding penalty update multipliers $\kappa \in \{0.1, 0.9, 0.99\}$ ($\hat{\pi}_{\mathrm{ADAPT}}$) and ρ in $\{0.1, 1.0, 2.0\}$ ($\hat{\pi}_{\mathrm{ALM}}$), and μ in $\{0, 2\}$.

B.4 HYPERPARAMETER SELECTION

In this section, we list the final selected hyperparameters per experiment. Additionally, in all experiments the Quadratic-Logarithmic penalty-barrier function φ_{QL} showed better performance as part of the grid, except for **($\mathbf{E}_3$)**, where the Quadratic-Reciprocal penalty barrier function φ_{QR} won out. We mention this here to streamline the tables. LR stands for the learning rate.

Exp.	Method	LR	Dual LR or γ	Penalty (ALM: ρ)	Penalty update (SPBM: κ)	μ
($\mathbf{E}_1$)	Adam	0.001	–	–	–	–
	SSw	0.001	0.01	–	–	1.0
	SSL-ALM	0.001	0.5	0.0	–	2.0
	SPBM	0.001	0.9	–	$\hat{\pi}_{\text{ADAPT}}$, $\kappa = 0.9$	2.0
($\mathbf{E}_2$)	Adam	0.001	–	–	–	–
	SSw	0.001	0.01	–	–	0.0
	SSL-ALM	0.005	0.01	0.0	–	2.0
	SPBM	0.005	0.9	–	$\hat{\pi}_{\text{ADAPT}}$, $\kappa = 0.9$	2.0
($\mathbf{E}_3$)	Adam	0.005	–	–	–	–
	SSw	0.05	0.001	–	–	0.0
	SSL-ALM	0.02	0.01	0.0	–	1.0
	SPBM	0.02	0.99	–	$\hat{\pi}_{\text{ADAPT}}$, $\kappa = 0.5$	1.0
($\mathbf{E}_4$)	Adam	0.001	–	–	–	–
	SSw	0.001	0.001	–	–	0.0
	SSL-ALM	0.01	0.01	1.0	–	2.0
	SPBM	0.005	0.9	–	$\hat{\pi}_{\text{ADAPT}}$, $\kappa = 0.9$	2.0
($\mathbf{E}_5$)	Adam	0.01	–	–	–	–
	SSw	0.001	0.001	–	–	1.0
	SSL-ALM	0.01	0.01	1.0	–	1.0
	SPBM	0.005	0.999	–	$\hat{\pi}_{\text{ADAPT}}$, $\kappa = 0.9$	1.0

Exp.	Method	LR	Dual LR or γ	Penalty (Adam (reg.): β, SSL-ALM: ρ)	Penalty update (SPBM: κ or ρ)	μ
($\mathbf{E}_6$)	Adam (reg.)	0.005	–	10.0	–	–
	SSw	0.01	0.01	–	–	0.0
	SSL-ALM	0.01	0.05	0.1	–	2.0
	SPBM	0.01	0.1	–	$\hat{\pi}_{\text{ALM}}$, $\rho = 2.0$	0.0
($\mathbf{E}_7$)	Adam (reg.)	0.05	–	10.0	–	–
	SSw	0.005	0.0005	–	–	0.0
	SSL-ALM	0.005	0.01	2.0	–	0.0
	SPBM	0.005	0.9	–	$\hat{\pi}_{\text{ADAPT}}$, $\kappa = 0.9$	0.0
($\mathbf{E}_8$)	Adam (reg.)	0.001	–	0.5	–	–
	SSw	0.001	0.01	–	–	0.0
	SSL-ALM	0.005	0.01	0.0	–	0.0
	SPBM	0.005	0.0	–	$\hat{\pi}_{\text{ALM}}$, $\rho = 2.0$	2.0

C Remaining Results

C.1 Ablation Study - SPBM

We perform an ablation study of SPBM to determine the impact of each component of the method on its performance. We identify γ, κ, and μ as parameters of interest, and run experiments analogous to **($\mathbf{E}_1$)**–**($\mathbf{E}_8$)** while setting each to 0. The results for fairness experiments are shown in Table 4, and the results for PINN experiments are shown in Table 5.

Table 4: Results of an ablation study of SPBM on the fairness benchmark. As earlier, we pick the hyperparameter configuration that produces the lowest training loss while being feasible within 10% tolerance; if no such configurations exist, we pick one with the lowest constraint violation.

Exp.	Method	Best loss	Max constraint viol.	Mean constraint
($\mathbf{E}_1$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	0.373 ± 0.001	**0.000** ± 0.002	1.402 ± 0.036
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	**0.372** ± 0.001	0.001 ± 0.003	1.433 ± 0.024
	SPBM ($\gamma = 0$)	**0.372** ± 0.001	**0.000** ± 0.001	1.425 ± 0.035
	SPBM ($\kappa = 0$)	0.374 ± 0.001	**0.000** ± 0.002	1.399 ± 0.041
	SPBM ($\mu = 0$)	0.376 ± 0.002	**0.000** ± 0.001	1.430 ± 0.068
($\mathbf{E}_2$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	0.359 ± 0.001	**0.000** ± 0.002	0.000 ± 0.002
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	**0.350** ± 0.003	0.002 ± 0.007	0.000 ± 0.005
	SPBM ($\gamma = 0$)	0.386 ± 0.004	**0.000** ± 0.009	0.000 ± 0.006
	SPBM ($\kappa = 0$)	<u>0.356</u> ± 0.002	**0.000** ± 0.005	0.000 ± 0.003
	SPBM ($\mu = 0$)	0.366 ± 0.003	0.004 ± 0.004	0.000 ± 0.003
($\mathbf{E}_3$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	0.487 ± 0.009	0.001 ± 0.020	0.000 ± 0.010
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	**0.467** ± 0.003	0.001 ± 0.004	0.000 ± 0.006
	SPBM ($\gamma = 0$)	0.487 ± 0.009	0.001 ± 0.020	0.000 ± 0.010
	SPBM ($\kappa = 0$)	<u>0.475</u> ± 0.002	**0.000** ± 0.006	0.000 ± 0.004
	SPBM ($\mu = 0$)	**0.467** ± 0.003	0.001 ± 0.004	0.000 ± 0.006
($\mathbf{E}_4$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	0.750 ± 0.041	**0.000** ± 0.011	0.000 ± 0.014
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	**0.016** ± 0.004	**0.000** ± 0.013	0.000 ± 0.007
	SPBM ($\gamma = 0$)	2.117 ± 0.106	**0.000** ± 0.024	0.000 ± 0.025
	SPBM ($\kappa = 0$)	<u>0.187</u> ± 0.017	**0.000** ± 0.018	0.000 ± 0.015
	SPBM ($\mu = 0$)	0.235 ± 0.012	**0.000** ± 0.052	0.000 ± 0.028
($\mathbf{E}_5$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	4.487 ± 0.041	**0.000** ± 0.029	0.000 ± 0.018
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	**0.001** ± 0.000	**0.000** ± 0.002	0.000 ± 0.000
	SPBM ($\gamma = 0$)	4.604 ± 0.001	**0.000** ± 0.042	0.000 ± 0.013
	SPBM ($\kappa = 0$)	0.017 ± 0.006	**0.000** ± 0.012	0.000 ± 0.005
	SPBM ($\mu = 0$)	<u>0.002</u> ± 0.000	**0.000** ± 0.002	0.000 ± 0.000

Table 5: Results of an ablation study of SPBM on the PINN benchmark. As earlier, we pick the hyperparameter configuration that produces the lowest training loss while being feasible within 10% tolerance; if no such configurations exist, we pick one with the lowest constraint violation.

Exp.	Method	Best loss	Max constraint viol.	Mean constraint
($\mathbf{E}_6$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	**4010.463** $\pm$ 3274.490	**0.0000** $\pm$ 0.0002	0.0001 $\pm$ 0.0002
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	6680.601 $\pm$ 7.048	**0.0000** $\pm$ 0.0002	0.0001 $\pm$ 0.0002
	SPBM ($\gamma = 0$)	5347.308 $\pm$ 2673.634	**0.0000** $\pm$ 0.0002	0.0001 $\pm$ 0.0002
	SPBM ($\kappa = 0$)	6684.120 $\pm$ 0.008	**0.0000** $\pm$ 0.0000	0.0000 $\pm$ 0.0000
	SPBM ($\mu = 0$)	**4010.463** $\pm$ 3274.490	**0.0000** $\pm$ 0.0002	0.0001 $\pm$ 0.0002
($\mathbf{E}_7$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	**0.000** $\pm$ 0.000	**0.0000** $\pm$ 0.0000	0.0000 $\pm$ 0.0000
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	**0.000** $\pm$ 0.000	**0.0000** $\pm$ 0.0000	0.0000 $\pm$ 0.0000
	SPBM ($\gamma = 0$)	**0.000** $\pm$ 0.000	**0.0000** $\pm$ 0.0000	0.0000 $\pm$ 0.0000
	SPBM ($\kappa = 0$)	0.004 $\pm$ 0.005	0.0001 $\pm$ 0.0001	0.0001 $\pm$ 0.0001
	SPBM ($\mu = 0$)	**0.000** $\pm$ 0.000	**0.0000** $\pm$ 0.0000	0.0000 $\pm$ 0.0000
($\mathbf{E}_8$)	SPBM ($\hat{\pi}_\rho^{ALM}$)	**0.252** $\pm$ 0.109	**0.0003** $\pm$ 0.0004	0.0002 $\pm$ 0.0002
	SPBM ($\hat{\pi}_\rho^{ADAPT}$)	1815.751 $\pm$ 3577.152	0.0607 $\pm$ 0.1055	0.0352 $\pm$ 0.0605
	SPBM ($\gamma = 0$)	**0.252** $\pm$ 0.109	**0.0003** $\pm$ 0.0004	0.0002 $\pm$ 0.0002
	SPBM ($\kappa = 0$)	4163.698 $\pm$ 5079.520	0.1199 $\pm$ 0.1423	0.0573 $\pm$ 0.0691
	SPBM ($\mu = 0$)	3.079 $\pm$ 2.051	0.0018 $\pm$ 0.0008	0.0009 $\pm$ 0.0005

C.2 Convergence Plots

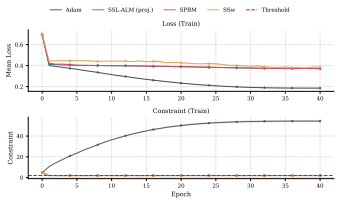


Figure 2: ($\mathbf{E}_1$): Convergence plot.

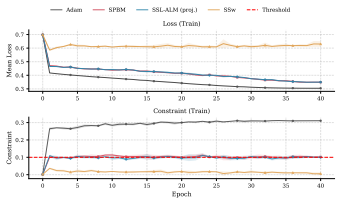


Figure 3: ($\mathbf{E}_2$): Convergence plot.

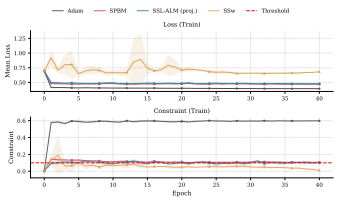


Figure 4: ($\mathbf{E}_3$): Convergence plot.

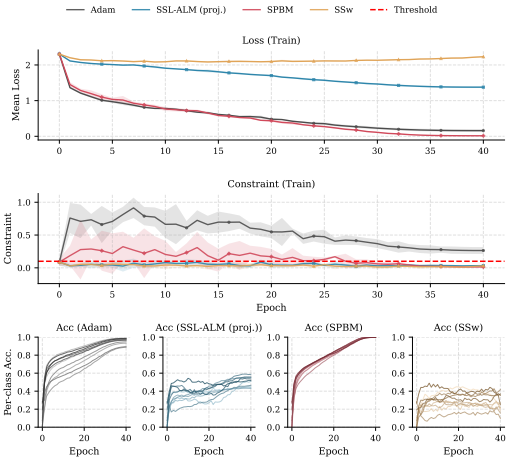


Figure 5: ($\mathbf{E}_4$): Convergence plot. Bottom row shows per-class accuracy convergence (i.e. 10 accuracy curves per subplot).

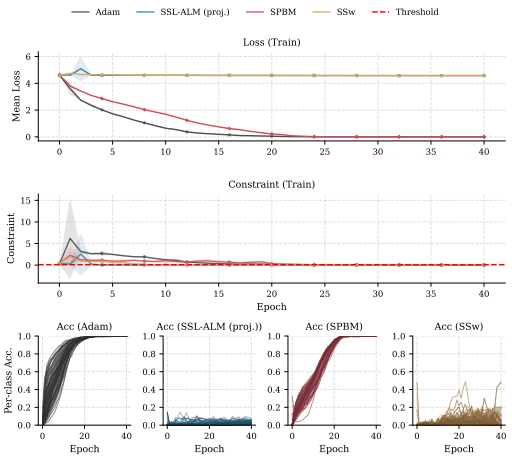


Figure 6: ($\mathbf{E}_5$): Convergence plot. Bottom row shows per-class accuracy convergence (i.e. 100 accuracy curves per subplot).

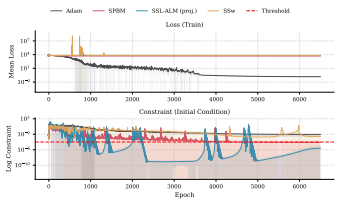


Figure 7: ($\mathbf{E}_6$): Convergence plot.

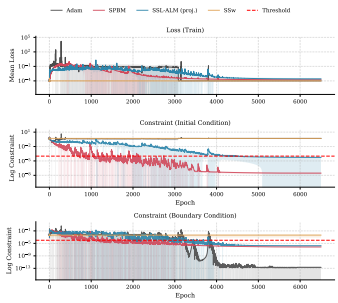


Figure 8: ($\mathbf{E}_7$): Convergence plot.

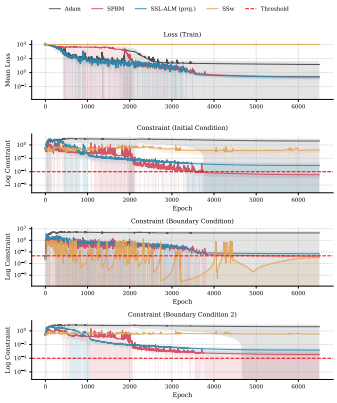


Figure 9: ($\mathbf{E}_8$): Convergence plot.

C.3 Computational Benchmark

We provide a full table of the computational benchmark in Table 6. The shortened version of the table was presented in Table 3.

Table 6: Computational benchmark based on ($\mathbf{E}_5$): per-epoch runtime (s), mean $\pm$ std over epochs and init seeds.

# constraints (m)	Adam	Adam + Reg.	SPBM	SSL-ALM (proj.)	SSw
100	1.74 ± 0.10	1.79 ± 0.09	2.20 ± 0.01	1.98 ± 0.01	1.98 ± 0.09
1000	1.72 ± 0.01	1.84 ± 0.04	2.23 ± 0.01	1.99 ± 0.01	1.95 ± 0.02
2000	1.71 ± 0.01	1.79 ± 0.02	2.23 ± 0.01	1.98 ± 0.01	1.95 ± 0.02
3000	1.71 ± 0.01	1.81 ± 0.02	2.25 ± 0.01	1.98 ± 0.01	1.96 ± 0.02
4000	1.72 ± 0.01	1.81 ± 0.03	2.26 ± 0.01	1.98 ± 0.01	1.96 ± 0.02
5000	1.71 ± 0.00	1.82 ± 0.03	2.26 ± 0.01	1.99 ± 0.01	1.96 ± 0.03
6000	1.71 ± 0.01	1.85 ± 0.03	2.26 ± 0.01	1.99 ± 0.01	1.97 ± 0.05
7000	1.72 ± 0.01	1.82 ± 0.03	2.25 ± 0.01	1.99 ± 0.01	1.96 ± 0.02
8000	1.73 ± 0.04	1.83 ± 0.05	2.26 ± 0.01	1.98 ± 0.01	1.96 ± 0.03
9000	1.72 ± 0.01	1.81 ± 0.01	2.26 ± 0.01	1.99 ± 0.01	1.96 ± 0.02
9900	1.71 ± 0.00	1.84 ± 0.02	2.26 ± 0.01	1.98 ± 0.01	1.96 ± 0.02

D PRELIMINARIES

In this section we give an overview of some of the mathematical ingredients used in the theoretical results in section E below. We do not aim to give here an exhaustive account of all prerequisites used; instead we focus on some topics at the interface of *o-minimality* and *nonsmooth nonconvex analysis/optimization*, especially as it relates to the convergence results below.

Section D.1 introduces o-minimality and explains its relevance to our setting. Sections D.2-D.4 recall Minkowski sums, Clarke subgradients, and conservative fields, leading to the chain-rule and critical-value properties used in the convergence analysis. Section D.5 records the definition of weak convexity and an auxiliary sequence convergence lemma.

D.1 TAME THEORY

From a certain birds-eye point of view, the primary objects of study here are objects that live in Euclidean space; e.g., subsets of $\mathbb{R}^n$, functions $\mathbb{R}^n \to \mathbb{R}^m$, and set-valued maps $\mathbb{R}^n \rightrightarrows \mathbb{R}^m$. Indeed, the main convergence result Theorem 3 below is ultimately a statement about the expected performance of the algorithm SPBM given the fixed locally Lipschitz functions $\bar{f}(\cdot;\mathbf{j})$ and $\bar{g}_i(\cdot;\mathbf{j})$. From this perspective, the following two comments are in order:

1. Completely arbitrary locally Lipschitz functions and their corresponding subgradient dynamics can be highly pathological. For example, see Borwein & Wang (2000); Fusek et al. (2002); Daniilidis & Drusvyatskiy (2020); Rios-Zertuche (2022); Daniilidis et al. (2024). Consequently, we wouldn't expect Theorem 3 to be true in this most general setting, although we make no claims in this regard.
2. We are content with establishing Theorem 3 for a very broad class of functions which actually show up in practice. In particular, the functions $\bar{f}$ and $\bar{g}_i$ we have in mind are not arbitrary functions, but instead are functions (NNs) built as a large composition of familiar pieces: affine functions, activation functions such as ReLU, GeLU, and Mish, common loss functions, etc.

To address these, we adopt the framework of *o-minimality*, also called *tame geometry*. This framework arrived in 1984 in van den Dries (1984) and Pillay & Steinhorn (1984) as a means of extending semialgebraic geometry to broader settings (including a setting containing the exponential function).

Below our primary reference for connections between o-minimality and deep learning is the survey Bareilles et al. (2025); although these ideas are also discussed in, e.g., Davis et al. (2018); Ji & Telgarsky (2020); Bolte & Pauwels (2021); Kranz et al. (2026); Kinn & Petersen (2026); Kratsios et al. (2026).

Definition 1. *A **structure** on $\mathbb{R}$ is a sequence $\mathcal{O} = (\mathcal{O}_n)_{n\geq 0}$ such that for each $n \geq 0$:*

(S1) *$\mathcal{O}_n$ is a boolean algebra of subsets of $\mathbb{R}^n$;*
(S2) *if $A \in \mathcal{O}_n$, then $\mathbb{R} \times A$ and $A \times \mathbb{R}$ belong to $\mathcal{O}_{n+1}$;*
(S3) *$\{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 = x_n\} \in \mathcal{O}_n$;*
(S4) *if $A \in \mathcal{O}_{n+1}$, then $\pi(A) \in \mathcal{O}_n$, where $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the projection map on the first n coordinates.*

*A structure $\mathcal{O}$ is **o-minimal** if it satisfies:*

(O1) *$\{(x, y) \in \mathbb{R}^2 : x < y\} \in \mathcal{O}_2$;*
(O2) *the sets in $\mathcal{O}_1$ are exactly the finite unions of intervals and points.*

Finally, we will always additionally assume that all structures satisfy:

(S5) *$\{a\} \in \mathcal{O}_1$ for every $a \in \mathbb{R}$;*
(S6) *$\mathcal{O}_3$ contains the graphs of addition and multiplication $+, \cdot : \mathbb{R}^2 \to \mathbb{R}$.*

Given a structure $\mathcal{O}$ (o-minimal or otherwise) and a set $A \subseteq \mathbb{R}^n$, we say that A is definable *(in $\mathcal{O}$) if $A \in \mathcal{O}_n$. Likewise, given a function $f : X \to \mathbb{R}^m$ with $X \subseteq \mathbb{R}^n$, we say that the function f is* definable *if its graph $\Gamma(f) \subseteq \mathbb{R}^{n+m}$ is a definable set; a similar convention holds also for set-valued maps $\mathbb{R}^n \rightrightarrows \mathbb{R}^m$.*

The rest of D.1 is organized around following three ideas:

1. Definability in a structure is preserved by many relevant mathematical constructions (D.1.1)

2. Many (most?) functions which are built from standard NN primitives are definable in some known o-minimal structure (D.1.2)
3. The o-minimality axiom (O2) has powerful consequences for definable sets, including higher-dimensional definable sets (D.1.3)

D.1.1 Closure under common constructions

One of the advantages of assuming our Euclidean objects are definable in a common structure $\mathcal{O}$ (as we do in Assumption 2 below for $\tilde{f}$ and $\bar{g}_i$) is that many related mathematical constructions from the objects will also be definable in $\mathcal{O}$. To give some basic examples of what we mean by that, we record the following facts from (van den Dries, 1998, 1.§2): Suppose $\mathcal{O}$ is a structure, then:

- if functions $f : X \to Y$ and $g : Y \to Z$ are definable, then so is the composition $g \circ f : X \to Z$;
- if $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$ are definable, then so is the cartesian product $A \times B$;
- if $f : \mathbb{R} \to \mathbb{R}$ is definable, then so is the set $\mathcal{R}_f := \{x \in \mathbb{R} : f \text{ is differentiable at } x\}$, as well as the function $x \mapsto f'(x) : \mathcal{R}_f \to \mathbb{R}$.

More generally, we quote from Ioffe (2009):

> *...the class [of definable sets and functions] is closed with respect to practically all operations used in variational analysis...*

Why do *structures* support such a large amount of mathematical constructions? The short answer is that the S-axioms of Definition 1 are expressive enough to permit a vast arsenal of mathematical constructions from their judicious application. A more complete answer is that Definition 1 encodes the geometric/semantic content of the concept *definability in first-order logic*; namely, a set is definable if and only if there is a valid *first-order formula* which defines it. A first-order formula is a syntactic object built in a disciplined way from variables $x, y, z, \ldots$, primitive relations and functions $(=, +, \cdot, \ldots)$, logical connectives $\wedge, \vee, \neg, \ldots$, and first-order quantifiers $\forall x, \exists x$ (where x may only range over $\mathbb{R}$).

A rigorous definition and overview of this connection is outside the scope of this paper; our main reference is (Aschenbrenner et al., 2015, Appendix B); see also standard texts in first-order logic and model theory, e.g., Enderton (2001); Hils & Loeser (2019); Marker (2002).

Regardless, we can already see how the axioms relate to their syntactic counterparts, e.g.,

- Axiom (S1) reflects the fact that we have the basic logical connectives $\wedge, \vee, \neg, \ldots$ at our disposal.
- Axioms (S3) and (S6) give us the basic primitives $=, +, \cdot$.
- Axiom (S4) expresses that we may do existential quantification $\exists x_{n+1}$ over the last variable; in fact, quantification over any of the variables can be done, although this requires a small argument.

Here is a typical example of the usage of first-order formulas:

Example 1. *Suppose $\mathcal{O}$ is an arbitrary structure satisfying (S1)-(S4), (S5)-(S6). Then $\mathcal{O}$ also satisfies (O1). Indeed, we may define the binary relation $<$ with the following first-order formula:*

$$\varphi(x, y) \;:=\; \exists z\big[\neg(x = y) \wedge x + z^2 = y\big]$$

We claim that φ is a valid first-order formula which defines $<$, which thus establishes that axiom (O1) holds for $\mathcal{O}$. Implicit here is the claim that the theory of first-order definability permits us to "compile down" the formula into an argument which only uses the assumed axioms (S1)-(S4), (S5)-(S6).

Thus in practice, in order to identify a construction as being definable, one usually needs to simply write the definition of the construction as a valid first-order formula. This requires learning the careful rules of *first-order definability* and some practice, but otherwise it is usually a routine matter. The constructions we are most concerned with come from topology, differential calculus, and variational analysis; for this, the following insight usually suffices:

- if a definition involves quantifying over infinite sequences x_k, an equivalent ε-δ-style definition should be used instead since first-order quantifiers $\forall x, \exists x$ cannot quantify over sequences.

Specifically, we make claims that definability is preserved by a given construction in Lemmas 1, 5, 12, and 13, and Proposition 6; the goal of the preceding discussion was to give some intuition as to *why* such claims ought to be true.

We conclude by emphasizing that nothing in the preceding discussion in D.1.1 required or used the o-minimality axiom (O2).

D.1.2 Definability of functions used in practice

We now briefly survey three o-minimal structures which are perhaps most relevant to the NN examples we have in mind:

$$\mathbb{R}_{\text{alg}} \subseteq \mathbb{R}_{\text{exp}} \subseteq \mathbb{R}_{\text{Pfaff}}$$

Example 2 ($\mathbb{R}_{\text{alg}}$ - the semialgebraic sets)**.** *The structure $\mathbb{R}_{\text{alg}}$ can be defined as the smallest structure satisfying axioms (S1)-(S4) and (S5)-(S6). This structure is easily seen to also satisfy (O1) and it is essentially a theorem of Tarski (1948) that this structure is o-minimal, i.e., satisfies (O2). The definable sets of $\mathbb{R}_{\text{alg}}$ are those which can be defined by boolean combinations of sets of the form $\{f(x) \diamond 0\}$, where $f(x)$ is a multivariable polynomial and $\diamond$ is either $=$ or $<$. Consequently, any NN built as a composition of affine layers and semialgebraic activations (such as* ReLU *and* Softsign*) will be semialgebraic; likewise, the barrier function φ_{QR} is also semialgebraic.*

As it turns out, no restriction of the real-valued functions $\exp$ or $\log$ will be semialgebraic. For this we turn to:

Example 3 ($\mathbb{R}_{\text{exp}}$ - the subexponential sets)**.** *The structure $\mathbb{R}_{\text{exp}}$ can be defined as the smallest structure expanding $\mathbb{R}_{\text{alg}}$ and containing the (graph of) the real exponential function $\exp : \mathbb{R} \to \mathbb{R}$. The o-minimality of $\mathbb{R}_{\text{exp}}$ was established by Wilkie (1996). In $\mathbb{R}_{\text{exp}}$, the barrier function φ_{QL} becomes definable, as does any NN which now uses activations involving $\exp$ and $\log$ (e.g., $\tanh$,* Logistic*,* Swish*, and* Softplus*).*

One notable drawback of $\mathbb{R}_{\text{exp}}$ is that the activation function $\mathsf{GELU}(x) = \frac{x}{2}\big(1 + \text{erf}(x/\sqrt{2})\big)$ is not definable; this is because the *error function* $\text{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2}\, ds$ is not definable. For this we turn to:

Example 4 ($\mathbb{R}_{\text{Pfaff}}$ - the Pfaffian closure of $\mathbb{R}_{\text{alg}}$)**.** *This is a further expansion of $\mathbb{R}_{\text{exp}}$ whose o-minimality was shown in Speissegger (1999). We forego the definition here; for our purposes it suffices to know that some additional activation functions such as $\arctan$ and* GELU *become available here.*

We also have the following main obstruction to being not-definable in an o-minimal structure:

Non-example 1 (Infinite oscillation)**.** *If an activation function exhibits infinite oscillation, then it cannot be definable in any o-minimal structure. For example, if the sine function $\sin : \mathbb{R} \to \mathbb{R}$ were definable, then so would the set $\{x : \sin(x) = 0\}$ of zeros; although this clearly contradicts the o-minimality axiom (O2). Even in this case, if the sine function is only being evaluated on a fixed compact set (e.g., $[0, 2\pi]$), then the restriction would still be definable in the o-minimal structures $\mathbb{R}_{\text{an}} \subseteq \mathbb{R}_{\text{an,exp}}$ (see e.g., Dries et al. (1994) for definitions of these structures).*

In conclusion, as a rule of thumb, if a function is constructed from familiar constructions alluded to in D.1.1 above, using familiar pieces which do not exhibit infinite oscillation, then most likely that function will be definable in some known o-minimal structure. This is the basis for our belief that *definable in an o-minimal structure* is an appropriate hypothesis for our theoretical results. We make no general claims about the ubiquity of this hypothesis in practice, although as a heuristic we point to (Bareilles et al., 2025, Remark 3.19) where it is estimated that around 89% of the activation functions listed in the survey Kunc & Kléma (2024) are definable in some known o-minimal structure.

D.1.3 Consequences of the o-minimality axiom

To avoid giving the wrong impression, we emphasize the following:

> *Although it might be trivial to recognize that a given object is definable in a known o-minimal structure $\mathcal{O}$; the proof of o-minimality of $\mathcal{O}$ is non-trivial.*

In this subsection we briefly address the question:

> *What can be gained from knowing an object is definable in an o-minimal structure?*

The short answer is that an o-minimal structure gives rise to a coherent world of Euclidean objects free from undesirable pathologies (e.g., infinite oscillation, or worse): where all definable objects exhibit tame asymptotic and finiteness properties, and the world is closed under all first-order mathematical constructions.

As a simple example, we mention that *non-convergence* is impossible in the following sense:

- if $f : [0, \infty) \to \mathbb{R}^n$ is a bounded curve definable in an o-minimal structure, then $\lim_{t\to\infty} f(t)$ always exists (cf. (van den Dries, 1998, 3.§1)).

For "higher dimensional" definable sets in $\mathbb{R}^n$ ($n \geq 1$), we have the following:

- Every definable subset $X \subseteq \mathbb{R}^n$ can be assigned a unique and well-defined *dimension* $\dim X$, where *dimension* satisfies many reasonable properties of a *dimension function* (van den Dries, 1998, 4.§1), (Bareilles et al., 2025, 3.34); this dimension agrees with the usual notions of dimension where applicable, e.g., linear dimension and manifold dimension.
- For every $r \geq 1$, every definable subset $X \subseteq \mathbb{R}^n$ can be partitioned into finitely many definable C^r-submanifolds of $\mathbb{R}^n$. There are various results of this form, most notably *smooth cell decomposition* (van den Dries, 1998, 7.§3) and various *stratification theorems* (e.g., Whitney-(a)/(b), Verdier); see Loi (1998).

For our purposes, we point out that the stratification results enable Bolte et al. (2007) to prove the *projection formula* for the set-valued *Clarke subgradient*, and this was key for Davis et al. (2018) to prove the convergence of the *stochastic subgradient method* (SSM). This was later extended to the set-valued notion of *conservative field* in Bolte & Pauwels (2021), partially to justify convergence even though standard AD implementations such as PyTorch and TensorFlow need not always return an element of the Clarke subgradient.

Our convergence proof (Theorem 3) follows this same playbook. Namely, we first define an enveloping subgradient-like set-valued map (see Lemma 13):

$$\mathcal{D}_p(\cdot, \lambda) \ : \ \mathbb{R}^n \rightrightarrows \mathbb{R}^n.$$

Next, using the tame stratification machinery outlined above, we show that this map is in fact a *definable conservative field* in the sense of Bolte & Pauwels (2021) (cf. Lemmas 8 and 13). This permits us to adapt the main stochastic approximation / differential inclusion convergence arguments of Davis et al. (2018) and Bolte & Pauwels (2021); after verifying a descent property and a Weak Sard property (via Theorem 2), we can conclude Theorem 3.

We conclude by emphasizing that everything in the preceding discussion in D.1.3 assumed the o-minimality axiom (O2). When dropping this axiom, cracks appear in the world just described.

D.2 Minkowski sums

Minkowski sums provide the notation for combining set-valued subgradients, in particular when averaging over mini-batches in Section E.3.

Definition 2 (Minkowski Sum)**.** *The Minkowski sum of sets $A, B \subseteq \mathbb{R}^n$ is defined as*

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

Constructing the Minkowski sum preserves definability:

Lemma 1. *If $A, B \subseteq \mathbb{R}^n$ are definable in a structure, then so is the Minkowski sum $A + B$.*

Proof sketch. Let $\varphi_A(x)$, $\varphi_B(y)$ be first-order formulas which define A and B in $\mathcal{O}$. Then the following formula defines $A + B$ in $\mathcal{O}$:

$$\varphi_{A+B}(z) \ := \ \exists x \exists y \big[\varphi_A(x) \ \wedge \ \varphi_B(y) \ \wedge \ x + y = z\big]$$

We emphasize that this lemma and the argument do not depend on the o-minimality axiom. □

We also need the following two facts about Minkowski sums:

Lemma 2. *Minkowski sum preserves inclusion, if $A_j \subseteq B_j \subseteq \mathbb{R}^n$ for $j = 1, \dots, N$, then*

$$\sum_{j=1}^{N} A_j \subseteq \sum_{j=1}^{N} B_j.$$

Proof. This is an instance of the following basic fact about functions: given a function $f : X \to Y$ and an inclusion $A \subseteq B \subseteq X$ of subsets of the domain, we have an inclusion $f[A] \subseteq f[B]$ of direct images in the codomain. In this case, the function $f := \sum_{j=1}^{N} : \mathbb{R}^{nN} \to \mathbb{R}^n$ is multilinear vector addition, and the domain inclusion is $A := A_1 \times \cdots \times A_N \subseteq B := B_1 \times \cdots \times B_N$. □

Lemma 3. *Let $C \subseteq \mathbb{R}^n$ be convex, $N \geq 1$, and $c_1, \dots, c_N \geq 0$. Then $\sum_{j=1}^{N} c_j C = \big(\sum_{j=1}^{N} c_j\big) C$. In particular, $\frac{1}{N} \sum_{j=1}^{N} K\,\mathbb{B} = K\,\mathbb{B}$ for $K \geq 0$.*

Proof. The inclusion $\big(\sum_j c_j\big) C \subseteq \sum_j c_j C$ holds for any set, since $\big(\sum_j c_j\big) u = \sum_j c_j u$ for $u \in C$. Conversely, let $v = \sum_j c_j u_j$ with $u_j \in C$ and put $s := \sum_j c_j$. If $s = 0$ then $v = 0$ and the claim is trivial; otherwise $v = s \sum_j (c_j/s) u_j$, and $\sum_j (c_j/s) u_j \in C$ because the coefficients c_j/s are nonnegative and sum to one and C is convex. □

D.3 The Clarke subgradient

In this subsection we fix a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$.

An arbitrary locally Lipschitz function need not have a well-defined gradient at every point. However, Rademacher's theorem (see Nekvinda & Zajíček (1988) for the statement with simple proof) guarantees that the set:

$$\mathcal{R}_f := \{x \in \mathbb{R}^n : f \text{ is differentiable at } x\} \subseteq \mathbb{R}^n$$

has full measure, i.e., the complement $\mathbb{R}^n \setminus \mathcal{R}_f$ has Lebesgue measure zero. In particular, at a given $x \in \mathbb{R}^n$, the gradient of f exists at almost every point near x; this motivates considering the set-valued *Clarke subgradient* $\partial f(x)$ of f at x since the actual gradient $\nabla f(x)$ might not actually exist:

Definition 3 (Clarke subgradient)**.** *The* Clarke subgradient *$\partial f(x)$ of f at x is the set:*

$$\partial f(x) = \operatorname{conv}\Big\{ v \in \mathbb{R}^n,\ \exists\, x^k \underset{k\to\infty}{\to} x \text{ with } x^k \in \mathcal{R}_f,\ v_k = \nabla f(x_k) \underset{k\to\infty}{\to} v \Big\} \subseteq \mathbb{R}^n$$

Thus we may construe the Clarke subgradient of f as a set-valued map $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

We next recall two continuity notions for set-valued maps that will be used when passing to limits in subgradient inclusions.

Definition 4 (Outer-Semicontinuity by Rockafellar & Wets (2009))**.** *A set-valued map $G : \mathcal{X} \to \mathbb{R}^n$ is outer-semicontinuous at a point $x \in \mathcal{X}$ if for any sequence $x^i \to x$ and $v^i \in G(x^i)$ converging to a vector $v \in \mathbb{R}^n$, it holds $v \in G(x)$.*

Outer semicontinuity on $\mathcal{X}$ is equivalent to the graph of G being closed in $\mathcal{X} \times \mathbb{R}^n$.

Definition 5 (Upper-Semicontinuity by Clarke (1990))**.** *Let $G : X \to Y$ be a set-valued map; when X, Y are Banach spaces, the upper-semicontinuity property at x is defined as: for all $\varepsilon > 0$, there exists $\eta > 0$ such that*

$$G(x') \subseteq G(x) + \varepsilon \mathbb{B}_Y, \quad \forall x' \in x + \eta \mathbb{B}_X.$$

Here are the basic properties of the Clarke subgradient:

Lemma 4. *∂f has nonempty compact convex values and is locally bounded, outer semicontinuous, and upper semicontinuous. More precisely, if f is L-Lipschitz on an open set $U \subseteq \mathbb{R}^n$, then*

$$\partial f(x) \subseteq L\mathbb{B} \quad \textit{for every } x \in U,$$

where $\mathbb{B}$ is the closed Euclidean unit ball. These conclusions require no definability assumption.

Proof. Nonemptiness, compactness, convexity, and the stated bound follow from (Clarke, 1990, Proposition 2.1.2(a)); applying the bound on a Lipschitz neighbourhood of each point gives local boundedness. Outer semicontinuity and uppersemicontinuity follow from (Clarke, 1990, Proposition 2.1.5(b) and (d)), respectively. □

Lemma 5. *If $f : \mathbb{R}^n \to \mathbb{R}$ is definable in a structure $\mathcal{O}$, then so is its Clarke subgradient ∂f.*

Proof idea. This is a routine exercise in *first-order definability*; see (Ji & Telgarsky, 2020, Lemma B.8) for an argument. We also emphasize here that the lemma and its proof do not depend on the o-minimality axiom (O2). □

When f (hence also ∂f) is definable in an *o-minimal* structure, then we can say much more about the properties of ∂f (cf. Lemma 8 below); for this it is convenient to first discuss the more general notion of *conservative fields*.

D.4 Conservative Fields

A conservative field of a function f is a set-valued mapping which generalizes the notion of derivates (such as Clarke subgradients Clarke (1990)) introduced in Bolte & Pauwels (2021). From the practical point of view, as stated in Bolte & Pauwels (2021), Automatic Differentiation (AD) is proven to yield conservative fields, making the idea of conservative fields a very useful tool.

Definition 6 (Conservative Fields, Definition 1 Bolte & Pauwels (2021))**.** *Let $\mathcal{D} : \mathbb{R}^n \to \mathbb{R}^n$ be a set-valued map. $\mathcal{D}$ is a conservative (set-valued) field whenever it is outer-semicontinuous, has nonempty compact values, and for any absolutely continuous loop $x : [0,1] \to \mathbb{R}^n$ such that $x(0) = x(1)$, we have*

$$\int_0^1 \max_{v \in \mathcal{D}(x(t))} \langle \dot{x}(t), v \rangle dt = 0$$

or equivalently

$$\int_0^1 \min_{v \in \mathcal{D}(x(t))} \langle \dot{x}(t), v \rangle dt = 0.$$

Conservative fields determine an underlying function (unique up to a constant), which Bolte & Pauwels (2021) call a *potential function*.

Definition 7 (Definition 2 Bolte & Pauwels (2021))**.** *Let $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a conservative field. For any absolutely continous curve x with $x(0) = 0, x(1) = s$, a function Φ defined as*

$$\Phi(s) - \Phi(0) = \int_0^1 \langle \dot{x}(t), \mathcal{D}(x(t)) \rangle dt$$

is called a potential function for $\mathcal{D}$, or $\mathcal{D}$ is a conservative field for Φ.

Remark 1. *There can be many conservative fields for a given potential function, whereas, a conservative field determines a unique potential function up to a constant.*

Also, if a function is a potential function of some conservative field, that function is called *path differentiable*; this terminology is motivated by the following:

Lemma 6 (Lemma 2 Bolte & Pauwels (2021))**.** *Let $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a locally bounded, outer-semicontinuous set valued map and $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. Then $\mathcal{D}$ is a conservative field for Φ, if and only if for any absolutely continous curve $x : [0,1] \to \mathbb{R}^n$, the function $t \mapsto \Phi(x(t))$ satisfies for almost all $t \in [0,1]$*

$$\frac{d}{dt}\Phi(x(t)) = \langle v, \dot{x}(t) \rangle \quad \forall v \in \mathcal{D}(x(t)).$$

The following lemma is the key step in the proof of the descent property. It is a restatement of (Davis et al., 2018, Lemma 5.2) for an arbitrary set-valued map G admitting a chain rule for f, rather than for the Clarke subdifferential; the argument is theirs.

Lemma 7. *Let $\mathcal{X} \subseteq \mathbb{R}^n$, let $G : \mathcal{X} \rightrightarrows \mathbb{R}^n$ be a set-valued map with nonempty values, and let $f : \mathbb{R}^n \to \mathbb{R}$. Let $z : \mathbb{R}_+ \to \mathcal{X}$ be absolutely continuous and suppose that for a.e. $t \geq 0$*

$$\dot{z}(t) \in G(z(t)) \qquad \textit{and} \qquad \tfrac{d}{dt} f(z(t)) = \langle v, \dot{z}(t) \rangle \quad \textit{for all } v \in G(z(t)).$$

Then for a.e. $t \geq 0$,

$$\|\dot{z}(t)\| = \operatorname{dist}\big(0, G(z(t))\big) \qquad \textit{and} \qquad \tfrac{d}{dt} f(z(t)) = \|\dot{z}(t)\|^2 = \operatorname{dist}^2\big(0, G(z(t))\big).$$

Proof. Fix $t \geq 0$ at which $\dot{z}(t)$ exists and both hypotheses hold, and write $S := G(z(t))$, $y := \dot{z}(t)$, $W := \operatorname{span}(S - S)$.

Taking $v = y \in S$ in the chain rule gives $\frac{d}{dt} f(z(t)) = \langle y, y \rangle = \|y\|^2$.

Since the chain rule assigns the same value $\langle v, y \rangle$ to every $v \in S$, we get $\langle s - s', y \rangle = 0$ for all $s, s' \in S$, hence $\langle w, y \rangle = 0$ for every $w \in W$; that is, $y \in W^\perp$.

Consequently, for every $w \in W$,

$$\|y + w\|^2 = \|y\|^2 + 2\langle y, w \rangle + \|w\|^2 = \|y\|^2 + \|w\|^2 \;\geq\; \|y\|^2,$$

and since $y \in y + W$ we conclude $\|y\| = \operatorname{dist}(0, y + W)$.

Moreover $y \in S$, so $s - y \in S - y \subseteq W$ for every $s \in S$, i.e. $S \subseteq y + W$. For a distance function, it holds that

$$\operatorname{dist}(0, S) \;\geq\; \operatorname{dist}(0, y + W) \;=\; \|y\| = \|\dot{z}(t)\|.$$

The reverse inequality $\operatorname{dist}(0, G(z(t))) \leq \|\dot{z}(t)\|$ is immediate from $\dot{z}(t) \in G(z(t))$. Combining the two gives the claim. □

For a locally Lipschitz potential Φ, Rademacher's theorem gives almost-everywhere differentiability; conservativity further ensures that $D(x) = \{\nabla \Phi(x)\}$ Lebesgue almost everywhere (Bolte & Pauwels, 2021, Theorem 1). Moreover, the Clarke subgradient is contained in the convex hull of any conservative field for Φ (Bolte & Pauwels, 2021, Corollary 1):

$$\partial \Phi(x) \subseteq \operatorname{conv}(\mathcal{D}(x)). \tag{29}$$

And as it turns out, any function that is locally Lipschitz and definable in some o-minimal structure, is also path differentiable.

Lemma 8. *Let $\Phi : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function and definable in some o-minimal structure, then all three hold:*

1. *Φ admits a chain rule for Clarke subdifferential,*
2. *$\partial \Phi$ is a conservative field,*
3. *Φ path differentiable.*

Proof. By Van den Dries & Miller (1996), a definable function admits Whitney C^1-stratification. And so by [Theorem 5.8.]Davis et al. (2018), we get that a locally Lipschitz function that is Whitney C^1-stratifiable, admits a chain rule for Clarke subdifferential. Finally, by [Corollary 2]Bolte & Pauwels (2021), and by the fact that f has a chain rule for a Clarke subdifferential, then Clarke of Φ is a conservative field and Φ is path differentiable, which finishes the proof. □

We state some other important properties in the following theorem and corollary.

Theorem 2 (Theorem 5, Bolte & Pauwels (2021))**.** *Let $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a conservative field for $\Phi : \mathbb{R}^n \to \mathbb{R}$ and assume that both Φ and $\mathcal{D}$ are definable in some o-minimal structure $\mathcal{O}$. Then the set of $\mathcal{D}$-critical values*

$$\big\{\, \Phi(x) \;:\; x \in \mathbb{R}^n,\; 0 \in \mathcal{D}(x) \,\big\}$$

is finite.

Corollary 1 (Corollary 4, Bolte & Pauwels (2021))**.** *Let $\Phi_1, \ldots, \Phi_N : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz functions and let $\mathcal{D}_i : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a conservative field for Φ_i, $i = 1, \ldots, N$. Then $\sum_{i=1}^N \mathcal{D}_i$ is a conservative field for $\sum_{i=1}^N \Phi_i$.*

D.5 WEAK CONVEXITY AND AN AUXILIARY CONVERGENCE LEMMA.

Definition 8 (Weak Convexity). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is C-weakly convex for $C \geq 0$, if the function $x \mapsto f(x) + \frac{C}{2}\|x\|^2$ is convex. It is* weakly convex *if it is C-weakly convex for some C.*

The following sequence lemma will be used in Lemma 18 to show that the proximal center asymptotically tracks the primal iterates.

Lemma 9 (Lemma 2.1 Xu (2004)). *Let $\{a^k\}$ be a sequence of non-negative real numbers such that*

$$a^{k+1} \leq (1 - \gamma^k)\, a^k + \delta^k,$$

where $\gamma^k \in (0, 1)$, and suppose

(a) $\sum_{k=0}^{\infty} \gamma^k = \infty$, *and*

(b) either $\limsup_{k\to\infty} \frac{\delta^k}{\gamma^k} \leq 0$ *or* $\sum_k |\delta^k| < \infty$.

Then $a^k \to 0$.

E THEORETICAL RESULTS

E.1 BIAS INTRODUCED BY PENALTY/BARRIER FUNCTIONS

Let us recall and prove Proposition 1.

Proposition 1 (Transformed-feasibility gives original-feasibility)**.** *Consider a penalty/barrier function $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$ (8), a positive integer $B \in \mathbb{N}$, and random variables $\xi_1, \dots, \xi_B \overset{iid}{\sim} \Xi$. Consider the feasible set of the original problem (1) and of the transformed problem (13), defined as*

$$\mathcal{S} := \bigcap_{i=1}^{m} \big\{x \in \mathbb{R}^n : \mathbb{E}_{\xi}\big[g_i(x,\xi)\big] \le 0\big\}, \quad \mathcal{S}^B_{\mathrm{transform}} := \bigcap_{i=1}^{m} \big\{x \in \mathbb{R}^n : \mathbb{E}_{\xi_{1:B}}\big[p_i \varphi\big(\bar{g}_i(x)/p_i\big)\big] \le 0\big\}.$$

Then, $\mathcal{S}^B_{\mathrm{transform}} \subseteq \mathcal{S}$.

Proof of Proposition 1. Let $\hat{x} \in \mathcal{S}^B_{\mathrm{transform}}$ and fix an arbitrary $i \in \{1, \dots, m\}$. By feasibility,

$$\mathbb{E}_{\xi_{1:B}}\big[p_i \varphi\big(\bar{g}_i(\hat{x})/p_i\big)\big] \le 0.$$

Since $p_i > 0$, dividing by p_i gives

$$\mathbb{E}_{\xi_{1:B}}\big[\varphi\big(\bar{g}_i(\hat{x})/p_i\big)\big] \le 0,$$

and Jensen's inequality yields

$$\varphi\big(\mathbb{E}_{\xi_{1:B}}\big[\bar{g}_i(\hat{x})\big]/p_i\big) \le \mathbb{E}_{\xi_{1:B}}\big[\varphi\big(\bar{g}_i(\hat{x})/p_i\big)\big] \le 0.$$

By the definition of $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$, we have $\varphi(t) \le 0$ if and only if $t \le 0$; hence $\mathbb{E}_{\xi_{1:B}}\big[\bar{g}_i(\hat{x})\big]/p_i \le 0$, and since $p_i > 0$, also $\mathbb{E}_{\xi_{1:B}}\big[\bar{g}_i(\hat{x})\big] \le 0$. Finally, $\bar{g}_i(\hat{x})$ is an unbiased estimator of the constraint, i.e., $\mathbb{E}_{\xi_{1:B}}\big[\bar{g}_i(\hat{x})\big] = \mathbb{E}_{\xi}\big[g_i(\hat{x},\xi)\big]$, so $\mathbb{E}_{\xi}\big[g_i(\hat{x},\xi)\big] = g_i(\hat{x}) \le 0$. As i was arbitrary, $\hat{x} \in \mathcal{S}$. □

In the next proof, we use the fact that the two particular penalty/barrier functions have the following second derivatives:

$$\varphi''_{\mathrm{QL}}(t) = \begin{cases} 1 & \text{if } t \ge -\frac{1}{2}, \\ \frac{1}{4t^2} & \text{if } t < -\frac{1}{2}, \end{cases} \quad \varphi''_{\mathrm{QR}}(t) = \begin{cases} 1 & \text{if } t \ge -\frac{1}{3}, \\ \frac{64}{27(1-t)^3} & \text{if } t < -\frac{1}{3}. \end{cases} \tag{30}$$

For the particular penalty/barrier functions defined in Eq. (8), the bounded curvature of φ yields an explicit bound on the Jensen gap. We recall here Proposition 2:

Proof of Proposition 2. From Eq. (30), we have $\varphi''(t) \le 1$ for all $t \in \mathbb{R}$. Hence

$$h(t) = \tfrac{1}{2}t^2 - \varphi(t)$$

satisfies $h''(t) = 1 - \varphi''(t) \ge 0$ and is therefore convex. Applying Jensen's inequality to h with a square-integrable random variable X (finite second moment is guaranteed by a finite variance) gives $\mathbb{E}[h(X)] \ge h(\mathbb{E}[X])$. Substituting for h gives

$$\frac{1}{2}\mathbb{E}[X^2] - \mathbb{E}\big[\varphi(X)\big] \ge \frac{1}{2}\big(\mathbb{E}[X]\big)^2 - \varphi\big(\mathbb{E}[X]\big),$$

which directly implies

$$\frac{1}{2}\operatorname{Var}(X) = \frac{1}{2}\Big(\mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2\Big) \ge \mathbb{E}\big[\varphi(X)\big] - \varphi\big(\mathbb{E}[X]\big).$$

Setting $X := \bar{g}_i(x)/p_i$, with the expectation taken over $\xi_{1:B}$, yields

$$\mathbb{E}_{\xi_{1:B}}\big[\varphi\big(\bar{g}_i(x)/p_i\big)\big] - \varphi\big(\mathbb{E}_{\xi_{1:B}}[\bar{g}_i(x)]/p_i\big) \;\le\; \frac{1}{2p_i^2}\operatorname{Var}\big(\bar{g}_i(x)\big). \tag{31}$$

Multiplying (31) by p_i and using $\operatorname{Var}(\bar{g}_i(x)) = \frac{1}{B}\operatorname{Var}(g_i(x,\xi))$, which holds since $\xi_1, \dots, \xi_B$ are i.i.d., gives the claim with $\mathbb{E}_{\xi_{1:B}}[\bar{g}_i(x)]$ in place of $\mathbb{E}_{\xi}[g_i(x,\xi)]$; these coincide because $\bar{g}_i(x)$ is an unbiased estimator of the constraint, which finishes the proof. □

Proposition 2 quantifies how the bias introduced by the penalty/barrier transformation depends on the batch size and the penalty parameters. Moreover, the bound is tight: on the quadratic branch of φ, where $\varphi(t) = t + \frac{1}{2}t^2$, inequality (31) holds with equality (this is immediate by substituting for φ in (31)), so the Jensen gap equals $\operatorname{Var}(g_i(x,\xi))/(2Bp_i)$ whenever $\bar{g}_i(x)/p_i$ lies in the quadratic region almost surely.

E.2 SPBM Reduces to SSL-ALM-like Primal-Dual Update

Ben-Tal & Zibulevsky (1997) showed that PBM reduces to the standard augmented Lagrangian method Hestenes (1969); Powell (1969) for a particular choice of penalty-updating rule. We show the reduction survives mini-batching. Assume *(i)* the penalty-updating function (12), *(ii)* no dual damping, $\gamma^k \equiv 0$, *(iii)* $\rho\lambda^k \in [\ell_c, u_c]^m$, and *(iv)* that $\bar{g}_i(x^k)/p_i^k$ lies in the quadratic branch of φ (i.e. $\bar{g}_i(x^k) \geq -p_i^k/2$ for φ_{QL} or $\bar{g}_i(x^k) \geq -p_i^k/3$ for φ_{QR}), where the quadratic branch of $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$ satisfies

$$\varphi(t) = t + \tfrac{1}{2}t^2, \qquad \varphi'(t) = 1 + t.$$

Then the Lagrangian becomes the augmented Lagrangian with penalty ρ:

$$\begin{aligned}
\overline{\mathcal{L}}_{p^k}(x^k, \lambda^k) &= \bar{f}(x^k) + \sum_{i=1}^m \lambda_i^k p_i^k \varphi\big(\bar{g}_i(x^k)/p_i^k\big) = \bar{f}(x^k) + \sum_{i=1}^m \lambda_i^k p_i^k \left(\frac{\bar{g}_i(x^k)}{p_i^k} + \frac{\bar{g}_i^2(x^k)}{2(p_i^k)^2}\right) \\
&= \bar{f}(x^k) + \sum_{i=1}^m \left(\lambda_i^k \bar{g}_i(x^k) + \frac{\lambda_i^k \bar{g}_i^2(x^k)}{2p_i^k}\right) \\
&= \bar{f}(x^k) + (\lambda^k)^\top \bar{g}(x^k) + \frac{1}{2\rho}\left\|\bar{g}(x^k)\right\|^2.
\end{aligned}$$

The dual update gives

$$\lambda^{k+1} = \lambda^k \odot \varphi'\big(\bar{g}(x^k) \oslash p^k\big) = \lambda^k + \lambda^k \odot \frac{\bar{g}(x^k)}{\lambda^k \rho} = \lambda^k + \frac{1}{\rho}\,\bar{g}(x^k),$$

which is the ALM multiplier update. The reduction is sample-path-wise: the algebra is pointwise in the realized $\bar{g}(x^k)$, so it is unaffected by mini-batching. Mini-batching does, however, bias the augmented Lagrangian itself. Since $\bar{f}$ is unbiased and $\lambda^\top \bar{g}$ is linear in $\bar{g}$, only the quadratic term contributes:

$$\mathbb{E}\big[\tfrac{1}{2\rho}\bar{g}_i^2(x)\big] - \tfrac{1}{2\rho}\big(\mathbb{E}[\bar{g}_i(x)]\big)^2 = \tfrac{1}{2\rho}\operatorname{Var}\big(\bar{g}_i(x)\big) = \frac{\operatorname{Var}\big(g_i(x,\xi)\big)}{2\rho B},$$

recovering the $1/B$ decay of Proposition 2 in this special case. It is worth noting that the recovered update is that of the *equality*-constrained ALM: for strongly inactive constraints $\bar{g}_i < -p_i^k/2$ (or $\bar{g}_i < -p_i^k/3$ for φ_{QR}), φ leaves its quadratic branch and SPBM decays λ_i multiplicatively compared to ALM, so the methods differ there. Lastly, since SPBM utilizes the Moreau envelope, setting the penalty-updating function to (12) yields the SSL-ALM algorithm up to the order of the dual and primal update steps, the prox-center index, and the coupling of the dual learning rate and the penalty regularizer. Note that κ becomes a redundant hyperparameter and is instead replaced by ρ.

E.3 Convergence Analysis

The analysis applies to Algorithm 2, a variant of the benchmarked Algorithm 1 that drives the dual and penalty averaging coefficients to a limit along fixed schedules.

We work on a probability space $(\Omega, \mathcal{F}, P)$ carrying the mini-batch draws $\{\mathbf{j}^k\}_{k\geq 0}$, with the filtration $\mathcal{F}^k := \sigma(\mathbf{j}^0, \ldots, \mathbf{j}^{k-1})$. At step k the algorithm returns the direction $\bar{y}^k$ of line 4, which we express as a sum of its conditional mean and the noise of the problem:

$$y^k := \mathbb{E}\big[\bar{y}^k \mid \mathcal{F}^k\big], \qquad \zeta^k := \bar{y}^k - y^k,$$

so that the primal update of line 5 reads

$$x^{k+1} = x^k + \alpha^k\big(y^k + \zeta^k\big). \tag{32}$$

Here x^k, λ^k, p^k, s^k and y^k are $\mathcal{F}^k$-measurable, while $\bar{y}^k$ and ζ^k are $\mathcal{F}^{k+1}$-measurable. Let $\{x^k, \lambda^k, p^k, s^k\}_{k\geq 0}$ be generated by Algorithm 2. We show that the duals and the penalty parameters converge almost surely to limits $\hat{\lambda}$ and $\hat{p}$, and that the primal iterates asymptotically track the differential inclusion

$$\dot{x}(t) \in \mathcal{D}_{\hat{p}}\big(x(t), \hat{\lambda}\big),$$

whose equilibria are the $\mathcal{D}_{\hat{p}}(\cdot,\hat{\lambda})$-critical points appearing in Theorem 3. Here $\mathcal{D}_{\hat{p}}(\cdot,\hat{\lambda})$ is a conservative field of $-\mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot,\hat{\lambda})$, where

$$\mathcal{L}^{\mathrm{emp}}_{p}(\cdot,\lambda) := \mathbb{E}\big[\bar{\mathcal{L}}_p(\cdot,\lambda)\big]$$

is the expected value of the Lagrangian of the transformed problem (13). The schedules are required to satisfy $\gamma^k, \kappa^k \to 1$ with $\sum_k (1-\gamma^k) < \infty$ and $\sum_k (1-\kappa^k) < \infty$, and $\sum_k \alpha^k = \infty$ with $\sum_k (\alpha^k)^2 < \infty$. For instance, we suggest

$$\kappa^{k+1} = 1 - \frac{1}{\lfloor k/L \rfloor^{1+\varepsilon_\kappa} + \frac{1}{1-\kappa^0}}, \qquad \varepsilon_\kappa > 0, \tag{33}$$

$$\gamma^{k+1} = 1 - \frac{1}{\lfloor k/L \rfloor^{1+\varepsilon_\gamma} + \frac{1}{1-\gamma^0}}, \qquad \varepsilon_\gamma > 0, \tag{34}$$

$$\alpha^{k+1} = \frac{1}{(1+k)^{\varepsilon_\alpha}}, \qquad \varepsilon_\alpha \in (\tfrac{1}{2}, 1]. \tag{35}$$

Boundedness of the duals and penalties is enforced rather than assumed. The duals are projected onto $[0,\Lambda]^m$ in line 6, and the penalties implicitly onto $[\ell_c, u_c]^m$ by both suggested updating methods (11) and (12).

Algorithm 2 Stochastic Penalty-Barrier Method (SPBM) with schedules. Schedules: γ^k, κ^k follow (33),(34) ($\to$ 1); α^k follows a square-summable, but not summable schedule (35)($\to$ 0). `PrimalUpdate` is standard SGD step (32). Since κ^k is now scheduled, we redefine (11) as $\hat{\pi}^{\mathrm{ADAPT}}(\bar{g}, p, \lambda, \kappa^k)$. We assume ρ is fixed, so for consistency we simply redefine (12) as $\hat{\pi}^{\mathrm{ALM}}_{\rho}(\bar{g}, p, \lambda, \cdot)$.

1: **Input:** $x^0 \in \mathbb{R}^n$; $\lambda^0 \in \mathbb{R}^m_{++}$; $p^0 = \mathbf{1}$(or $p^0 = \rho\,\lambda^0$ if $\hat{\pi} = \hat{\pi}^{\mathrm{ALM}}_{\rho}$); $s^0 = x^0$; schedules α^0, γ^0 and κ^0(alternatively $\rho \in \mathbb{R}_{++}$, if $\hat{\pi} = \hat{\pi}^{\mathrm{ALM}}_{\rho}$); $\mu \geq 0$, $\delta \in [0,1)$; batch size B; $\hat{\pi} \in \{\hat{\pi}^{\mathrm{ALM}}_{\rho}, \hat{\pi}^{\mathrm{ADAPT}}\}$; $\Lambda \in \mathbb{R}_{++}$.
2: **for** $k = 0$ **to** K **do**
3: Sample $\xi_1, \ldots, \xi_B \overset{\text{iid}}{\sim} \Xi$
4: $\bar{y}^k = -\overline{\nabla}\mathcal{L}_{p^k}(x^k, \lambda^k) - \mu(x^k - s^k)$ {Moreau-augmented grad, (10)}
5: $x^{k+1} = x^k + \alpha^k \bar{y}^k$ {Primal step, (5), (32)}
6: $\lambda^{k+1} = \mathrm{Proj}_{[0,\Lambda]^m}\left[\gamma^k \lambda^k + (1-\gamma^k)\,\lambda^k \odot \varphi'(\overline{g}(x^k) \oslash p^k)\right]$ {Dual variable step, (6)}
7: $p^{k+1} = \hat{\pi}(\overline{g}(x^k), p^k, \lambda^{k+1}, \kappa^k)$ {Penalty param. step, (11)/(12)}
8: $s^{k+1} = \delta s^k + (1-\delta) x^{k+1}$ {Prox-center update}
9: Advance schedules $\alpha^k, \gamma^k, \kappa^k$ {Schedule updates}
10: **end for**

The section is organised as follows. Section E.3.1 states the two standing assumptions and their direct implications. Section E.3.2 then proves the convergence Theorem 3, pointing to lemmas that are stated and proved afterwards. Lastly, Section E.3.3 establishes convergence of the duals and penalties together with condition on limiting properties of y^k, and Section E.3.4 the Lyapunov properties behind the descent and the Weak-Sard property.

E.3.1 Assumptions and Direct Implications

Let $(\Omega, \mathcal{F}, P)$ be a probability space carrying the independent mini-batch draws $\{\mathbf{j}^k\}_{k\geq 0}$, such that $\mathbf{j}^k \in \{1,\ldots,M\}^B$ where $M \in \mathbb{N}$ is the number of samples in our dataset. And let $\{\mathcal{F}^k\}_{k\geq 0}$ be the filtration generated by the batches drawn before iteration k:

$$\mathcal{F}^k := \sigma\big(\mathbf{j}^0, \ldots, \mathbf{j}^{k-1}\big), \qquad \mathcal{F}^0 := \{\emptyset, \Omega\}, \qquad \mathcal{F}_\infty := \sigma\Big(\bigcup_{k\geq 0} \mathcal{F}^k\Big),$$

so that $\mathcal{F}^k \subseteq \mathcal{F}^{k+1} \subseteq \mathcal{F}$. This gives that x^k, λ^k, p^k, s^k and y^k are $\mathcal{F}^k$-measurable but $\bar{y}^k$ need not be, and the draw $\mathbf{j}^k$ is independent of $\mathcal{F}^k$.

Specifically, assume a finite dataset $(\xi^{(1)}, \dots, \xi^{(M)})$, possibly with repetitions, and Ξ its empirical measure; i.e. Ξ is the law of $\xi^{(l)}$ for $l \sim \mathrm{Unif}\{1, \dots, M\}$, in particular $\operatorname{supp} \Xi$ is finite. So drawing $\xi \sim \Xi$ is the same as drawing an index uniformly from $\{1, \dots, M\}$, and a mini-batch of B i.i.d. draws is encoded by a multi-index

$$\mathbf{j} = (j_1, \dots, j_B) \in \{1, \dots, M\}^B.$$

We accordingly write $\bar{g}_i(x; \mathbf{j}) := \frac{1}{B} \sum_{b=1}^{B} g_i(x, \xi^{(j_b)})$, and likewise $\bar{f}(\cdot\,; \mathbf{j})$. This connects the notation of Section E.1 with the one used below. Concretely, the run may then be realised on the probability space with $\Omega = (\{1, \dots, M\}^B)^{\mathbb{N}_0}$, equipped with its product sigma algebra and the product of uniform measures P. In particular, $P(\mathbf{j}^k = \mathbf{j}) = M^{-B}$.

Assumption 1 (Assumption A.2 in Davis et al. (2018))**.** *There exists a deterministic constant R such that, almost surely,*

$$\sup_{k \geq 0} \|x^k\| \leq R.$$

Remark 2. *Assumption 1 is a standard boundedness requirement in stochastic approximation. We impose no restriction on the size of R, only that it is finite. Conditions, under which such assumption can be guaranteed, are discussed in (Davis et al., 2018, Section 6.1).*

Unlike the duals and penalties, we do not project the primal iterates. Doing so would change the main statement being proved. By projecting onto $R\mathbb{B}$, the limiting dynamics would change into a projected inclusion, changing the $D_{\hat{p}}(\cdot, \hat{\lambda})$-criticality being established.

Definition 9 (Bounded primal set)**.** *Let R be the constant from Assumption 1 and set*

$$\mathcal{X} := \big\{x \in \mathbb{R}^n : \|x\| \leq R + 1\big\},$$

a compact convex set.

Corollary 2 (Condition A.1 in Davis et al. (2018))**.** *By Assumption 1, $x^k \in R\mathbb{B} \subset \operatorname{int} \mathcal{X}$ for every k, almost surely, and hence every limit point of $\{x^k\}$ lies in $R\mathbb{B} \subset \mathcal{X}$.*

Remark 3. *The padding by 1 is not cosmetic: Lemma 13, Lemma 19 and Lemma 20 bound the Clarke subdifferential at x. Since a Clarke subdifferential is determined by the nearby points, this requires a whole neighborhood of each x to lie in $\mathcal{X}$. Any fixed $\varepsilon > 0$ in place of 1 would work too.*

Lemma 10 (Schedules and projection of Algorithm 2 satisfy A.3)**.**

1. *The dual variables and penalty variables are within compact boxes almost surely*

$$\lambda^k \in [0, \Lambda]^m, \quad p^k \in [\ell_c, u_c]^m.$$

2. *(A.3 in Davis et al. (2018)) The sequence of primal learning rates $\{\alpha^k\}$ is nonnegative, square summable, but not summable:*

$$\alpha^k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha^k = \infty, \qquad \text{and} \qquad \sum_{k=1}^{\infty} (\alpha^k)^2 < \infty.$$

3. *The proposed schedule of the dual variable (34), and of the penalty parameter (33) have the property*

$$\sum_{k=0}^{\infty} (1 - \kappa^k) < \infty, \qquad \sum_{k=0}^{\infty} (1 - \gamma^k) < \infty.$$

Specifically, in the limit $\gamma^k, \kappa^k \to 1$.

Proof. (1) The duals and penalties are ensured to be within compact sets $\lambda^k \in [0, \Lambda]^m$, $p^k \in [\ell_c, u_c]^m$ through projections onto bounded boxes (see Algorithm 2). Therefore by Assumption 1, it follows that all three random variables x^k, p^k, λ^k lie in a compact set almost surely. *(2)* Follows from the suggested schedule (35) by p-series test. *(3)* It holds $1 - \gamma^{k+1} = O\left(k^{-1-\varepsilon_\gamma}\right)$ and consequently, by p-series test, the sequence is summable $\sum_{k=0}^{\infty}(1 - \gamma^k) < \infty$. The same applies to summability of $\sum_{k=0}^{\infty}(1 - \kappa^k)$. Note that when $k \leq L$, then $(\gamma^k, \kappa^k) = (\gamma^0, \kappa^0)$. □

Lemma 11 (Regularity of penalty-barriers)**.** *The function $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$ is convex,* finite-valued on all of $\mathbb{R}$*, and definable in the o-minimal structure $\mathbb{R}_{\exp}$. Moreover, $\varphi \in C^2$ and $\varphi' > 0$.*

Proof. By construction $\varphi \in \{\varphi_{\mathrm{QL}}, \varphi_{\mathrm{QR}}\}$ is convex and C^2, locally Lipschitz and definable. Specifically, using the notation taken from Bareilles et al. (2025), φ_{QR} is semi-algebraic, hence definable in $\mathbb{R}_{\mathrm{alg}} \subseteq \mathbb{R}_{\exp}$; φ_{QL} involves the logarithm on its second branch and is therefore definable in $\mathbb{R}_{\exp}$. $\varphi' > 0$ follows from (8). □

Assumption 2 (Definability, finite support, and the subgradient oracle)**.**

1. *For every $\mathbf{j} \in \{1, \dots, M\}^B$ the mini-batch estimators $\bar{f}(\cdot; \mathbf{j})$ and $\bar{g}_i(\cdot; \mathbf{j})$ are locally Lipschitz and definable in the common $\mathcal{O} \supseteq \mathbb{R}_{\exp}$. We fix $\mathcal{O}$ from this point on.*
2. *dataset $(\xi^{(1)}, \dots, \xi^{(M)})$ is finite, and Ξ is the law of $\xi^{(l)}$ for j being uniformly distributed over M samples, which gives*

$$\mathcal{L}_p^{\mathrm{emp}}(x, \lambda) := \mathbb{E}_{\xi_{1:B}}\left[\bar{\mathcal{L}}_p(x, \lambda)\right] = \frac{1}{M^B} \sum_{\mathbf{j} \in \{1,\dots,M\}^B} \bar{\mathcal{L}}_p(x, \lambda; \mathbf{j}).$$

   In other words $\mathcal{L}_p^{\mathrm{emp}}(\cdot, \lambda)$ is the average of the mini-batch over all possible M^B batches, giving a deterministic value corresponding to the true expectation value of $\overline{\mathcal{L}}_p(\cdot, \lambda)$.
3. *(Subgradient Backpropagation Oracle) For each $\mathbf{j} \in \{1, \dots, M\}^B$ there is a Borel map $\mathrm{ch}(\cdot, \cdot, \cdot; \mathbf{j}) : \mathbb{R}^n \times [0, \Lambda]^m \times [\ell_c, u_c]^m \to \mathbb{R}^n$ with $\mathrm{ch}(x, \lambda, p; \mathbf{j}) \in -\partial_x \bar{\mathcal{L}}_p(x, \lambda; \mathbf{j})$, and line 4 of Algorithm 1 (and Algorithm 2) returns $\bar{y}^k = \mathrm{ch}(x^k, \lambda^k, p^k; \mathbf{j}^k) - \mu(x^k - s^k)$.*

Remark 4. *Regarding (1), we argue in Section D.1.2 that o-minimal setting is natural in the context of deep learning. Moreover, note that from Lemma 11 it follows that φ is definable in $\mathcal{O}$. For the rest of the proof, definable means definable in $\mathcal{O}$. Definability is later required for definability of the Lagrangian function, which propagates into the definability of the conservative field of the Lyapunov function; without this the Morse-Sard theorem could not be applied. Regarding (3), since we treat SPBM as a subgradient method, an oracle that gives the access to an element of the Clarke subdifferential is reasonable. Although many modern ML libraries implementing automatic differentiation, e.g. PyTorch and TensorFlow, do not necessarily output an element of a Clarke subdifferential, there exist implementations that do, such as Kakade & Lee (2018).*

Later, we analyze a Lyapunov function based on $\mathcal{L}_p^{\mathrm{emp}}(x, \lambda)$ with fixed p, λ. From Assumption 2, we can deduce properties of this function:

Lemma 12 (Definability and regularity of the Lagrangian Estimate)**.** *Under Assumption 1,Assumption 2:*

1. *for every fixed $\lambda \in [0, \Lambda]^m$ and $p \in [\ell_c, u_c]^m$, the map*

$$\mathcal{L}_p^{\mathrm{emp}}(x, \lambda) = \mathbb{E}_{\xi_{1:B}}\left[\bar{\mathcal{L}}_p(x, \lambda)\right] = \frac{1}{M^B} \sum_{\mathbf{j} \in \{1,\dots,M\}^B} \overline{\mathcal{L}}_p(x, \lambda; \mathbf{j})$$

   is locally Lipschitz, bounded from below on the compact set $\mathcal{X}$, and definable in $\mathcal{O}$; the same holds for each $\overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j})$;
2. *the joint maps $(x, \lambda, p) \mapsto \overline{\mathcal{L}}_p(x, \lambda; \mathbf{j})$ and $(x, \lambda, p) \mapsto \mathcal{L}_p^{\mathrm{emp}}(x, \lambda)$ are locally Lipschitz on $\mathbb{R}^n \times \mathbb{R}^m \times (0, \infty)^m$.*

Proof. Step 1 (Definable and locally Lipschitz on $\mathcal{X}$). Fix $\mathbf{j}$ and i. By Assumption 2(1) and $p_i > 0$, the map $x \mapsto \bar{g}_i(x; \mathbf{j})/p_i$ is definable and locally Lipschitz. Since φ is definable and is C^2 (hence locally Lipschitz; Lemma 11),

$$\text{the composition} \quad x \mapsto \varphi(\bar{g}_i(x; \mathbf{j})/p_i)$$

is definable (compositions of definable maps within the same structure are definable; see van den Dries (1998)) and locally Lipschitz (compositions of locally Lipschitz maps are locally Lipschitz). The same holds for $\bar{f}(\cdot; \mathbf{j})$. Finite sums over nonnegative combinations (over i) preserve both properties (o-minimal structures are closed under $+$; a local Lipschitz constant of the sum is bounded by the sum of the constants). Hence $\mathcal{L}_p^{\mathrm{emp}}(\cdot, \lambda)$ is locally Lipschitz, and definable. Since $\mathcal{X}$ is compact, and

$\mathcal{L}_p^{\text{emp}}(x,\lambda)$ is locally Lipschitz, by the extreme value theorem, we get that $\mathcal{L}_p^{\text{emp}}(x,\lambda)$ is bounded from below on $\mathcal{X}$. Finally, a locally Lipschitz function definable in an o-minimal structure admits a Whitney (b)-stratification compatible with the projection, hence admits a chain rule (Davis et al., 2018). Note that all of the above also applies to $\overline{\mathcal{L}}_p(x,\lambda;\mathbf{j})$.

Step 2 (Locally Lipschitz joint maps). Fix $\mathbf{j}$ and i. The argument of *Step 1* applies once the three maps involving λ and p are checked to be locally Lipschitz jointly rather than in x alone. First, $(x,p)\mapsto \bar{g}_i(x;\mathbf{j})/p_i$ is locally Lipschitz on $\mathbb{R}^n\times(0,\infty)^m$, being a composition of locally Lipschitz maps whose denominator is locally bounded away from zero. Second, φ is C^2 by Lemma 11, hence locally Lipschitz, so composing preserves the property. Third, $\lambda_i p_i$ is no longer a constant, but a product of locally Lipschitz maps of (x,λ,p) is locally Lipschitz. The remaining steps: local Lipschitzness of $\bar{f}(\cdot;\mathbf{j})$, closure of the property under finite sums over i, and averaging over $\mathbf{j}$ remain the same as in *Step 1*, which finishes the proof. □

Lemma 13 (Membership and boundedness of y^k)**.** *The direction $\bar{y}^k$ generated in line 4 of Algorithm 2 satisfies*

$$y^k \;:=\; \mathbb{E}\big[\bar{y}^k \mid \mathcal{F}^k\big] \;\in\; \mathcal{D}_{p^k}(x^k,\lambda^k) - \mu\,(x^k - s^k), \quad \textit{almost surely},$$

where $\mathcal{D}_{p^k}(\cdot,\lambda^k):\mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is convex- and compact-valued, upper-semicontinuous, definable conservative field of a potential function $-\mathcal{L}^{\text{emp}}_{p^k}(\cdot,\lambda^k)$. Moreover, $\sup_{k\geq 0}\max\{||y^k||,||\bar{y}^k||\}$ is bounded by a deterministic constant $C_y<\infty$ almost surely.

Proof. Step 1 (Membership and properties of the conservative field). Let $\mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j})$ be the subgradient backpropagation oracle from Assumption 2(3). Then

$$\bar{y}^k = \mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j}^k) - \mu(x^k - s^k).$$

Taking the conditional expected value, we get by using measurability of x^k, s^k

$$\mathbb{E}\big[\bar{y}^k \mid \mathcal{F}^k\big] = \mathbb{E}[\mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j}^k) \mid \mathcal{F}^k] - \mu(x^k - s^k).$$

By the finiteness of $\{1,\dots,M\}^B$ (Assumption 2(2)), we rewrite the random variable of choice as

$$\mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j}^k) = \sum_{\mathbf{j}\in\{1,\dots,M\}^B} \mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j})\mathbf{1}_{\{\mathbf{j}=\mathbf{j}^k\}}.$$

From Assumption 2(3), we have that *(i)* $\mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j})$ is $\mathcal{F}^k$-measurable, being the Borel map $\mathrm{ch}(\cdot,\cdot,\cdot;\mathbf{j})$ composed with the $\mathcal{F}^k$-measurable vector (x^k,λ^k,p^k), *(ii)* the definition of an indicator random variable $\mathbf{1}_{\{\mathbf{j}=\mathbf{j}^k\}}$ gives that $\mathbf{1}_{\{\mathbf{j}=\mathbf{j}^k\}}\in L^1(\Omega,\mathcal{F},P)$, *(iii)* since $\partial\overline{\mathcal{L}}_{p^k}(x^k,\lambda^k;\mathbf{j})$ is locally bounded (Clarke subdifferential of a locally Lipschitz is locally bounded), x^k,λ^k,p^k lie in a compact set almost surely, and the indicator function is bounded by 1, then $\mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j})\mathbf{1}_{\{\mathbf{j}=\mathbf{j}^k\}}\in L^1(\Omega,\mathcal{F},P)$. Combining *(i)-(iii)*, we apply (Dembo, 2019, Proposition 4.2.10) to pull-out $\mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j})$, and then in the second step we use the independence of the draw at iterate k of previous draws, to get

$$\begin{aligned}\mathbb{E}\big[\bar{y}^k \mid \mathcal{F}^k\big] &= \sum_{\mathbf{j}\in\{1,\dots,M\}^B} \mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j})\mathbb{E}[\mathbf{1}_{\{\mathbf{j}=\mathbf{j}^k\}} \mid \mathcal{F}^k] - \mu(x^k - s^k)\\ &= \frac{1}{M^B}\sum_{\mathbf{j}\in\{1,\dots,M\}^B} \mathrm{ch}(x^k,\lambda^k,p^k;\mathbf{j}) - \mu(x^k - s^k).\end{aligned}$$

By the definition of the Minkowski sum and by Assumption 2(3), we have that

$$y^k \;:=\; \mathbb{E}\big[\bar{y}^k \mid \mathcal{F}^k\big] \;\in\; \frac{1}{M^B}\sum_{\mathbf{j}\in\{1,\dots,M\}^B} -\partial_x\overline{\mathcal{L}}_{p^k}(x^k,\lambda^k;\mathbf{j}) - \mu\,(x^k - s^k), \quad \text{almost surely.}$$

From Lemma 12(1), we have that $-\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})$ is a locally Lipschitz, definable function, thus $-\frac{1}{M^B}\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})$ is a definable, locally Lipschitz function. Therefore its Clarke subdifferential is a definable set-valued mapping (Bolte et al., 2007). And by Clarke calculus (Clason & Valkonen, 2026, Theorem 13.10), it holds

$$y^k \;:=\; \mathbb{E}\big[\bar{y}^k \mid \mathcal{F}^k\big] \;\in\; \sum_{\mathbf{j}\in\{1,\dots,M\}^B} \partial_x\left(-\frac{1}{M^B}\overline{\mathcal{L}}_{p^k}(x^k,\lambda^k;\mathbf{j})\right) - \mu\,(x^k - s^k), \quad \text{almost surely,}$$

where each element of the sum is a Clarke definable set-valued subdifferential. Moreover, by Lemma 8, we have that a Clarke subdifferential of a locally Lipschitz, definable function is also a conservative field. Furthermore, by Corollary 1 (Bolte & Pauwels, 2021, Corollary 4), we get that

$$\mathcal{D}_{p^k}(\cdot,\lambda^k) = \sum_{\mathbf{j}\in\{1,\dots,M\}^B} \partial_x\left(-\frac{1}{M^B}\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})\right) = -\frac{1}{M^B}\sum_{\mathbf{j}\in\{1,\dots,M\}^B}\partial_x\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})$$

is a conservative field of a potential function $-\mathcal{L}^{\text{emp}}_{p^k}(\cdot,\lambda^k)$, hence outer-semicontinuous and locally bounded, hence upper-semicontinuous. Moreover, the sum of sets is a Minkowski sum, which preserves convexity, and so $\mathcal{D}_{p^k}(\cdot,\lambda^k)$ is convex-valued. Moreover, Minkowski sum also preserves definability (van den Dries, 1998, Ch. 1) . Lastly $\mathcal{D}_{p^k}(\cdot,\lambda^k)$ is compact as it is a property of a conservative field, which completes the proof.

Step 2 (Boundedness of y^k). We first bound the proximal term. Since $s^0 = x^0$ and $s^{k+1} = \delta s^k + (1-\delta)x^{k+1}$ with $\delta \in [0,1)$, an induction shows that each s^k is a convex combination of $x^0,\dots,x^k$. By Assumption 1 these lie in the convex set $R\mathbb{B}$, hence $\|s^k\| \le R$ almost surely and

$$\|\mu(x^k - s^k)\| \;\le\; 2\mu R \;<\; \infty. \tag{36}$$

Moreover, $\mathcal{D}_{p^k}(x^k,\lambda^k)$ is a Minkowski sum of Clarke subdifferentials of the maps $\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})$ evaluated at x^k. By Lemma 12(2) these are locally Lipschitz on $\mathbb{R}^n\times\mathbb{R}^m\times(0,\infty)^m$, which contains the compact set $\mathcal{X}\times[0,\Lambda]^m\times[\ell_c,u_c]^m$. Hence the constant

$$S := \max_{\mathbf{j}\in\{1,\dots,M\}^B}\ \sup_{\lambda\in[0,\Lambda]^m,\,p\in[\ell_c,u_c]^m} \operatorname{Lip}_{\mathcal{X}}\big(\overline{\mathcal{L}}_p(\cdot,\lambda;\mathbf{j})\big)$$

is finite. By Corollary 2, almost surely $x^k\in\operatorname{int}\mathcal{X}$, and by Lemma 10(1) p^k,λ^k are within compact sets $[0,\Lambda]^m,[\ell_c,u_c]^m$ almost surely. So each $\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})$ is Lipschitz near x^k with constant S, and by (Clason & Valkonen, 2026, Lemma 13.2) every element of $\mathcal{D}_{p^k}(x^k,\lambda^k)$ has a norm at most S. Together with (36), it holds $\|y^k\|\le C_y<\infty$ almost surely for the deterministic $C_y = S+2\mu R$. Note that by using the same arguments the same applies to the oracle value $\operatorname{ch}(x^k,\lambda^k,p^k;\mathbf{j}^k)$ for any k, hence by Assumption 2(3), also $\|\bar{y}^k\|\le C_y$ which finishes the proof. □

Corollary 3 (Dependence of a conservative field $\mathcal{D}$ on λ,p)**.** *Note that we did not assume anything specific about λ^k,p^k, hence the conservative field $\mathcal{D}_p(\cdot,\lambda)$ has the desired properties for any λ,p. Also, note that none of the properties rely on the negative sign of each element, hence*

$$-\mathcal{D}_{p^k}(\cdot,\lambda^k) = \frac{1}{M^B}\sum_{\mathbf{j}\in\{1,\dots,M\}^B}\partial\overline{\mathcal{L}}_{p^k}(\cdot,\lambda^k;\mathbf{j})$$

is a convex-valued, upper-semicontinuous and definable conservative field of the potential function $\mathcal{L}^{\text{emp}}_{p^k}(\cdot,\lambda^k)$.

Proof. Follows immediately from the proof of Lemma 13 showing no assumptions on the fixed p,λ, and from the sign properties of a Clarke subdifferential. □

Lemma 14 (Bounded conditional second moment of $\bar{y}^k$)**.** *Let Assumption 1, Assumption 2 hold. Then the subgradient element approximation has a bounded variance*

$$\mathbb{E}\left[||\bar{y}^k - y^k||^2 \mid \mathcal{F}^k\right] \le \sigma^2_{\bar{y}}, \quad a.s.$$

Proof. By using the triangle inequality, $\mathcal{F}^k$-measurability of y^k, and by the boundedness property given by Lemma 13, it holds

$$\begin{aligned}\mathbb{E}\left[||\bar{y}^k - y^k||^2 \mid \mathcal{F}^k\right] &\le \mathbb{E}\left[||\bar{y}^k||^2 \mid \mathcal{F}^k\right] + 2\mathbb{E}\left[||\bar{y}^k|| \mid \mathcal{F}^k\right]\,||y^k|| + ||y^k||^2\\ &\le 4C_y^2 := \sigma_y^2 < \infty.\end{aligned}$$

□

We express the noise $\zeta^k := \bar{y}^k - y^k$ of the recursion (32). We show that the noise sequence is a martingale sequence difference.

Lemma 15 (ζ^k is a martingale difference sequence)**.** *Let Assumption 1, Assumption 2 hold. Then $\{\zeta^k\}$ is a martingale difference sequence w.r.t. to the increasing σ-fields $\mathcal{F}^k$. Specifically, there exists a function $h : \mathbb{R}^n \to [0, \infty)$ bounded on bounded sets, so that almost surely, for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}[\zeta^k \mid \mathcal{F}^k] = 0 \quad \textit{and} \quad \mathbb{E}[\|\zeta^k\|^2 \mid \mathcal{F}^k] \le h(x^k) \quad a.s.$$

Proof. The first condition follows from $\mathcal{F}^k$-measurability of y^k:

$$\mathbb{E}[\zeta^k \mid \mathcal{F}^k] = \mathbb{E}[\bar{y}^k \mid \mathcal{F}^k] - y^k = 0 \quad a.s.$$

By definition of ζ^k and Lemma 14

$$\mathbb{E}[\|\zeta^k\|^2 \mid \mathcal{F}^k] = \mathbb{E}[\|\bar{y}^k - y^k\|^2 \mid \mathcal{F}^k] \le \sigma_{\bar{y}}^2,$$

hence by taking $h = \sigma_y^2$, the second condition holds almost surely. □

Lemma 16 (Condition A.4 in Davis et al. (2018))**.** *Let Assumption 1, Assumption 2 hold. There exists a finite random vector q such that the following holds almost surely*

$$\lim_{j\to\infty} \sum_{k=1}^{j} \alpha^k \zeta^k = q.$$

Proof. Since $\{\alpha^k\}$ is square summable (Lemma 10(2)), and from Lemma 14 it follows

$$\sum_{i=0}^{\infty} (\alpha^i)^2 \mathbb{E}[\|\zeta^i\|^2 \mid \mathcal{F}^i] \le \sigma_{\bar{y}}^2 \sum_{i=0}^{\infty} (\alpha^i)^2 < \infty.$$

Hence $X_k = \sum_{i=0}^{k} \alpha^i \zeta^i$ satisfies $\mathbb{E}\|X_k\|^2 < \infty$ for every k. Since $\mathbb{E}[\zeta^k \mid \mathcal{F}^k] = 0$ and α^k is deterministic, $\{X_k\}$ is a square-integrable, $\{\mathcal{F}^{k+1}\}$-measurable martingale. Therefore the martingale X_k is in L^2. Its predictable compensator is

$$\langle X \rangle_j = \sum_{k=1}^{j} \mathbb{E}\left[\|\sum_{i=0}^{k} \alpha^i \zeta^i - \sum_{i=0}^{k-1} \alpha^i \zeta^i\|^2 \mid \mathcal{F}^k \right] = \sum_{k=1}^{j} \mathbb{E}\left[\|\alpha^k \zeta^k\|^2 \mid \mathcal{F}^k \right] \le \sigma_{\bar{y}}^2 \sum_{k=1}^{j} (\alpha^k)^2.$$

And so $\langle X \rangle_\infty$ is finite almost surely. By (Dembo, 2019, Theorem 5.3.33) this gives that X_k converges to a finite limit almost surely, which completes the proof. The arguments of this lemma are mostly taken from (Davis et al., 2018, Lemma 4.1.) and extended to our setting. □

E.3.2 Proof of the Main Theorem

Theorem 3 (Primal criticality and transformed Lagrangian convergence)**.** *Let Assumption 1 and Assumption 2 hold. Let $\{x^k, \lambda^k, p^k, s^k\}_{k\ge 0}$ be generated by Algorithm 2. Then almost surely there exist random $\hat{\lambda} \in [0, \Lambda]^m$, $\hat{p} \in [\ell_c, u_c]^m$ such that*

1. *$\lambda^k \to \hat{\lambda}$, $p^k \to \hat{p}$;*
2. *$\|x^k - s^k\| \to 0$;*
3. *the function values $\{\mathcal{L}_{\hat{p}}^{\mathrm{emp}}(x^k, \hat{\lambda})\}_{k\ge 0}$ converge;*
4. *every limit point $\hat{x}$ of $\{x^k\}_{k\ge 0}$ satisfies $0 \in \mathcal{D}_{\hat{p}}(\hat{x}, \hat{\lambda})$,*

where $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field of the potential function $-\mathcal{L}_{\hat{p}}^{\mathrm{emp}}(\cdot, \hat{\lambda})$.

Proof of Theorem 3. The proof reduces SPBM to the stochastic approximation framework of Davis et al. (2018). Since the duals and penalties change only slightly in the limit, the iteration asymptotically tracks SGD applied to the fixed function $\mathcal{L}_{\hat{p}}^{\mathrm{emp}}(\cdot, \hat{\lambda})$, and we may verify their conditions [A.1]–[A.5] and [B.1]–[B.2] for the pair $\left(\Phi_{\hat{p}}(\cdot, \hat{\lambda}),\ \mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})|_{\mathcal{X}}\right)$ consisting of a Lyapunov function and a restricted conservative field. All references to [A.·] and [B.·] below are assumptions of Davis et al. (2018).

Step 1 (Preliminaries). By Lemma 17 the duals and penalties converge almost surely to some $\hat{\lambda} \in [0, \Lambda]^m$ and $\hat{p} \in [\ell_c, u_c]^m$, and by Lemma 18 the proximal term vanishes, $\|x^k - s^k\| \to 0$ almost surely. This proves items *(1)* and *(2)* of the theorem.

Step 2 (The limiting objects). The limits $\hat{\lambda}, \hat{p}$ are random, hence so are the objects built from them, whereas the framework of Davis et al. (2018) is stated for a fixed set-valued map. We therefore argue path-wise. Let Ω_0 be the intersection of the almost-sure events of Assumption 1, Lemma 10, Lemma 13, Lemma 14, Lemma 15, Lemma 16, Lemma 17, Lemma 18, Lemma 20, Lemma 21; being a finite intersection of events of full measure, Ω_0 still maintains full measure: $P(\Omega_0) = 1$. We fix $\omega \in \Omega_0$ and write $\hat{\lambda} = \hat{\lambda}(\omega)$, $\hat{p} = \hat{p}(\omega)$ for the remainder of the proof.

Set $G := \mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})|_{\mathcal{X}}$ and $\Phi_{\hat{p}}(\cdot, \hat{\lambda}) := \max\{c,\ \mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot, \hat{\lambda})\}$ with c as in Lemma 22. We cannot take $\mathcal{X} = \mathbb{R}^n$ in the proof: with $\Phi_{\hat{p}}(\cdot, \hat{\lambda}) = \max\{\mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot, \hat{\lambda}), c\}$ the Lyapunov function is constant on $\{\mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot, \hat{\lambda}) < c\}$, so the strict descent required by [B.2] is not available on curves entering that set, whereas without the maximum, Assumption [B] requires $\mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot, \hat{\lambda})$ to be bounded from below on all of $\mathbb{R}^n$, which Lemma 12(1) gives only on compact sets. The choice $\mathcal{X} = \{x : \|x\| \le R + 1\}$ is admissible because [A.1] requires only that the limit points of $\{x^k\}$ lie in $\mathcal{X}$, which holds by Assumption 1. Following Davis et al. (2018, Section 3.1), a trajectory of G is an arc $x : \mathbb{R}_+ \to \mathbb{R}^n$ satisfying $\dot{x}(t) \in G\big(x(t)\big)$ for a.e. $t \ge 0$; since the inclusion is meaningful only where G is defined, every such arc takes its values in $\mathcal{X}$, and [B.1]–[B.2] are accordingly conditions on curves in $\mathcal{X}$ alone. Since Theorem 2 (Bolte & Pauwels, 2021, Theorem 5) and Lemma 6 (Bolte & Pauwels, 2021, Lemma 2) are stated for conservative fields on $\mathbb{R}^n$, we verify [B.1] and [B.2] for the unrestricted field $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$ and transfer them to G in the proof, ultimately proving [B.1]-[B.2] for G.

Step 3 (Conditions [A.1]–[A.5]).

Table 7: Each condition is established by the result named below.

	Condition	**Established by**
[A.1]	limit points of $\{x^k\}$ lie in $\mathcal{X}$	Corollary 2
[A.2]	$\sup_k \|x^k\| < \infty$, $\sup_k \|y^k\| < \infty$	Assumption 1, Lemma 13
[A.3]	$\sum_k \alpha^k = \infty$, $\sum_k (\alpha^k)^2 < \infty$	Lemma 10 (2)
[A.4]	$\sum_k \alpha^k \zeta^k$ converges	Lemma 16
[A.5]	Cesàro averages of y^{k_j} approach $G(\hat{x})$	Lemma 21

Step 4 (Conditions [B.1]–[B.2]). $\Phi_{\hat{p}}(\cdot, \hat{\lambda})$ is continuous and bounded below by Lemma 22. The weak Sard property [B.1] is Lemma 23 and the descent property [B.2] is Lemma 24; both are proven for G, as described in *Step 2*.

Step 5 (Conclusion). All hypotheses of Davis et al. (2018, Theorem 3.2) hold for ω, so every limit point $\hat{x}$ of $\{x^k\}$ satisfies $0 \in G(\hat{x})$, and the sequence $\{\Phi_{\hat{p}}(x^k, \hat{\lambda})\}_{k \ge 0}$ converges. Since $x^k \in \mathcal{X}$, Lemma 22 gives $\Phi_{\hat{p}}(x^k, \hat{\lambda}) = \mathcal{L}^{\mathrm{emp}}_{\hat{p}}(x^k, \hat{\lambda})$, which proves item *(3)*; and $G(\hat{x}) = \mathcal{D}_{\hat{p}}(\hat{x}, \hat{\lambda})$ for $\hat{x} \in \mathcal{X}$, which proves item *(4)*. As $\omega \in \Omega_0$ was arbitrary and $P(\Omega_0) = 1$, the conclusions hold almost surely.

□

Remark 5 (Dual and penalty paths have finite length by construction)**.** *The bounds of the dual variables and of penalty parameters are not influenced by the progress of the algorithm. Instead, convergence of the duals and penalty parameters is imposed by our summable schedules, and both $\hat{\lambda}, \hat{p}$ end up within a fixed distance from the initial iterates (Lemma 17). Summability is however critical for the analysis: without it, there are no guarantees that the random $\hat{\lambda}, \hat{p}$ exist. The convergence theorem reduces to the framework of Davis et al. (2018) which assumes a path-wise fixed set-valued map G, that we build with fixed $\hat{\lambda}, \hat{p}$. Setting instead $\sum_k (1 - \gamma^k) = \infty, \sum_k (1 - \kappa^k) = \infty$ would leave the duals and penalty variables moving, making the framework unavailable. We leave this extension for future work.*

Corollary 4 (Meaning of criticality in the smooth setting)**.** *Let $\hat{x}$ be a limit point of $\{x^k\}_{k\geq 0}$ as in Theorem 3, and suppose that $\bar{f}(\cdot;\mathbf{j})$ and $\bar{g}_i(\cdot;\mathbf{j})$ are continuously differentiable on a neighbourhood of $\hat{x}$ for every $\mathbf{j}\in\{1,\dots,M\}^B$ and every $i\in\{1,\dots,m\}$. Then item (4) of Theorem 3 gives*

$$0 \;=\; \frac{1}{M^B}\sum_{\mathbf{j}\in\{1,\dots,M\}^B}\Bigg[\nabla \bar{f}(\hat{x};\mathbf{j}) \;+\; \sum_{i=1}^{m}\underbrace{\hat{\lambda}_i\,\varphi'\big(\bar{g}_i(\hat{x};\mathbf{j})/\hat{p}_i\big)}_{=:\,\bar{\lambda}_i^{\mathbf{j}}\geq 0}\nabla \bar{g}_i(\hat{x};\mathbf{j})\Bigg].$$

That is, $\hat{x}$ is a stationary point of the averaged Lagrangian with non-negative batch-dependent effective multipliers $\bar{\lambda}_i^{\mathbf{j}}$.

Proof. Fix $\mathbf{j}$. By Lemma 11 $\varphi\in C^2$, and by the hypothesis on the continuous differentiability of $\bar{f}(\cdot;\mathbf{j})$ and $\bar{g}_i(\cdot;\mathbf{j})$, it holds that $\overline{\mathcal{L}}_{\hat{p}}(\cdot,\hat{\lambda};\mathbf{j})$ is continuously differentiable on a neighbourhood of $\hat{x}$ with

$$\nabla_x\overline{\mathcal{L}}_{\hat{p}}(\hat{x},\hat{\lambda};\mathbf{j}) = \nabla \bar{f}(\hat{x};\mathbf{j}) + \sum_{i=1}^{m}\hat{\lambda}_i\,\varphi'\big(\bar{g}_i(\hat{x};\mathbf{j})/\hat{p}_i\big)\,\nabla \bar{g}_i(\hat{x};\mathbf{j}).$$

Since the Lagrangian is continuously differentiable near $\hat{x}$, it is also strictly differentiable. By strict differentiability and by (Clarke, 1990, Proposition 2.2.4), it holds that $\partial_x\overline{\mathcal{L}}_{\hat{p}}(\hat{x},\hat{\lambda};\mathbf{j}) = \{\nabla_x\overline{\mathcal{L}}_{\hat{p}}(\hat{x},\hat{\lambda};\mathbf{j})\}$, hence $\mathcal{D}_{\hat{p}}(\hat{x},\hat{\lambda})$ is a singleton given by the negative of the average and item (4) of Theorem 3 is the stated equation. Finally $\hat{\lambda}_i\geq 0$ by Lemma 10(1) and $\varphi'>0$ by Lemma 11, hence $\bar{\lambda}_i^{\mathbf{j}}\geq 0$. □

For a better understanding, we give Table 8 that links each lemma and the proof of the main theorem.

Table 8: Structure of the convergence proof. Throughout, $\mathcal{D}_{\hat{p}}(\cdot,\hat{\lambda})$ is the conservative field of $-L^{\mathrm{emp}}_{\hat{p}}(\cdot,\hat{\lambda})$ from Lemma 13, and G is its restriction to $\mathcal{X}$. The Lyapunov function $\Phi_{\hat{p}}(\cdot,\hat{\lambda})$ of Lemma 22 coincides with $L^{\mathrm{emp}}_{\hat{p}}(\cdot,\hat{\lambda})$ on $\mathcal{X}$. The upper blocks show the list of fixed-path-wise objects and the prerequisites specific to SPBM; the lower block lists the conditions of Davis et al. (2018). Conditions A.1–A.5 concern the recursion and the map G; B.1–B.2 concern the pair $(\Phi_{\hat{p}}(\cdot,\hat{\lambda}),G)$. Together they yield Theorem 3 via (Davis et al., 2018, Theorem 3.2).

	Statement	**Established by**
Objects, fixed path-wise on Ω_0		
	$\mathcal{D}_{\hat{p}}(\cdot,\hat{\lambda}):\mathbb{R}^n\rightrightarrows\mathbb{R}^n$ conservative field of $-L^{\mathrm{emp}}_{\hat{p}}(\cdot,\hat{\lambda})$	Lemma 13
	$G:=\mathcal{D}_{\hat{p}}(\cdot,\hat{\lambda})\big\vert_{\mathcal{X}}$, *Step 2* of the proof Theorem 3	Theorem 3
	$\Phi_{\hat{p}}(\cdot,\hat{\lambda})=\max\{\mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot,\hat{\lambda}),c\}=\mathcal{L}^{\mathrm{emp}}_{\hat{p}}(\cdot,\hat{\lambda})$ on $\mathcal{X}$	Lemma 22
Prerequisites for SPBM		
	$\lambda^k\to\hat{\lambda}$, $p^k\to\hat{p}$ a.s.	Lemma 17
	$\|x^k-s^k\|\to 0$ a.s.	Lemma 18
	$\mathbb{E}[\bar{y}^k\mid\mathcal{F}^k]\in\mathcal{D}_{p^k}(x^k,\lambda^k)-\mu(x^k-s^k)$ a.s.	Lemma 13
	$\mathrm{dist}\big(y^k,\mathcal{D}_{\hat{p}}(x^k,\hat{\lambda})\big)\to 0$ a.s.	Lemma 20
Conditions of Davis et al. (2018)		
[A.1]	Limit points lie in $\mathcal{X}$	Def. of $\mathcal{X}$, Corollary 2
[A.2]	$\sup_k\|x^k\|$, $\sup_k\|y^k\|<\infty$	Assumption 1, Lemma 13
[A.3]	$\alpha^k\geq 0$, $\sum_k\alpha^k=\infty$, $\sum_k(\alpha^k)^2<\infty$	Lemma 10(2)
[A.4]	$\sum_k\alpha^k\zeta^k$ convergent	Lemma 16
[A.5]	$\mathrm{dist}\big(\frac{1}{N}\sum_{j\leq N}y^{k_j},\,G(\hat{x})\big)\to 0$	Lemma 21
[B.1]	Weak Sard: $\Phi^{-1}(v)\cap G^{-1}(0)=\emptyset$ for a dense set of v	Lemma 23
[B.2]	Descent along trajectories of $\dot{x}\in G(x)$	Lemma 24
	$\Phi_{\hat{p}}(\cdot,\hat{\lambda})$ continuous, bounded below	Lemma 22
	$\mathcal{D}_{\hat{p}}(\cdot,\hat{\lambda})$ convex compact valued, definable, usc; inherited by G on $\mathcal{X}$	Lemma 13

Remark 6. *The definability in some common o-minimal structure is critical in exactly one place: to prove the Weak Sard property. Unlike the descent property, which follows from the properties of a general conservative field. Weak Sard proof, by contrast, needs Morse-Sard Theorem 2 (Bolte & Pauwels, 2021, Theorem 5), which requires both the potential function and the conservative field to be definable. The second is not implied by the first. As Bolte & Pauwels (2021) shows, by setting a conservative field*

$$D(s) = \begin{cases} [0,1], & \text{if } s \in \mathbb{Z}, \\ \{0\}, & \text{otherwise,} \end{cases}$$

D is a conservative field of a zero function, which is definable. Yet, D is not definable as $\mathbb{Z}$ is not definable in any o-minimal structure. We were able to prove that our choice of a conservative field is definable since we take $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$ to be a sum of Clarke subdifferentials of a definable function, which are always definable Bolte et al. (2007).

Remark 7. *Theorem 3 establishes only D-criticality, $0 \in \mathcal{D}_{\hat{p}}(\hat{x}, \hat{\lambda})$ which is weaker than Clarke criticality. Indeed, one can see from inclusion (29) (Bolte & Pauwels, 2021, Corollary 1), that not every D-critical point is necessarily Clarke critical. Nevertheless, the use D-criticality has been of interest in the research community recently (Bianchi et al., 2022; Bolte et al., 2023; Bolte & Pauwels, 2021).*

E.3.3 Proving Dual Convergence and A.5 of Davis et al. (2018)

As it turns out, dual variables and penalty parameter iterates form Cauchy sequences almost surely, giving that their limit points exist almost surely.

Lemma 17 (Convergence of the dual and penalty variables)**.** *Suppose Assumption 1 and Assumption 2 hold, let $\{x^k, \lambda^k, p^k, s^k\}_{k\geq 0}$ be generated by Algorithm 2. Then, almost surely, $\gamma^k, \kappa^k \to 1$ and there exist (random) $\hat{\lambda} \in [0, \Lambda]^m$, $\hat{p} \in [\ell_c, u_c]^m$ with $\lambda^k \to \hat{\lambda}$ and $p^k \to \hat{p}$. Specifically, the path lengths are finite, for a deterministic $C < \infty$:*

$$\sum_{k=0}^{\infty} \|\lambda^{k+1} - \lambda^k\| \leq \Lambda\sqrt{m}\, C \sum_{k=0}^{\infty}(1-\gamma^k),$$

for penalty schedule Eq. (11), it holds

$$\sum_{k=0}^{\infty} \|p^{k+1} - p^k\| \leq u_c\, \sqrt{m} C \sum_{k=0}^{\infty}(1-\kappa^k),$$

whereas for penalty schedule Eq. (12), it holds:

$$\sum_{k=0}^{\infty} \|p^{k+1} - p^k\| \leq \rho\Lambda\sqrt{m}\, C \sum_{k=0}^{\infty}(1-\gamma^k).$$

Proof. $\gamma^k, \kappa^k \to 1$ is immediate from (33)–(34). We know that by Assumption 1 that $x^k \in \operatorname{int} \mathcal{X}$ almost surely. Moreover, by construction $p_i^k \geq \ell_c > 0$. We can conclude that

$$\text{the arguments} \quad \bar{g}_i(x^k)\,,\ \bar{g}_i(x^k)/p_i^k \quad \text{lie in a compact set.} \tag{37}$$

Since $\bar{g}_i(\cdot; \mathbf{j})$ is locally Lipschitz (Assumption 2(1)) for any $\mathbf{j} \in \{1, \ldots, M\}^B$, and the derivative $\varphi' > 0$ is also locally Lipschitz (can be seen from $\varphi'' \leq 1$; see Eq. (30)), following the same arguments from the Pf. of Lemma 12(1), we can conclude:

$$\text{both} \quad \varphi'\big(\bar{g}_i(x^k)/p_i^k\big), \varphi'\big(\bar{g}_i(x^k)\big) \quad \text{are locally Lipschitz, hence continuous.} \tag{38}$$

By using (37), (38), and the extreme value theorem, we can conclude that there is a deterministic $C < \infty$ with, for all k, i,

$$\left|\varphi'\big(\bar{g}_i(x^k)/p_i^k\big) - 1\right| \leq C, \qquad \left|\tfrac{1}{\varphi'(\bar{g}_i(x^k))+\varepsilon} - 1\right| \leq C.$$

Dual Norm Bounds. By the ascent update rule of the dual variable, we can equivalently rewrite the iterative formula as

$$\lambda^{k+1} = \operatorname{Proj}_{[0,\Lambda]^m}\left[\lambda^k \odot \left((1-\gamma^k)\left(\varphi'(\bar{g}(x^k)/p^k) - 1\right) + 1\right)\right].$$

Equivalently, we can write (by $\mathrm{Proj}_{[0,\Lambda]^m}\left[\lambda^k\right] = \lambda^k$)

$$\lambda^{k+1} - \lambda^k = \mathrm{Proj}_{[0,\Lambda]^m}\left[\lambda^k \odot \left((1-\gamma^k)\left(\varphi'(\bar{g}(x^k)/p^k) - 1\right) + 1\right)\right] - \mathrm{Proj}_{[0,\Lambda]^m}\left[\lambda^k\right].$$

Note that we enforce the dual random variable λ^k to lie within the compact set $[0,\Lambda]^m$. Thus, by taking norms on each side of the equality, we get by the non-expansiveness of the Proj operator:

$$\begin{aligned}\|\lambda^{k+1} - \lambda^k\| &= \|\,\mathrm{Proj}_{[0,\Lambda]^m}\left[\lambda^k \odot \left((1-\gamma^k)\left(\varphi'(\bar{g}(x^k)/p^k) - 1\right) + 1\right)\right] - \mathrm{Proj}_{[0,\Lambda]^m}\left[\lambda^k\right]\| \\ &\leq \|\lambda^k \odot \left((1-\gamma^k)\left(\varphi'(\bar{g}(x^k)/p^k) - 1\right)\right)\| \\ &\leq \Lambda\sqrt{m}\,(1-\gamma^k)\,C. \end{aligned} \tag{39}$$

We can take the sums of each side of (39)

$$\sum_{k=0}^{\infty} \|\lambda^{k+1} - \lambda^k\| \leq \Lambda\sqrt{m}\,C \sum_{k=0}^{\infty}(1-\gamma^k),$$

which by our assumption of the lemma on summability $\sum_k (1-\gamma^k) < \infty$ gives that the distance traveled by the dual variable is bounded, which gives that $\{\lambda^k\}$ has summable increments. Its tails therefore vanish. Hence for any $\varepsilon_\lambda > 0$, there exists K_0 such that

$$\sum_{k=K_0}^{\infty} \|\lambda^{k+1} - \lambda^k\| \leq \varepsilon_\lambda,$$

Thus by the triangle inequality for any $q_1, q_2 > K_0$:

$$\|\lambda^{q_1} - \lambda^{q_2}\| \leq \varepsilon_\lambda.$$

Since all of the preceding steps hold almost surely, we conclude that $\{\lambda^k\}$ is Cauchy, hence $\lambda^k \to \hat{\lambda} \in [0,\Lambda]^m$ almost surely.

Penalty Norm Bounds. Rewrite the penalty updating formula (11)

$$p^{k+1} = \mathrm{Proj}_{[\ell_c,u_c]^m}\left[p^k \odot \left((1-\kappa^k)(\tfrac{1}{\varphi'(\bar{g}(x^k))+\varepsilon} - 1) + 1\right)\right]$$

The operator Proj is nonexpansive (see the same arguments when proving (39)) and fixes p^k, hence

$$\begin{aligned}\|p^{k+1} - p^k\| &= \|\,\mathrm{Proj}_{[\ell_c,u_c]^m}\left[p^k \odot \left((1-\kappa^k)\big(\tfrac{1}{\varphi'(\bar{g}(x^k))+\varepsilon} - 1\big) + 1\right)\right] - \mathrm{Proj}_{[\ell_c,u_c]^m}\left[p^k\right]\| \\ &\leq \|p^k \odot (1-\kappa^k)\big(\tfrac{1}{\varphi'(\bar{g}(x^k))+\varepsilon} - 1\big)\| \leq u_c\,\sqrt{m}C\,(1-\kappa^k)\end{aligned}$$

As $\sum_k (1-\kappa^k) < \infty$, the increments are summable, so $\{p^k\}$ is Cauchy almost surely and $p^k \to \hat{p} \in [\ell_c, u_c]^m$ almost surely.

When considering the penalty-updating formula (12), we know from the *Dual Norm Bounds*, that $\lambda^k \to \hat{\lambda}$ almost surely. Hence $p^{k+1} = \mathrm{Proj}_{[\ell_c,u_c]^m}\left[\rho\,\lambda^{k+1}\right]$ for $\rho > 0$, so by the continuity of the projection and $\lambda^k \to \hat{\lambda}$ we get $p^k \to \hat{p}$ almost surely. Using the non-expansiveness

$$\sum_{k=0}^{\infty} \|p^{k+1} - p^k\| \leq \rho \sum_{k=0}^{\infty} \|\lambda^{k+1} - \lambda^k\| \leq \rho\Lambda\sqrt{m}\,C \sum_{k=0}^{\infty}(1-\gamma^k),$$

so the path length is finite as well. In either case of (11) and (12), it holds $p^k \to \hat{p}$ almost surely.

□

Lemma 18 (Prox center collapses)**.** *Suppose Assumption 1 and Assumption 2 hold, let* $\{x^k, \lambda^k, p^k, s^k\}_{k\geq 0}$ *be generated by Algorithm 2. Then almost surely*

$$\|x^k - s^k\| \to 0.$$

Proof. Let $e^k := \|x^k - s^k\| \geq 0$. From $s^{k+1} = \delta s^k + (1-\delta)x^{k+1}$,

$$s^{k+1} - x^{k+1} = \delta\,(s^k - x^{k+1}) = \delta\,(s^k - x^k) + \delta\,(x^k - x^{k+1}),$$

so by the triangle inequality

$$e^{k+1} \leq \delta\, e^k + \delta\,\|x^{k+1} - x^k\|.$$

Now $\|x^{k+1} - x^k\| = \|\alpha^k(y^k + \zeta^k)\| \leq \alpha^k\|y^k\| + \|\alpha^k\zeta^k\| \to 0$, since $\alpha^k \to 0$ (as $\sum_k (\alpha^k)^2 < \infty$), $\|y^k\|$ is bounded almost surely (Lemma 13), and $\alpha^k\zeta^k \to 0$ (the terms of the convergent series by Lemma 16 vanish); the same argument is used in (Davis et al., 2018, Lemma 3.3). Setting $b^k := \delta\|x^{k+1} - x^k\| \geq 0$ we obtain

$$e^{k+1} \leq \delta e^k + b^k, \qquad \delta \in [0, 1), \qquad b^k \to 0. \tag{40}$$

We argue that $e^k \to 0$. Indeed, by observing Lemma 9 (Xu, 2004, Lemma 2.1), we can rewrite (40) as

$$e^{k+1} \leq \big(1 - (1-\delta)\big)\, e^k + b^k.$$

hence both

$$\sum_{k=1}^{\infty}(1-\delta) = \infty, \qquad \limsup_{k\to\infty} \frac{b^k}{1-\delta} = 0, \qquad \text{hold,}$$

hence by applying Lemma 9, we have $e^k \to 0$. All steps hold on the almost-sure event of Assumption 1, so $e^k \to 0$ almost surely. Note that Lemma 9 (Xu, 2004, Lemma 2.1) requires $\delta \in (0, 1)$. We can see that for $\delta = 0$, the almost sure convergence $e^k \to 0$ is immediate. Consequently, along any subsequence with $x^{k_j} \to \hat{x}$ we have $\|s^{k_j} - \hat{x}\| \leq e^{k_j} + \|x^{k_j} - \hat{x}\| \to 0$, so $s^{k_j} \to \hat{x}$; i.e. every limit point $(\hat{x}, \hat{s})$ of $\{(x^k, s^k)\}$ satisfies $\hat{s} = \hat{x}$, and therefore $\mu(\hat{x} - \hat{s}) = 0$. □

Now that we have proven the convergence of dual variables, and that the proximal term vanishes, one last lemma that we need is one that states a desired property of the approximate evaluation y^k of G. For that, we need to prove that a residual function is locally Lipschitz in x. For the next lemma, define the finite constants

$$G_{\max} := \max_{i\leq m,\, \mathbf{j}\in\{1,\dots,M\}^B} \max_{x\in\mathcal{X}} \big|\bar{g}_i(x;\mathbf{j})\big|, \qquad L_g := \max_{i\leq m,\, \mathbf{j}\in\{1,\dots,M\}^B} \mathrm{Lip}_{\mathcal{X}}\big(\bar{g}_i(\cdot;\mathbf{j})\big),$$
$$S_1 := \max_{t\in T} \varphi'(t), \qquad S_2 := \max_{t\in T} \varphi''(t), \qquad T := \big[-G_{\max}/\ell_c,\ G_{\max}/\ell_c\big].$$

Finiteness of L_g and $G_{\max}$ follows from local Lipschitzness on the compact set $\mathcal{X}$ together with the finiteness of the dataset (Assumption 2(1)–(2)); finiteness of S_1, S_2 from $\varphi \in C^2$ and compactness of T. For both barriers one has in fact $S_2 \leq 1$, since $\varphi'' \equiv 1$ on the quadratic branch and $\varphi'' < 1$ below it.

Lemma 19 (Perturbation in the duals and penalties). *Let* $(\lambda, p), (\lambda', p') \in [0, \Lambda]^m \times [\ell_c, u_c]^m$ *and set* $r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')} : \mathbb{R}^n \to \mathbb{R}$ *as*

$$r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')} := \overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j}) - \overline{\mathcal{L}}_{p'}(\cdot, \lambda'; \mathbf{j}).$$

then

1. *for all* $x, x' \in \mathcal{X}$, *and any* $\mathbf{j} \in \{1, \dots, M\}^B$

$$\left| r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')}(x) - r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')}(x') \right| \;\le\; C_{\mathcal{X}} \big(\|\lambda - \lambda'\| + \|p - p'\| \big) \, \|x - x'\|,$$

2. *for all* $x \in \operatorname{int} \mathcal{X}$, $\eta \in \mathbb{R}$, *and any* $\mathbf{j} \in \{1, \dots, M\}^B$

$$\eta \, \partial_x \overline{\mathcal{L}}_p(x, \lambda; \mathbf{j}) \;\subseteq\; \eta \, \partial_x \overline{\mathcal{L}}_{p'}(x, \lambda'; \mathbf{j}) \;+\; |\eta| \, C_{\mathcal{X}} \big(\|\lambda - \lambda'\| + \|p - p'\| \big) \, \mathbb{B}.$$

Specifically, item 1 gives $\operatorname{Lip}_{\mathcal{X}} \big(r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')} \big) \le C_{\mathcal{X}} \big(\|\lambda - \lambda'\| + \|p - p'\| \big)$.

Proof. Step 1 (Lipschitz constant). Write $r_{\mathbf{j}} = r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')}$. Since the objective terms cancel, the fundamental theorem of calculus yields

$$\begin{aligned} r_{\mathbf{j}}(x) - r_{\mathbf{j}}(x') = &\sum_{i=1}^{m} (\lambda_i - \lambda_i') \int_{\bar{g}_i(x';\mathbf{j})}^{\bar{g}_i(x;\mathbf{j})} \varphi'(\tau / p_i) \, d\tau \\ &+ \sum_{i=1}^{m} \lambda_i' \int_{\bar{g}_i(x';\mathbf{j})}^{\bar{g}_i(x;\mathbf{j})} \left[\varphi'(\tau / p_i) - \varphi'(\tau / p_i') \right] d\tau. \end{aligned}$$

For every τ in these intervals, $|\tau| \le G_{\max}$ and $\tau / p_i, \tau / p_i' \in T$. Hence

$$|\varphi'(\tau / p_i)| \le S_1.$$

Since φ' is S_2-Lipschitz on T,

$$\left| \varphi'(\tau / p_i) - \varphi'(\tau / p_i') \right| \le S_2 \left| \frac{\tau}{p_i} - \frac{\tau}{p_i'} \right| = S_2 \, |\tau| \, \frac{|p_i - p_i'|}{p_i p_i'} \le S_2 \, \frac{G_{\max}}{\ell_c^2} \, |p_i - p_i'|.$$

Moreover, by

$$|\bar{g}_i(x; \mathbf{j}) - \bar{g}_i(x'; \mathbf{j})| \le L_g \|x - x'\|,$$

and $\lambda_i' \le \Lambda$, we obtain

$$|r_j(x) - r_j(x')| \le L_g \left(S_1 \sum_{i=1}^{m} |\lambda_i - \lambda_i'| + \frac{\Lambda S_2 G_{\max}}{\ell_c^2} \sum_{i=1}^{m} |p_i - p_i'| \right) \|x - x'\|.$$

Using $\sum_i |a_i| \le \sqrt{m} \, \|a\|$ (holds by Cauchy-Schwarz) applied to $a = \lambda - \lambda'$ and $a = p - p'$ separately gives the result with

$$C_{\mathcal{X}} = \sqrt{m} \, L_g \max \left\{ S_1, \frac{\Lambda S_2 G_{\max}}{\ell_c^2} \right\}.$$

Step 2 (Subdifferential inclusion). Write $\overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j}) = \overline{\mathcal{L}}_{p'}(\cdot, \lambda'; \mathbf{j}) + r_{\mathbf{j}}^{(\lambda,p),(\lambda',p')}(\cdot)$, with residual $r_{\mathbf{j}}$ defined as in *Step 1*. The Clarke sum rules (Clason & Valkonen, 2026, Theorem 13.10, Theorem 13.20) give for any $\eta \in \mathbb{R}$

$$\eta \, \partial_x \overline{\mathcal{L}}_p(x, \lambda; \mathbf{j}) \subseteq \eta \, \partial_x \overline{\mathcal{L}}_{p'}(x, \lambda'; \mathbf{j}), + \partial \big(\eta \, r_{\mathbf{j}}(x) \big).$$

Since $x \in \operatorname{int} \mathcal{X}$ (Corollary 2), there is $\varepsilon > 0$ with $\varepsilon \mathbb{B} + x \subseteq \mathcal{X}$, so ηr_j is Lipschitz near x with constant $|\eta| \operatorname{Lip}_{\mathcal{X}}(r_{\mathbf{j}})$. (Clason & Valkonen, 2026, Lemma 13.2) gives that the subdifferential is bounded by $|\eta| \operatorname{Lip}_{\mathcal{X}}(r_{\mathbf{j}})$. *Step 1* bounds $\operatorname{Lip}_{\mathcal{X}}(r_{\mathbf{j}})$ by $C_{\mathcal{X}}(\|\lambda - \lambda'\| + \|p - p'\|)$ which completes the proof. □

Lemma 20 (Asymptotic consistency of the search directions)**.** *Let Assumption 1 and Assumption 2 hold, let summability properties as stated in Lemma 17 hold, let $\hat{\lambda}, \hat{p}$ be the limits defined by Lemma 17, and set $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as*

$$D(x) := \mathcal{D}_{\hat{p}}(x, \hat{\lambda}). \tag{41}$$

Then, almost surely,

$$\lim_{k\to\infty} \operatorname{dist}\big(y^k,\, D(x^k)\big) = 0.$$

Proof. Fix path-wise some ω on the almost sure event on which Assumption 1, Lemma 17, Lemma 18, Lemma 13 simultaneously hold; being a finite intersection of events of full measure, it has also full measure. Fix k.

Step 1 (splitting off the proximal term). By Lemma 13 there is $w^k \in \mathcal{D}_{p^k}(x^k, \lambda^k)$ with $y^k = w^k - \mu\,(x^k - s^k)$. Hence, for every $t \in D(x^k)$,

$$\|y^k - t\| \;=\; \|(w^k - \mu\,(x^k - s^k)) - t\| \;\le\; \|w^k - t\| + \mu\,\|x^k - s^k\|,$$

and taking the infimum over $t \in D(x^k)$,

$$\operatorname{dist}\big(y^k, D(x^k)\big) \;\le\; \operatorname{dist}\big(w^k, D(x^k)\big) + \mu\,\|x^k - s^k\|. \tag{42}$$

Step 2 (bounding the conservative fields). Write $K := C_{\mathcal{X}}\big(\|\lambda - \lambda'\| + \|p - p'\|\big)$ and, for each $\mathbf{j}$, $r_{\mathbf{j}} := \overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j}) - \overline{\mathcal{L}}_{p'}(\cdot, \lambda'; \mathbf{j})$, so that $\overline{\mathcal{L}}_p(\cdot, \lambda; \mathbf{j}) = \overline{\mathcal{L}}_{p'}(\cdot, \lambda'; \mathbf{j}) + r_{\mathbf{j}}$ as in Lemma 19. Since $x \in \operatorname{int} \mathcal{X}$, by Lemma 19 (2) with $\eta = -1$, it holds

$$-\partial_x \overline{\mathcal{L}}_p(x, \lambda; \mathbf{j}) \;\subseteq\; -\partial_x \overline{\mathcal{L}}_{p'}(x, \lambda'; \mathbf{j}) + K\,\mathbb{B} \qquad \text{for every } \mathbf{j}.$$

Furthermore, Minkowski summation preserves inclusions (Lemma 2), hence

$$\mathcal{D}_p(x, \lambda) = \frac{1}{M^B} \sum_{\mathbf{j}\in\{1,\dots,M\}^B} \big(-\partial_x \overline{\mathcal{L}}_p(x, \lambda; \mathbf{j})\big) \;\subseteq$$
$$\frac{1}{M^B} \sum_{\mathbf{j}\in\{1,\dots,M\}^B} \Big(-\partial_x \overline{\mathcal{L}}_{p'}(x, \lambda'; \mathbf{j}) + K\,\mathbb{B}\Big) = \mathcal{D}_{p'}(x, \lambda') + K\,\mathbb{B},$$

where the last equality uses $\frac{1}{M^B}\sum_{j=1}^{M^B} K\,\mathbb{B} = K\,\mathbb{B}$, valid since $\mathbb{B}$ is convex (see Lemma 3).

Step 3. Therefore by setting (λ, p) as (λ^k, p^k), and (λ', p') as $(\hat{\lambda}, \hat{p})$, we bound the distance using inequality (42) as

$$\operatorname{dist}\big(y^k, D(x^k)\big) \le K + \mu\|x^k - s^k\| = C_{\mathcal{X}}\big(\|\lambda^k - \hat{\lambda}\| + \|p^k - \hat{p}\|\big) + \mu\|x^k - s^k\| \longrightarrow 0 \quad a.s.$$

□

Lemma 21 (A.5 Davis et al. (2018))**.** *Let Assumption 1, Assumption 2 hold. Let the hypotheses of Lemma 20 hold and let D and $G := D|_{\mathcal{X}}$ be as in (41). Then, almost surely, for every increasing unbounded sequence $\{k_j\} \subset \mathbb{N}$ such that $x^{k_j} \to \hat{x}$ for some $\hat{x} \in \mathbb{R}^n$,*

$$\lim_{N\to\infty} \operatorname{dist}\Big(\frac{1}{N}\sum_{j=1}^{N} y^{k_j},\; G(\hat{x})\Big) = 0.$$

Proof. Let Ω_0 be the almost-sure event on which Lemma 20 holds. Fix $\omega \in \Omega_0$ and a sequence $\{k_j\}$ together with its limit $\hat{x} \in \mathcal{X}$ and the iterates $x^{k_j} \in \operatorname{int} \mathcal{X}$ (by Corollary 2). Therefore, $G(x^{k_j}) = \mathcal{D}_{\hat{p}}(x^{k_j}, \hat{\lambda})$ for any k_j and also $G(\hat{x}) = \mathcal{D}_{\hat{p}}(\hat{x}, \hat{\lambda})$, in other words, the two set-valued maps coincide in our setting, hence we prove the property for $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$ which gives that the property also holds for G.

Step 1. Write $D := \mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$. The set-valued map $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is nonempty, convex, compact and upper-semicontinuous by Lemma 13. From compactness and upper-semicontinuity we get: for every $\varepsilon/2 > 0$ there is $\varrho > 0$ such that

$$\|x - \hat{x}\| \le \varrho \quad\Longrightarrow\quad D(x) \subseteq D(\hat{x}) + \tfrac{\varepsilon}{2}\mathbb{B}. \tag{43}$$

Fix $\varepsilon > 0$ and choose ϱ as in (43). Since both

$$x^{k_j} \;\to\; \hat{x}$$

and also, by Lemma 20,

$$\operatorname{dist}(y^{k_j}, D(x^{k_j})) \;\to\; 0,$$

there exists J such that for all $j \geq J$ both $\|x^{k_j} - \hat{x}\| \leq \varrho$ and $\operatorname{dist}\big(y^{k_j}, D(x^{k_j})\big) \leq \varepsilon/2$. For such j, (43) gives

$$b^j := \operatorname{dist}\big(y^{k_j}, D(\hat{x})\big) \leq \operatorname{dist}\big(y^{k_j}, D(x^{k_j})\big) + \tfrac{\varepsilon}{2} \leq \varepsilon.$$

As $\varepsilon > 0$ was arbitrary, $b^j \to 0$.

Step 2 (averaging). The set $D(\hat{x})$ is nonempty and compact (Lemma 13), so for each j there is $v^j \in D(\hat{x})$ with $\|y^{k_j} - v^j\| = b^j$. Because $D(\hat{x})$ is *convex*, the average $\bar{v}^N := \frac{1}{N}\sum_{j=1}^N v^j$ again lies in $D(\hat{x})$, and therefore

$$\operatorname{dist}\Big(\frac{1}{N}\sum_{j=1}^N y^{k_j},\, D(\hat{x})\Big) \;\leq\; \Big\|\frac{1}{N}\sum_{j=1}^N \big(y^{k_j}\big) - \bar{v}^N\Big\| \;=\; \Big\|\frac{1}{N}\sum_{j=1}^N \big(y^{k_j} - v^j\big)\Big\| \;\leq\; \frac{1}{N}\sum_{j=1}^N b^j. \quad (44)$$

It remains to observe that $b^j \to 0$ forces its Cesàro averages (44) to vanish, *irrespective of whether the series* $\sum_j b^j$ *converges*. Let us prove the convergence of its Cesàro average to 0. Pick an arbitrary $\varepsilon > 0$. Find J such that $b^j < \frac{\varepsilon}{2}$ for all $j > J$. Then, we get for $N > J$

$$\frac{1}{N}\sum_{j=1}^N b^j = \frac{1}{N}\sum_{j=1}^J b^j + \frac{1}{N}\sum_{j=J+1}^N b^j < \frac{1}{N}\sum_{j=1}^J b^j + \frac{\varepsilon}{2}.$$

Now fix N' such that both $N' > J$ and $\frac{1}{N'}\sum_{j=1}^J b^j < \frac{\varepsilon}{2}$ hold. This is always possible since the sum $\sum_{j=1}^J b^j$ is a finite number and N' can be arbitrarily large. Therefore,

$$\frac{1}{N}\sum_{j=1}^N b^j < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \quad \forall N > N',$$

and ε was arbitrary. This completes the proof that the Cesàro average (44) has a 0 limit with $b^j \to 0$. This also finishes the proof of lemma as the sequence $\{k^j\}$ was arbitrary and $\omega \in \Omega_0$. □

E.3.4 Proving Boundedness from Below, Descent and Morse-Sard Properties of a Potential Function

To construct a Lyapunov function, we need the boundedness from below on the entire domain. This is approached in the next lemma.

Lemma 22 (Continuous, bounded-below Lyapunov function)**.** *Fix $\lambda \in [0, \Lambda]^m$ and $p \in [\ell_c, u_c]^m$, set*

$$c(\lambda, p) := \min_{x \in \mathcal{X}} \mathcal{L}_p^{\mathrm{emp}}(x, \lambda), \qquad \Phi_p(x, \lambda) := \max\big\{\mathcal{L}_p^{\mathrm{emp}}(x, \lambda),\, c(\lambda, p)\big\}.$$

Then

1. *$c(\lambda, p) \in \mathbb{R}$;*
2. *$\Phi_p(\cdot, \lambda)$ is locally Lipschitz, hence continuous, on $\mathbb{R}^n$, and satisfies $\Phi_p(\cdot, \lambda) \geq c(\lambda, p)$, hence is bounded from below on $\mathbb{R}^n$;*
3. *$\Phi_p(x, \lambda) = \mathcal{L}_p^{\mathrm{emp}}(x, \lambda)$ for every $x \in \mathcal{X}$.*

Proof. By Lemma 12(1), $\mathcal{L}_p^{\mathrm{emp}}(\cdot, \lambda)$ is locally Lipschitz on $\mathbb{R}^n$, hence continuous. Since $\mathcal{X}$ is compact and nonempty, the extreme value theorem gives that the minimum defining $c(\lambda, p)$ is finite and attained, proving item 1.

For item 2, a maximum of two locally Lipschitz functions is locally Lipschitz, and a constant is locally Lipschitz, so $\Phi_p(\cdot, \lambda)$ is locally Lipschitz and therefore continuous. By definition of the maximum, $\Phi_p(x, \lambda) \geq c(\lambda, p)$ for every $x \in \mathbb{R}^n$, so $\Phi_p(\cdot, \lambda)$ is bounded from below on $\mathbb{R}^n$.

For item 3, let $x \in \mathcal{X}$. By definition of $c(\lambda, p)$ as the minimum over $\mathcal{X}$ we have $\mathcal{L}_p^{\mathrm{emp}}(x, \lambda) \geq c(\lambda, p)$, and hence $\Phi_p(x, \lambda) = \mathcal{L}_p^{\mathrm{emp}}(x, \lambda)$. □

Next, we prove the Weak-Sard property of the potential function Φ together with the conservative field $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$.

Lemma 23 (Weak-Sard, condition B.1 in Davis et al. (2018))**.** *Let Assumption 1 and Assumption 2 hold. Set* $G : \mathcal{X} \rightrightarrows \mathbb{R}^n$

$$G(x) = \left(\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})|_{\mathcal{X}}\right)(x)$$

and $\Phi(\cdot) := \Phi_{\hat{p}}(\cdot, \hat{\lambda})$ *defined in Lemma 22. Then for a dense set* V *of values* $v \in \mathbb{R}$*, the intersection of*

$$\Phi^{-1}(v) \cap G^{-1}(0)$$

is empty.

Proof. Since $\partial(-f) = -\partial f$ for a Clarke subdifferential (Clason & Valkonen, 2026, Theorem 13.10), and $\sum_j (-A_j) = -\sum_j A_j$ for Minkowski sums, $\mathcal{D}_{\hat{p}}(x, \hat{\lambda})$ is the pointwise negative of the conservative field $-\mathcal{D}_{\hat{p}}(x, \hat{\lambda})$. Since $0 \in A \iff 0 \in -A$, the two critical sets coincide:

$$\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})^{-1}(0) = \left(-\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})\right)^{-1}(0).$$

Step 1. (Apply Morse-Sard Theorem and construct a candidate for V)
By Lemma 12(1), we know that $\mathcal{L}^{\text{emp}}_{\hat{p}}(x, \hat{\lambda})$ is definable, and locally Lipschitz. By Lemma 13 and Corollary 3, $-\mathcal{D}_{\hat{p}}(x, \hat{\lambda})$ is a definable conservative field of $\mathcal{L}^{\text{emp}}_{\hat{p}}(x, \hat{\lambda})$. Combining these, Theorem 2 (Bolte & Pauwels, 2021, Theorem 5) gives that the set

$$\text{CritL} := \left\{ \mathcal{L}^{\text{emp}}_{\hat{p}}(x, \hat{\lambda}) \ : \ x \in \mathbb{R}^n, \ 0 \in -\mathcal{D}_{\hat{p}}(x, \hat{\lambda}) \right\}$$

is finite. Fix $c := \min_{x \in \mathcal{X}} \mathcal{L}^{\text{emp}}_{\hat{p}}(x, \hat{\lambda})$ as in Lemma 22, and define $V = \mathbb{R} \setminus (\text{CritL} \cup \{c\})$. We argue that

$$\Phi^{-1}(v) \cap \left(-\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})\right)^{-1}(0) = \emptyset. \tag{45}$$

Step 2. (Proving Eq. (45))
Indeed, if we write down the definition of $\Phi^{-1}(v)$

$$\Phi^{-1}(v) = \begin{cases} \mathcal{L}^{\text{emp}}_{\hat{p}}(\cdot, \hat{\lambda})^{-1}(v), \text{if } v > c, \\ \{x : \mathcal{L}^{\text{emp}}_{\hat{p}}(x, \hat{\lambda}) \leq c\}, \text{if } v = c, \\ \emptyset, \text{if } v < c. \end{cases}$$

Consider an arbitrary $v \in V$. Let us assume first *(i)* $v > c$, then if $\mathcal{L}^{\text{emp}}_{\hat{p}}(\cdot, \hat{\lambda})^{-1}(v) \cap \left(-\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})\right)^{-1}(0) \neq \emptyset$, that means that $v \in \text{CritL}$ which is a contradiction with the construction of V. Next *(ii)* assume that $v = c$ which is not possible by construction of V. Lastly, *(iii)* $v < c$, then $\Phi^{-1}(v) = \emptyset$, hence trivially the equality is true.

Step 3. (Proving density of V)
Next, we prove that V is dense. This is the same as proving that the interior of the complement is empty, i.e. the interior of

$$\mathbb{R} \setminus V = \mathbb{R} \setminus (\mathbb{R} \setminus (\text{CritL} \cup \{c\})) = \text{CritL} \cup \{c\}.$$

Since CritL is finite, $\text{CritL} \cup \{c\}$ is also finite, hence has empty interior.

Step 4. (proving Weak-Sard for G)
Finally, we show that the property also holds for G and not only for $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$. Since G is its restricted version, for any $q \in \mathbb{R}^n$

$$G^{-1}(q) = \{x \in \mathbb{R}^n \mid q \in \mathcal{D}_{\hat{p}}(x, \hat{\lambda}), x \in \mathcal{X}\} \subseteq \{x \in \mathbb{R}^n \mid q \in \mathcal{D}_{\hat{p}}(x, \hat{\lambda})\} = \mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})^{-1}(q).$$

But then by (45)

$$\Phi^{-1}(v) \cap G^{-1}(0) \subseteq \Phi^{-1}(v) \cap \mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})^{-1}(0) = \Phi^{-1}(v) \cap \left(-\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})\right)^{-1}(0) = \emptyset,$$

where the second last equality was proven in the beginning. □

Lemma 24 (Descent, condition B.2 in Davis et al. (2018)). *Let Assumption 1 and Assumption 2 hold and fix a realisation of $(\hat{\lambda}, \hat{p})$. Let $x : \mathbb{R}_+ \to \mathcal{X}$ be absolutely continuous with*

$$\dot{x}(t) \in \left(\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})|_{\mathcal{X}}\right)(x(t)) = G\big(x(t)\big) \qquad \text{for a.e. } t \geq 0.$$

Then for a.e. $t \geq 0$ the velocity is the minimal-norm element,

$$\|\dot{x}(t)\| = \operatorname{dist}\Big(0, G\big(x(t)\big)\Big),$$

and for every $T \geq 0$

$$\Phi_{\hat{p}}\big(x(T), \hat{\lambda}\big) - \Phi_{\hat{p}}\big(x(0), \hat{\lambda}\big) \;=\; -\int_0^T \operatorname{dist}^2\Big(0, G\big(x(t)\big)\Big)\, dt \;\leq\; 0.$$

In particular, if $0 \notin G\big(x(0)\big)$ there exists $T > 0$ with

$$\Phi_{\hat{p}}\big(x(T), \hat{\lambda}\big) \;<\; \sup_{t \in [0,T]} \Phi_{\hat{p}}\big(x(t), \hat{\lambda}\big) \;\leq\; \Phi_{\hat{p}}\big(x(0), \hat{\lambda}\big),$$

i.e. condition B.2 of Davis et al. (2018) holds for the pair $\big(\Phi_{\hat{p}}(\cdot, \hat{\lambda}),\, G(\cdot)\big)$.

Proof. Write $\Phi := \Phi_{\hat{p}}(\cdot, \hat{\lambda})$. By a property of a conservative field $\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})$ of a potential function $-\mathcal{L}^{\text{emp}}_{\hat{p}}(\cdot, \hat{\lambda})$, we apply Lemma 6 (Bolte & Pauwels, 2021, Lemma 2) to the absolutely continuous curve x, for a.e. t to get

$$\tfrac{d}{dt}\big[-\mathcal{L}^{\text{emp}}_{\hat{p}}(x(t), \hat{\lambda})\big] = \langle v, \dot{x}(t)\rangle \qquad \text{for } \textit{all } v \in \mathcal{D}_{\hat{p}}(x(t), \hat{\lambda}).$$

Since the absolutely continuous curve x is a solution to a DI defined through G, $x(t) \in \mathcal{X}$ must hold, therefore $\left(\mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})|_{\mathcal{X}}\right)(x(t)) = \mathcal{D}_{\hat{p}}(x(t), \hat{\lambda})$ for all t. Hence it must also hold that (note the restriction)

$$\tfrac{d}{dt}\big[-\mathcal{L}^{\text{emp}}_{\hat{p}}(x(t), \hat{\lambda})\big] = \langle v, \dot{x}(t)\rangle \qquad \text{for } \textit{all } v \in G\big(x(t)\big). \tag{46}$$

By using Lemma 7 with $G = \mathcal{D}_{\hat{p}}(\cdot, \hat{\lambda})|_{\mathcal{X}}$ and (46), we get that $\|\dot{x}(t)\| = \operatorname{dist}\Big(0, G\big(x(t)\big)\Big)$, and $\tfrac{d}{dt}\big[-\mathcal{L}^{\text{emp}}_{\hat{p}}(x(t), \hat{\lambda})\big] = \|\dot{x}(t)\|^2$, hence

$$\tfrac{d}{dt}\big[\mathcal{L}^{\text{emp}}_{\hat{p}}(x(t), \hat{\lambda})\big] = -\|\dot{x}(t)\|^2 = -\operatorname{dist}^2\Big(0, G\big(x(t)\big)\Big).$$

Since $\mathcal{L}^{\text{emp}}_{\hat{p}}(\cdot, \hat{\lambda})$ is locally Lipschitz, and x is absolutely continuous, the composition $\mathcal{L}^{\text{emp}}_{\hat{p}}(\cdot, \hat{\lambda}) \circ x$ is also absolutely continuous. Therefore, by integrating both sides of the equation, we get that

$$\mathcal{L}^{\text{emp}}_{\hat{p}}(x(T), \hat{\lambda}) - \mathcal{L}^{\text{emp}}_{\hat{p}}(x(0), \hat{\lambda}) = -\int_0^T \operatorname{dist}^2\Big(0, G\big(x(t)\big)\Big)\, dt \;\leq\; 0.$$

Since $x(t) \in \mathcal{X}$ holds, we have $\Phi\big(x(t)\big) = \mathcal{L}^{\text{emp}}_{\hat{p}}(x(t), \hat{\lambda})$ for all $t \geq 0$ by Lemma 22. Hence, we conclude that

$$\Phi\big(x(T)\big) - \Phi\big(x(0)\big) = -\int_0^T \operatorname{dist}^2\Big(0, G\big(x(t)\big)\Big)\, dt \;\leq\; 0.$$

Now assume that $0 \notin G\big(x(0)\big)$. Then there must exist $T > 0$ such that for any $t \in [0, T]$ it holds that $\operatorname{dist}\Big(0, G\big(x(t)\big)\Big) > 0$: in particular,

$$-\int_0^T \operatorname{dist}^2\Big(0, G\big(x(t)\big)\Big) < 0.$$

Suppose this is not true. Then there must exist a sequence $t_j \downarrow 0$, such that $v^j \in G\big(x(t_j)\big)$ and $v^j = 0$. Since this holds for G, then it also must hold for $\mathcal{D}_{\hat{p}}(x(t_j), \hat{\lambda})$. As x is continuous, we have that $x(t_j) \to x(0)$. By the outer-semicontinuity, we have that $0 \in D_{\hat{p}}(x(0), \hat{\lambda}) = G\big(x(0)\big)$ which is a contradiction.

Hence $t \mapsto \Phi\big(x(t)\big)$ is descending, so it must hold that $\Phi\big(x(T)\big) < \Phi\big(x(0)\big)$. Finally by monotonicity $\sup_{t \in [0,T]} \Phi\big(x(t)\big) = \Phi\big(x(0)\big)$. □

E.4 Counter Example - No subgradient-type method escapes the flat region.

In this section, we provide a simple counterexample that shows non-convergence of SPBM and also of the state-of-the-art methods for constrained machine learning (SSw, SSL-ALM). Both the objective and the constraint functions are made up of standard machine learning building blocks.

Consider a single-unit ReLU model $\hat{y} := [wz + b]_+$ with parameters (w, b) and a dataset of 1000 feature–label pairs (z_i, y_i) with $z_i = 1$ for all $i \in \{0, \dots, 999\}$, $y_i = 1$ for $i \in \{0, \dots, 99\}$, and $y_i = 0$ for $i \in \{100, \dots, 999\}$. Assume the preactivation $x := w + b$. We set the objective function f as a mean squared error:

$$f(x) = \frac{1}{1000} \sum_{i=0}^{999} (\hat{y}_i - y_i)^2$$
$$= 0.1\big([x]_+ - 1\big)^2 + 0.9\,[x]_+^2 = \begin{cases} x^2 - 0.2\,x + 0.1, & \text{if } x \geq 0, \\ 0.1, & \text{if } x < 0. \end{cases}$$

We assume a single constraint g – we bound the mean squared error on subgroup $\mathcal{R}$ (15 positive, 5 negative samples) within $\varepsilon := 0.25$:

$$g(x) = \frac{1}{|\mathcal{R}|} \sum_{(z_i, y_i) \in \mathcal{R}} \big([x]_+ - y_i\big)^2 - \varepsilon = \begin{cases} x^2 - 1.5\,x + 0.5, & \text{if } x \geq 0, \\ 0.5, & \text{if } x < 0. \end{cases}$$

The mixed labels in $\mathcal{R}$ make the constraint estimate $\bar{g}$ genuinely stochastic. For $x < 0$, and mini-batch size $B = 1$; the single-per-sample constraint estimate is $\bar{g}(x) = y_i^2 - \varepsilon \in \{-0.25,\, 0.75\}$. By estimating the constraint on the subgroup $\mathcal{R}$, $\bar{g}$ follows an ε-shifted Bernoulli distribution with a parameter $q := 0.75$, its moments read

$$\mathbb{E}\big[\bar{g}(x, \xi)\big] = q - \varepsilon = 0.5, \qquad \mathrm{Var}\,\big(\bar{g}(x, \xi)\big) = q(1 - q) = \tfrac{3}{16}.$$

We show later in Lemma 25 that the per-sample *gradients* are zero (the value of the mini-batched constraint $\bar{g}$ does not depend on w, b in any way): the method receives a noisy signal that the constraint is violated, but the gradient is completely flat. Same applies to the function estimate $\bar{f}$: for $x < 0$, we get a constant value independent of x.

Fig. 10 illustrates the problem. The objective, constraint, and Lagrangian functions are completely flat when $x < 0$ no matter the values of the dual parameters (λ, p). This leads to the primal minimization methods being stuck when the initial iterates w, b are chosen such that $x = w + b < 0$.

Let us prove some properties of this counter example.

Proposition 4. *The problem has a unique solution in $x^* = 0.5$, with $f(x^*) = 0.25$; it satisfies KKT conditions. The feasible set is $[0.5, 1]$, hence is non-empty and compact.*

Proof. For $x < 0$ we have $g \equiv 0.5 > 0$, so no such point is feasible. By computing the roots of $g(x)$ for $x \geq 0$, we get that the feasible set is $[0.5, 1]$, hence non-empty and compact. The constrained minimiser is the unique point $x^\star = 0.5$, with $f(x^\star) = 0.25$. Both f and g are continuously differentiable on a neighbourhood of $x^\star$. From the KKT stationarity condition, we find the multiplier $\lambda^* = \frac{8}{5}$, which gives dual feasibility and active constraint gives complementarity, therefore KKT holds. □

Proposition 5. *The functions g, f are **not** weakly-convex.*

Proof. Let us start with g. Reminding the definition of weak-convexity in Definition 8, we can fix an arbitrary $0 \leq C < \infty$. Now fix the convex combination $\alpha = 0.5$. We argue that for the fixed C, we can choose x_1, x_2 that breaks the convexity. Indeed, assume $x_1 = -x_2$ and $x_2 > 0$. That gives for $\hat{g} := g + \frac{C}{2}(\cdot)^2$

$$\hat{g}(\frac{1}{2}x_1 + \frac{1}{2}x_2) = \hat{g}(0) = \frac{1}{2}. \tag{47}$$

While also

$$\frac{1}{2}\hat{g}(x_1) + \frac{1}{2}\hat{g}(x_2) = x_2^2(\frac{1 + C}{2}) - \frac{3}{4}x_2 + \frac{1}{2}. \tag{48}$$

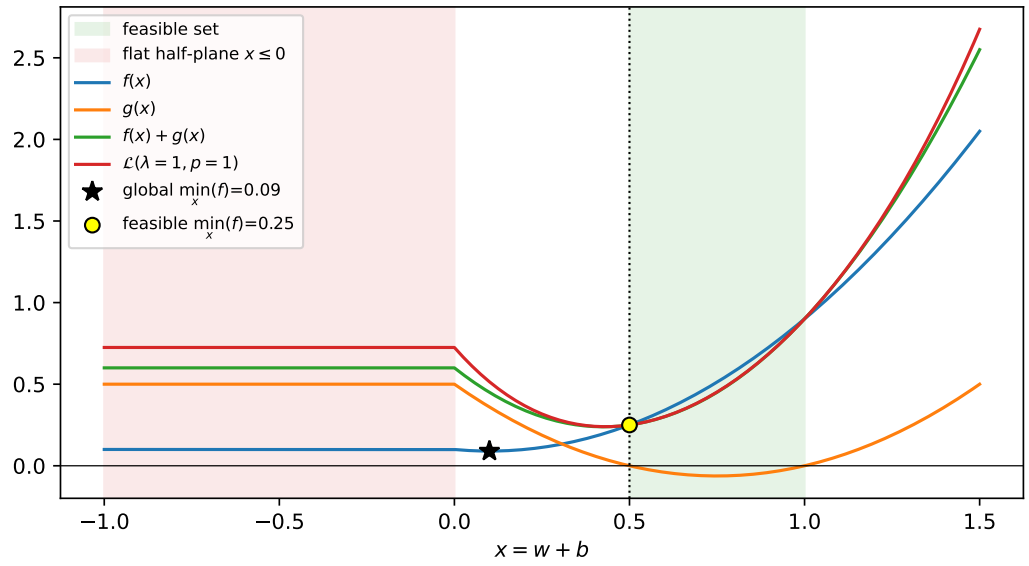


Figure 10: The counterexample problem. The objective function f (blue) and the constraint function g (orange) plotted against the *pre-activation* $x = w + b$. The feasible region is shaded green, and the "flat" region is shaded red. A classical Lagrangian $f(x) + \lambda g(x)$ with $\lambda = 1$ is plotted in green, and the SPBM Lagrangian with $\lambda = 1$, $p = 1$, and the Quadratic-Logarithmic penalty-barrier function is plotted in red.

Computing roots, we get that for any

$$x_2 \in (0, \frac{3}{2 + 2C})$$

it holds that Eq. (47) > Eq. (48), hence $\hat{g}$ is not convex for any finite C, therefore not weakly-convex. The same arguments apply to f with $x_2 \in (0, \frac{1}{5(1+C)})$, which completes the proof. □

Remark 8. *The problem of Section E.4 is a single-ReLU ERM instance, arguably the simplest neural-network model there is, yet by Proposition 5 neither its objective nor its constraint is weakly convex. The guarantees of methods that assume weak convexity, such as SSw and SSL-ALM, therefore do not apply to it and the methods fail here. This raises the question of which assumptions are broad enough to cover neural-network ERM problems of this kind. We suggest local Lipschitzness and definability, assumed in our paper and verified for this instance in Proposition 6.*

Proposition 6. *The problem defined through f, g in Section E.4 is an instance of Eq. (1).*

Proof. Let $B \in \mathbb{N}$ be a batch size, let g, f be functions defined in Section E.4. To show that the problem from Section E.4 is an instance of Eq. (1), we prove:

1. f, g are expectations of $\bar{f}, \bar{g}$,
2. $\bar{f}, \bar{g}$ are locally Lipschitz,
3. $\bar{f}, \bar{g}$ are definable.

(1) Averaging the constraint over $\mathcal{R}$ gives $\mathbb{E}_\xi[g(x, \xi)] = q([x]_+ - 1)^2 + (1 - q)[x]_+^2 - \varepsilon = g(x)$ with $q = 15/20$. Since $z_i \equiv 1$, the objective depends on a sample only through its label, and both labels occur in $\mathcal{R}$. Writing $\rho_1 = 0.1$, $\rho_0 = 0.9$ for the label frequencies in the full dataset, the weights $w(1) = \rho_1/q = 2/15$ and $w(0) = \rho_0/(1 - q) = 18/5$ give $f(x, \xi) := w(y(\xi))([x]_+ - y(\xi))^2$ with $\mathbb{E}_\xi[f(x, \xi)] = 0.1([x]_+ - 1)^2 + 0.9[x]_+^2 = f(x)$. Therefore, by sampling from $\mathrm{Unif}(\mathcal{R})$ and by objective reweighting, we get that both $\bar{f}, \bar{g}$ are unbiased estimators of f, g.

(2), (3) Both $\bar{g}$ and $\bar{f}$ are built from the ReLU $[\cdot]_+$ (semialgebraic and 1-Lipschitz), subtraction of a constant, squaring, multiplication by the fixed weights $w(\cdot)$, and averaging over the batch. Each of

these is locally Lipschitz and definable in $\mathcal{O}$, and both properties are preserved under composition and finite sums □

Lastly, we show that no method, considered in the Section 4, is able to converge to the feasible region in the case of $x^0 < 0$. We start with the following lemma.

Lemma 25. *Let $B \in \mathbb{N}$, and let $(w, b) \in \mathbb{R}^2$ satisfy $x := w + b < 0$. Then for every realisation ω, every $\lambda \in \mathbb{R}^m_+$ and every $p \in \mathbb{R}^m_{++}$,*

$$\partial_{(w,b)}\, \bar{f} = \partial_{(w,b)}\, \bar{g} = \partial_{(w,b)}\, \bar{\mathcal{L}}_p(\cdot, \lambda) = \{0\},$$

so any subgradient oracle returns $\overline{\nabla}\mathcal{L}_p = 0$. If in addition $s = (w, b)$, then the direction of line 4 of Algorithm 1 (identically Algorithm 2) satisfies $\bar{y} = 0$.

Proof. Since $z_i \equiv 1$, every model output is $[w + b]_+$, which vanishes on the open set $U := \{(w, b) : w + b < 0\}$. Hence $\bar{f}$ and $\bar{g}$ are constant on U for every realisation, and so is $\bar{\mathcal{L}}_p(\cdot, \lambda) = \bar{f} + \sum_i \lambda_i p_i \, \varphi(\bar{g}_i/p_i)$ for every fixed λ, p. A function constant on an open set is continuously differentiable there with zero gradient, hence strictly differentiable and so all three Clarke subdifferentials equal $\{0\}$; from which it follows $\overline{\nabla}\mathcal{L}_p = 0$ irrespective of the realisation. Moreover if $s = (w, b)$ the proximal term $\mu\big((w, b) - s\big)$ in SPBM vanishes, so $\bar{y} = -\overline{\nabla}\mathcal{L}_p - \mu\big((w, b) - s\big) = 0$. □

Proof of Proposition 3. Let $\mu \geq 0$, $\delta \in [0, 1)$, and let $(w^0, b^0) \in \mathbb{R}^2$ with $x^0 := w^0 + b^0 < 0$.

Step 1 (SPBM). By initialisation $s^0 = (w^0, b^0)$. We show by induction that $(w^k, b^k) = s^k = (w^0, b^0)$ for all k. The base case holds. Assume $(w^k, b^k) = s^k = (w^0, b^0)$; since $w^k + b^k = x^0 < 0$, Lemma 25 gives $\bar{y}^k = 0 \in \mathbb{R}^2$, hence $(w^{k+1}, b^{k+1}) = (w^k, b^k) + \alpha^k \bar{y}^k = (w^0, b^0)$ and $s^{k+1} = \delta s^k + (1 - \delta)(w^{k+1}, b^{k+1}) = (w^0, b^0)$, closing the induction.

Step 2 (SSw and SSL-ALM, and general first-order schemes). Both methods take primal steps on functions assembled from $\bar{f}$ and $\bar{g}$ alone: SSw on the constraint estimate $\bar{g}$, and SSL-ALM on the augmented Lagrangian $\bar{f} + \langle \lambda, \bar{g} \rangle + \frac{1}{2\rho}\|\bar{g}\|^2 + \mu(\,\cdot - s)$. By Lemma 25 we have $\partial_{(w,b)}\bar{f} = \partial_{(w,b)}\bar{g} = \{0\}$ on $\{w + b < 0\}$, so every such direction vanishes there, and the induction from SPBM applies to both cases.

More generally, the same holds for any scheme

$$(w^{k+1}, b^{k+1}) = (w^k, b^k) + h\big((w^k, b^k)\big)$$

whose direction satisfies $h = 0$ whenever $\partial_{(w,b)}\bar{f} = \partial_{(w,b)}\bar{g} = \{0\}$ and $s^k = (w^k, b^k)$.

Step 3 (Distance). Since $(w^k, b^k) = (w^0, b^0)$, the preactivation satisfies $x^k = w^k + b^k = x^0$ for every k. By Proposition 4 the feasible set is $[0.5, 1]$, so $\operatorname{dist}(x^k, \{g \leq 0\}) = 0.5 - x^0 = \operatorname{dist}(x^0, \{g \leq 0\})$ for all k, which can be made arbitrarily large by the choice of x^0. As the subdifferentials above are singletons, this holds for every realisation ω. □

Remark 9. *Note that both the dual variables λ^k and the penalty parameters p^k of SPBM **do** change, but their change does **not** play any role in the primal iterates, and therefore neither does it change the convergence result.*

What is proven in Proposition 3 is also supported by the numerics, as seen in Fig. 11.

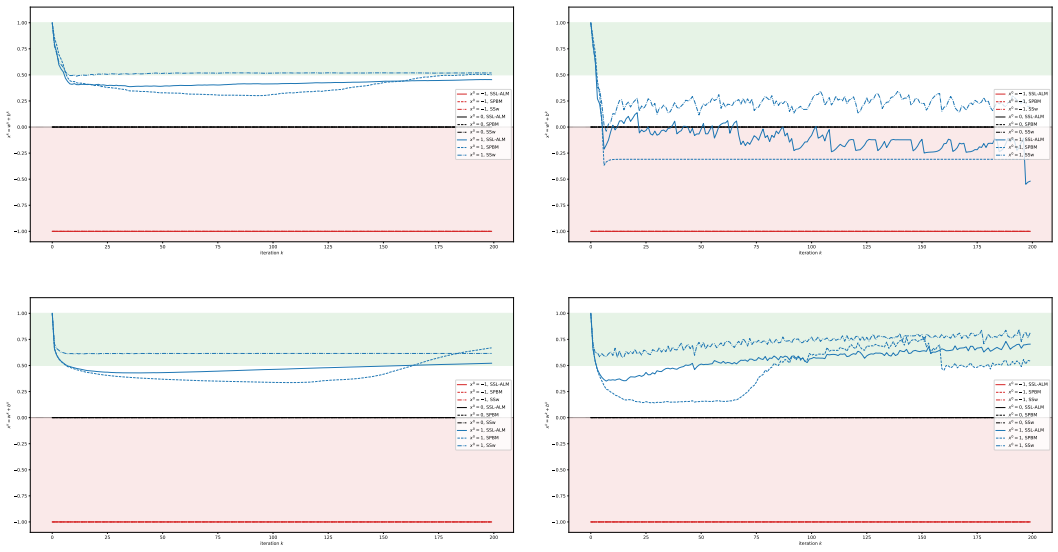


Figure 11: The mean trajectories of the iterates starting from $x^0 = -1$ (red), $x^0 = 0$ (black), and $x^0 = 1$ (blue), produced by SSw, SSL-ALM and SPBM over 10 sampling seeds in the mini-batch setting. The plots show that for $x^0 < 0$, neither of SPBM, SSL-ALM, SSw reach the feasible region $[0.5, 1]$. We fade the feasible region with a green color, the flat half plane with a red color. **Upper left**: Batch size $B = 1$ and diminishing stepsize $\alpha^0 = 0.1$, $\alpha_{k+1} = \frac{\alpha_k}{k}$. **Upper right**: Batch size $B = 1$ and constant stepsize $\alpha_k = 0.1$. **Lower left**: Batch size $B = 20$ and diminishing stepsize $\alpha^0 = 0.1$, $\alpha^{k+1} = \frac{\alpha^k}{k}$. **Lower right**: Batch size $B = 20$ and constant stepsize $\alpha_k = 0.1$.