A Convex-Analytic Reformulation for Poisoning Attacks on Logistic Regression

Carina Moreira Costa, Maria Eduarda Pinheiro, Martin Schmidt

Abstract. We study data poisoning attacks on logistic regression with ℓ_2-regularization, where an attacker chooses the features of a fixed number of injected observations with prescribed labels to maximize the logistic loss on the clean training data. Strong convexity of the defender's problem yields an exact reformulation through its first-order optimality conditions. Using convex conjugacy of the softplus function, we replace the sigmoid relations by a single convex optimality-gap inequality and express the attacker's objective as a concave function. A change of variables makes the stationarity condition affine, leaving bounded bilinear equalities as the only nonconvex constraints. We derive finite bounds on all variables involved in these equalities, including a global bound on the coefficient norm and tighter bounds obtained by preprocessing based on convex optimization. Experiments on eight binary classification datasets evaluate attack effectiveness and compare the two reformulations when solved with SCIP. Attacks obtained by local optimization degrade predictive performance on most datasets, while the convex-analytic formulation produces substantially stronger reported dual bounds within a fixed time limit.

1. Introduction

Machine learning models support automated decision-making across many fields and rely on training data, which can come from various sources, including the internet, humans, government databases, third-party sources, among others. A major concern regarding this is that data manipulations such as injection of incorrect or biased data into the training set can have severe consequences; e.g., the compromised model may develop security vulnerabilities, generate systematically biased or inaccurate predictions, and ultimately fail when used in critical applications, such as fraud detection, healthcare, etc.

One type of attack that a machine learning model can be subject to is the so-called *poisoning attack*. They occur when a malicious actor injects specially crafted attack data samples into the training data, forcing the model to learn erroneous patterns and degrading the predictive performance of the model on unseen data. In this work, we consider poisoning attacks against a logistic regression model. This attack is modeled as a bilevel optimization problem representing a game between an attacker and a defender. In this adversarial process, the attacker generates a set of poisoning points and the learner tries to learn the best predictive regression model.

The use of mathematical optimization to study machine learning models, including adversarial learning and machine teaching, is surveyed by Gambella et al. (2021). Adversarial manipulation of statistical inference is also studied by Caballero et al. (2025), who formulate attacks on conditional inference under multivariate Gaussian

2020 *Mathematics Subject Classification.* Primary 90C26; Secondary 90C25, 91A65, 62J12.

Key words and phrases. Data poisoning, Logistic regression, Bilevel optimization, Convex conjugacy, Global optimization, Bound tightening.

models and exploit their structure to develop solution methods. Their attacks modify the evidence supplied to a fixed probabilistic model. In this paper, the attacker instead injects training observations and thereby influences the parameters learned by the defender.

The formulation of poisoning attacks against some machine learning models as a bilevel optimization problem has already been considered in the literature. Poisoning attacks against support vector machines (SVMs) are studied by Biggio et al. (2012). By differentiating the KKT conditions of SVM, the method proposed in the paper iteratively computes the validation loss gradient to optimize the attack point, in order to maximize the test classification error. Xiao et al. (2015) explore the vulnerability of some feature selection methods, including LASSO, ridge regression, and elastic net, to training data poisoning. The attacker aims to increase classification error by maximizing the squared loss on the clean training or surrogate data plus a regularization term. A gradient-ascent strategy is used. Their experiments on PDF malware detection demonstrate that feature selection methods can be severely compromised under attack.

Poisoning attacks in logistic regression have also been considered, but the approach is different from the one we employ in this paper. In Mei and Zhu (2015), the idea was to mathematically formalize how an attacker can make the smallest possible modifications to the training set to force the algorithm to learn a beneficial target model for the attacker. The authors applied the idea to SVM, logistic regression, and linear regression. The authors derive single-level reformulations using KKT conditions and employ gradient descent for their attacks against SVMs and logistic regression. In the particular case of logistic regression, this framework was tested on a single dataset, the Spambase data set, and the experiment consisted of generating artificial target weights β^* to show that the optimization procedure was able to find a poisoned data set whose learned weights were close to β^*. Global optimality of the computed attack was not certified.

Carnerero-Cano et al. (2021) consider logistic regression with ℓ_2-regularization and the attack is modeled as a minimax bilevel optimization problem. To solve the problem, the authors employ a hypergradient based approach. Unlike previous approaches, in which the regularization parameter is assumed to remain fixed during an attack, the authors sought to model an optimal attack where both the poisoned data and the hyperparameters adjust dynamically under worst-case conditions. The experiments show that as the number of poisoning points increases, the learned regularization hyperparameter naturally increases to compensate for the instability the attacker attempts to introduce, but, on the other hand, it then forces the norm of the weight vector β to be closer to zero.

Steinhardt et al. (2017) derive minimax upper bounds and candidate attacks for sanitized empirical risk minimization over a constrained parameter set. Their surrogate uses the minimized contaminated loss; its test-loss interpretation additionally relies on train-test concentration and an approximation retaining all clean points during sanitization. Here, we maximize exactly the clean training loss under fixed poisoning labels and a quadratic defender penalty, without sanitization.

Koh et al. (2022), first circulated in 2018 and revised in 2021, develop influence, KKT, and minimax attacks, including logistic regression. Their KKT attack selects decoy parameters and then approximately enforces their optimality, rather than globally optimizing clean loss over defender parameters. Their two-point concentration theorem allows fractional copies and at most the original poisoning mass. It does not establish preservation of our prescribed integer class counts and exactly $n + q$ observations, so it does not by itself justify restricting our model to two poisoning locations.

Suya et al. (2021) construct model-targeted attacks with convergence to attainable targets under their assumptions and a lower bound on the number of poisoning points needed to reach a target. These guarantees concern target attainment as the budget grows, rather than global clean-loss maximization at a fixed budget.

More recently, Guedes-Ayala et al. (2025) and Şuvak et al. (2026) have studied poisoning attacks against ridge regression and linear regression, respectively. In the upper-level problem of Şuvak et al. (2026), the attacker selects the features of the poisoning samples to maximize the MSE of the regression model on the true training data. In the lower-level problem, the learner simultaneously optimizes its parameters, weights and bias, to minimize its loss function (with an ℓ_2-norm regularization term) over the poisoned training dataset. The experiments show that the proposed attack strategies are superior to a benchmark algorithm from the literature. Guedes-Ayala et al. (2025) propose a novel approach for poisoning attacks against ridge regression in which both numerical and categorical features are considered. The attack is framed as a mixed-integer bilevel optimization problem, where categorical variables are modeled as binary decisions that belong to SOS-1 sets. To solve the problem, it is first transformed into a single-level mixed-integer nonlinear optimization problem. A well-known computational bottleneck in single-level KKT reformulations is that they introduce lower-level variables that have no bounds. When solvers cannot bound these variables, their performance can be very poor. Thus, to improve solver performance, a method for bounding the lower-level variables is presented. However, even with tighter bounds, only toy examples could be solved to global optimality. The authors then propose a heuristic approach to quickly compute high-quality feasible solutions to the single-level reformulation. Finally, the authors emphasize the need to explore more robust exact approaches in the context of adversarial machine learning.

In this paper, we study poisoning attacks against ℓ_2-regularized logistic regression from a global optimization point of view. We first derive a single-level formulation of the bilevel problem using the optimality conditions of the defender's problem. KKT replacement, the softplus conjugate, and McCormick envelopes (McCormick 1976) are established ingredients. Our contribution is their integration into an exact formulation of the specified fixed-label, fixed-budget problem that combines a concave representation of the clean-loss objective, affine stationarity, bounded bilinear equalities as the only nonconvex constraints, and computable valid variable bounds. To this end, we use the convex conjugate of the softplus function to replace the sigmoid relations by a single convex inequality and introduce a change of variables that renders the stationarity condition affine. Moreover, we derive valid bounds on the classifier coefficients and the auxiliary variables, including tighter bounds obtained by convex preprocessing. These bounds allow us to construct convex relaxations of the remaining bilinear terms for spatial branch-and-bound. In our numerical study on eight binary classification datasets, we evaluate the effect of the attacks on predictive performance and compare the two single-level formulations using SCIP. The results show substantial improvements in the dual bounds reported by SCIP, while also illustrating the difficulty of certifying global optimality.

The remainder of the paper is structured as follows. In Section 2, we present the bilevel poisoning model and its reformulation using the first-order optimality conditions of the lower-level problem. The convex-analytic reformulation, the valid variable bounds, and the resulting convexity structure are discussed in Section 3. In Section 4, we present and discuss our numerical results. Finally, we close with some concluding remarks in Section 5.

2. Model

We consider a binary classification setting in which a defender trains a logistic regression model on a dataset that may have been partially corrupted by an adversary. The attacker's goal is to degrade the classifier's performance on clean data by injecting a small number of carefully crafted poisoning points into the training set.

Let $\mathcal{D}^{\text{tr}} = \{(x_i^{\text{tr}}, y_i^{\text{tr}})\}_{i=1}^n$ be the training set, with labels $y_i^{\text{tr}} \in \{0,1\}$ and features $x_i^{\text{tr}} \in [0,1]^d$. We assume throughout that all features have been scaled to $[0,1]$ as part of the preprocessing. The regularization term below is defined with respect to these scaled features. We fix $x_{i,1}^{\text{tr}} = 1$ for all i, so that the intercept is considered in the coefficient vector. A point x_i^{tr} is said to belong to the *positive class* if $y_i^{\text{tr}} = 1$, and to the *negative class* otherwise.

The logistic regression classifier assigns to each point x_i the probability

$$m_\beta(x_i) = \frac{1}{1 + e^{-\beta^\top x_i}} \in (0,1) \tag{1}$$

of belonging to the positive class, and the coefficients β are obtained by minimizing the negative log-likelihood

$$\min_{\beta \in \mathbb{R}^d} \ \mathcal{L}(\mathcal{D}^{tr}, \beta) + \lambda \|\beta\|_2^2, \tag{2}$$

where $\lambda > 0$ is a ridge penalty and

$$\mathcal{L}(\mathcal{D}^{tr}, \beta) := -\sum_{i=1}^{n} \big[y_i^{tr} \log m_\beta(x_i^{tr}) + (1 - y_i^{tr}) \log\big(1 - m_\beta(x_i^{tr})\big) \big].$$

The regularization term ensures coercivity of the lower-level objective and, as shown below, makes it 2λ-strongly convex. Consequently, the lower-level problem admits a unique minimizer.

2.1. Poisoning Attack.

The attacker constructs a set of poisoning points

$$\mathcal{D}^p = \{(x_k^p, y_k^p)\}_{k=1}^q, \quad y_k^p \in \{0,1\} \text{ fixed}, \quad x_k^p \in [0,1]^d,$$

whose labels are chosen in advance and whose features are the attacker's decision variables. As for the training data, we fix $x_{k,1}^p = 1$ for all k, so that the intercept convention is maintained. The attacker then injects $\mathcal{D}^p$ into the training set, so the defender unknowingly solves (2) on the contaminated data $\mathcal{D}^{tr} \cup \mathcal{D}^p$. Here and throughout, $\mathcal{D}^{tr} \cup \mathcal{D}^p$ denotes the combined indexed dataset of $n + q$ observations. Repeated feature-label pairs are retained and contribute separately to the loss.

The attacker's objective is to maximize the loss that the resulting classifier incurs on the clean training data. This yields the bilevel optimization problem

$$\max_{x_k^p} \quad \mathcal{L}(\mathcal{D}^{tr}, \hat{\beta}) \tag{P1a}$$

$$\text{s.t.} \quad x_{k,1}^p = 1, \quad x_{k,j}^p \in [0,1], \quad j = 2, \dots, d, \qquad k = 1, \dots, q \tag{P1b}$$

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \mathcal{L}(\mathcal{D}^{tr} \cup \mathcal{D}^p, \beta) + \lambda \|\beta\|_2^2 \right\}, \tag{P1c}$$

where

$$\begin{aligned} \mathcal{L}(\mathcal{D}^{tr} \cup \mathcal{D}^p, \beta) = &- \sum_{i=1}^{n} \big[y_i^{tr} \log m_\beta(x_i^{tr}) + (1 - y_i^{tr}) \log\big(1 - m_\beta(x_i^{tr})\big) \big] \\ &- \sum_{k=1}^{q} \big[y_k^{p} \log m_\beta(x_k^{p}) + (1 - y_k^{p}) \log\big(1 - m_\beta(x_k^{p})\big) \big]. \end{aligned} \tag{3}$$

Note that the upper-level objective (P1a) measures the classifier's logistic loss on the clean data only, whereas the lower-level problem (P1c) trains on the contaminated set. The regularization appears only in the lower level and it reflects the defender's training procedure, not the attacker's success criterion.

2.2. **Gradient and Hessian of the Lower-Level Objective Function.** To reformulate the bilevel problem as a single-level one, we need the first-order optimality conditions of the lower level. Let $\mathcal{L}(\beta) := \mathcal{L}(\mathcal{D}^{tr} \cup \mathcal{D}^{p}, \beta) + \lambda\|\beta\|_2^2$. Stacking training and poisoning data, we define

$$X = \begin{bmatrix} x_1^{tr\top} \\ \vdots \\ x_n^{tr\top} \\ x_1^{p\top} \\ \vdots \\ x_q^{p\top} \end{bmatrix} \in \mathbb{R}^{(n+q)\times d}, \quad y = \begin{pmatrix} y_1^{tr} \\ \vdots \\ y_n^{tr} \\ y_1^{p} \\ \vdots \\ y_q^{p} \end{pmatrix} \in \mathbb{R}^{n+q}, \quad m_\beta(X) = \begin{pmatrix} m_\beta(x_1^{tr}) \\ \vdots \\ m_\beta(x_n^{tr}) \\ m_\beta(x_1^{p}) \\ \vdots \\ m_\beta(x_q^{p}) \end{pmatrix} \in \mathbb{R}^{n+q}.$$

In this notation, the regularized lower-level objective is

$$\mathcal{L}(\beta) = -y^\top \log m_\beta(X) - (\mathbf{1} - y)^\top \log\big(\mathbf{1} - m_\beta(X)\big) + \lambda\|\beta\|_2^2,$$

where the logarithm is applied componentwise and $\mathbf{1} \in \mathbb{R}^{n+q}$ denotes the vector of all ones. The gradient and Hessian are then given by

$$\nabla\mathcal{L}(\beta) = X^\top \left(m_\beta(X) - y\right) + 2\lambda\beta, \quad \nabla^2\mathcal{L}(\beta) = X^\top \Gamma\, X + 2\lambda I, \tag{4}$$

where $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_{n+q})$ with

$$\gamma_i = \begin{cases} m_\beta(x_i^{tr})\left(1 - m_\beta(x_i^{tr})\right), & 1 \le i \le n, \\ m_\beta(x_{i-n}^{p})\left(1 - m_\beta(x_{i-n}^{p})\right), & n{+}1 \le i \le n{+}q. \end{cases} \tag{5}$$

2.3. **Strong Convexity and Single-Level Reformulation.**

Proposition 1. *For every $\lambda > 0$, the regularized loss*

$$\mathcal{L}(\mathcal{D}^{\mathrm{tr}} \cup \mathcal{D}^{\mathrm{p}}, \beta) + \lambda\|\beta\|_2^2$$

is 2λ-strongly convex in β.

Proof. By (4), the Hessian of the regularized loss is

$$\nabla^2\mathcal{L}(\beta) = X^\top \Gamma X + 2\lambda I.$$

Since each diagonal entry of Γ is of the form

$$\gamma_i = m_\beta(x_i)\big(1 - m_\beta(x_i)\big) \ge 0,$$

the matrix Γ is positive semidefinite. Hence, for every $v \in \mathbb{R}^d$,

$$v^\top \nabla^2\mathcal{L}(\beta) v = (Xv)^\top \Gamma (Xv) + 2\lambda\|v\|_2^2 \ge 2\lambda\|v\|_2^2.$$

Therefore,

$$\nabla^2\mathcal{L}(\beta) \succeq 2\lambda I$$

for every $\beta \in \mathbb{R}^d$. Thus, $\mathcal{L}$ is 2λ-strongly convex. □

Moreover, the lower-level objective is continuous and coercive, since

$$\mathcal{L}(\mathcal{D}^{\mathrm{tr}} \cup \mathcal{D}^{\mathrm{p}}, \beta) + \lambda\|\beta\|_2^2 \ge \lambda\|\beta\|_2^2 \to \infty \quad \text{as} \quad \|\beta\|_2 \to \infty.$$

Hence, the lower-level problem (P1c) admits a minimizer. By Proposition 1, this minimizer is unique and we denote it by $\hat{\beta}$. Since the lower-level objective is differentiable and strongly convex and the problem is unconstrained, $\hat{\beta}$ is characterized by the first-order optimality condition

$$\nabla \mathcal{L}(\hat{\beta}) = 0.$$

Hence, we obtain the single-level reformulation

$$\begin{aligned} \max_{x_k^p,\, \beta} \quad & \mathcal{L}(\mathcal{D}^{tr}, \beta) && \text{(P2a)} \\ \text{s.t.} \quad & x_{k,1}^p = 1, \quad x_{k,j}^p \in [0,1], \quad j = 2, \dots, d, \quad k = 1, \dots, q, && \text{(P2b)} \\ & X^\top \big(m_\beta(X) - y\big) + 2\lambda\beta = 0. && \text{(P2c)} \end{aligned}$$

3. Convex-Analytic Reformulation

The single-level formulation (P2) is nonconvex since the stationarity constraint (P2c) contains sigmoid evaluations and the poisoning features x_k^p interact bilinearly with the classifier coefficients β. In this section, we derive an equivalent reformulation in which the only nonconvexities are bounded bilinear terms.

We begin in Section 3.1 by rewriting the logistic loss using signed margins and softplus and by establishing the convex-conjugacy properties of the softplus. In Section 3.2, we use these properties to replace the nonconvex sigmoid evaluations in the stationarity condition by a convex Fenchel-gap inequality, and rewrite the upper-level objective in concave form. Section 3.3 eliminates the poisoning features x_k^p by a change of variables that renders the stationarity condition linear. Section 3.4 derives finite bounds on all variables from the problem data and the regularization parameter, without arbitrary big-M constants. The resulting reformulation (P3) and its equivalence with (P2) are presented in Section 3.5.

3.1. Signed Margins and Convex Conjugacy.

For a label $y \in \{0, 1\}$, define

$$a := 1 - 2y \in \{-1, 1\}. \tag{6}$$

Accordingly, we set

$$s_i^{tr} := 1 - 2y_i^{tr}, \quad s_k^p := 1 - 2y_k^p.$$

Moreover, let

$$\phi(\theta) := \log(1 + e^\theta), \quad \theta \in \mathbb{R}. \tag{7}$$

For every observation (x, y), its logistic-loss contribution can then be written as

$$-y \log m_\beta(x) - (1 - y) \log\big(1 - m_\beta(x)\big) = \phi\big((1 - 2y)\beta^\top x\big). \tag{8}$$

To see this, note that for $y = 0$ the left-hand side reduces to

$$-\log(1 - m_\beta(x)) = \log(1 + e^{\beta^\top x}) = \phi(\beta^\top x).$$

For $y = 1$ it reduces to $-\log m_\beta(x) = \log(1 + e^{-\beta^\top x})$, which equals $\phi(-\beta^\top x)$. In both cases the right-hand side of (8) matches. We therefore introduce signed-margin variables

$$\theta_i^{tr} = s_i^{tr} \beta^\top x_i^{tr}, \quad i = 1, \dots, n. \tag{9}$$

For the poisoning points, corresponding variables θ_k^p will be introduced below. The derivative of ϕ is the sigmoid

$$\phi'(\theta) = \frac{1}{1 + e^{-\theta}}.$$

The following convex-conjugacy relation will allow us to avoid introducing a nonconvex sigmoid equality.

Proposition 2. *The convex conjugate of ϕ is*

$$\phi^*(\alpha) = \alpha \log \alpha + (1-\alpha)\log(1-\alpha), \quad \alpha \in [0,1], \tag{10}$$

where $0 \log 0$ is interpreted as 0. Moreover, for every $\theta \in \mathbb{R}$ and $\alpha \in [0,1]$,

$$\phi(\theta) + \phi^*(\alpha) - \alpha\theta \geq 0, \tag{11}$$

and equality holds if and only if

$$\alpha = \phi'(\theta) = \frac{1}{1+e^{-\theta}}. \tag{12}$$

Proof. By definition,

$$\phi^*(\alpha) = \sup_{\theta \in \mathbb{R}} \{\alpha\theta - \log(1+e^{\theta})\}.$$

For $\alpha \in (0,1)$, define $g(\theta) := \alpha\theta - \log(1+e^{\theta})$. Its first and second derivatives are

$$g'(\theta) = \alpha - \frac{e^{\theta}}{1+e^{\theta}}, \quad g''(\theta) = -\frac{e^{\theta}}{(1+e^{\theta})^2} < 0 \quad \text{for all } \theta \in \mathbb{R}.$$

Hence, g is strictly concave, so any stationary point is the unique global maximizer and satisfies

$$\alpha = \frac{e^{\theta}}{1+e^{\theta}}, \tag{13}$$

which is equivalent to

$$\theta = \log \frac{\alpha}{1-\alpha}.$$

Since g is strictly concave and θ^* is its unique maximizer,

$$\phi^*(\alpha) = \sup_{\theta \in \mathbb{R}} g(\theta) = g(\theta^*).$$

Evaluating $g(\theta^*)$ with $\theta^* = \log \frac{\alpha}{1-\alpha}$ gives

$$\begin{aligned} g(\theta^*) &= \alpha \log \frac{\alpha}{1-\alpha} - \log\left(1 + \frac{\alpha}{1-\alpha}\right) \\ &= \alpha \log \frac{\alpha}{1-\alpha} - \log \frac{1}{1-\alpha} \\ &= \alpha \log \alpha - \alpha \log(1-\alpha) + \log(1-\alpha) \\ &= \alpha \log \alpha + (1-\alpha)\log(1-\alpha), \end{aligned}$$

which proves (10). The values at $\alpha = 0$ and $\alpha = 1$ follow by continuity.

Inequality (11) is Fenchel's inequality, which holds for any function and its convex conjugate. Since ϕ is differentiable and strictly convex, equality in Fenchel's inequality holds if and only if $\alpha = \phi'(\theta)$, which gives (12). □

3.2. **Reformulation of the Stationarity Condition and Upper-Level Objective.**

3.2.1. *Convex Optimality-Gap Constraint.* The stationarity condition of the lower-level problem involves evaluations of $1/(1+e^{-\theta})$, which is a nonconvex function of θ. Our goal in this section is to replace the stationarity condition by a system of constraints that is equivalent but whose only nonconvexities are bilinear terms.

For the moment, suppose that signed margins θ_i^{tr} and θ_k^p are available for all training and poisoning points. Using (8), each loss contribution can be written as $\phi(\theta)$, so the stationarity condition of the lower-level problem reads

$$\sum_{i=1}^{n} s_i^{tr}\, \phi'(\theta_i^{tr})\, x_i^{tr} + \sum_{k=1}^{q} s_k^p\, \phi'(\theta_k^p)\, x_k^p + 2\lambda\beta = 0. \tag{14}$$

This condition is nonconvex because of the evaluations of ϕ'. We add new variables

$$\alpha_i^{tr} \in [0,1], \quad i = 1,\dots,n, \quad \alpha_k^{p} \in [0,1], \quad k = 1,\dots,q, \tag{15}$$

to the problem and consider the bilinear condition

$$\sum_{i=1}^{n} s_i^{tr}\, x_i^{tr}\, \alpha_i^{tr} + \sum_{k=1}^{q} s_k^{p}\, x_k^{p}\, \alpha_k^{p} + 2\lambda\beta = 0, \tag{16}$$

which is obtained from (14) by writing α_i in place of $\phi'(\theta_i)$. Every point satisfying (14) also satisfies (16), by setting $\alpha_i = \phi'(\theta_i) \in (0,1)$, but the converse does not hold because Condition (16) accepts any $\alpha \in [0,1]$ and is therefore a relaxation of the original stationarity condition.

We now show that adding a single convex inequality restores the equivalence. Taking the scalar product of (16) with β and using $\theta_i^{tr} = s_i^{tr}\beta^\top x_i^{tr}$ and $\theta_k^{p} = s_k^{p}\beta^\top x_k^{p}$ gives

$$\sum_{i=1}^{n} \alpha_i^{tr}\, \theta_i^{tr} + \sum_{k=1}^{q} \alpha_k^{p}\, \theta_k^{p} + 2\lambda\|\beta\|_2^2 = 0. \tag{17}$$

It remains to recover the equivalence with the original stationarity condition (14), i.e., to ensure that $\alpha_i = \phi'(\theta_i)$ holds at every feasible point. Recall from Proposition 2 that $\alpha = \phi'(\theta)$ if and only if the Fenchel gap (11) vanishes. The sum of all individual Fenchel gaps can be written as

$$\begin{aligned} &\sum_{i=1}^{n} \left[\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr}) - \alpha_i^{tr}\,\theta_i^{tr}\right] + \sum_{k=1}^{q} \left[\phi(\theta_k^{p}) + \phi^*(\alpha_k^{p}) - \alpha_k^{p}\,\theta_k^{p}\right] \\ = &\sum_{i=1}^{n} \left[\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr})\right] + \sum_{k=1}^{q} \left[\phi(\theta_k^{p}) + \phi^*(\alpha_k^{p})\right] + 2\lambda\|\beta\|_2^2, \end{aligned} \tag{18}$$

where the equality follows from (17). By Proposition 2, every individual Fenchel gap is nonnegative, so the right-hand side in (18) satisfies

$$\sum_{i=1}^{n} \left[\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr})\right] + \sum_{k=1}^{q} \left[\phi(\theta_k^{p}) + \phi^*(\alpha_k^{p})\right] + 2\lambda\|\beta\|_2^2 \geq 0. \tag{19}$$

The left-hand side of (19) is convex, since ϕ, ϕ^*, and $\|\cdot\|_2^2$ are all convex. Now, we impose the reverse inequality as a constraint:

$$\sum_{i=1}^{n} \left[\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr})\right] + \sum_{k=1}^{q} \left[\phi(\theta_k^{p}) + \phi^*(\alpha_k^{p})\right] + 2\lambda\|\beta\|_2^2 \leq 0. \tag{20}$$

Together, (19) and (20) force the sum to be zero. Since every individual Fenchel gap is nonnegative and their sum is zero, each gap must vanish. By (12) in Proposition 2, this implies $\alpha_i^{tr} = \phi'(\theta_i^{tr})$ and $\alpha_k^{p} = \phi'(\theta_k^{p})$.

Therefore, the original stationarity condition (14) in the variables (x^p, β) is equivalent to conditions (15), (16), and (20) in the extended variables (x^p, β, α): the bilinear relaxation (16) enlarges the feasible set by introducing the α variables (15), and the convex inequality (20) compensates this exactly, forcing $\alpha_i = \phi'(\theta_i)$ at every feasible point.

3.2.2. *Concave Form of the Objective.* Finally, the upper-level objective can be rewritten in a concave way. As shown above, Conditions (19) and (20) force all Fenchel gaps (11) to vanish. In particular, for each $i = 1,\dots,n$,

$$\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr}) - \alpha_i^{tr}\,\theta_i^{tr} = 0$$

holds. Summing over the training points gives

$$\sum_{i=1}^{n} \phi(\theta_i^{tr}) = \sum_{i=1}^{n} \alpha_i^{tr}\, \theta_i^{tr} - \sum_{i=1}^{n} \phi^*(\alpha_i^{tr}), \tag{21}$$

and by (8) and (9) the left-hand side equals $\mathcal{L}(\mathcal{D}^{tr}, \beta)$.

Similarly, for each poisoning point, $k = 1, \ldots, q$, the vanishing Fenchel gap gives $\alpha_k^p\, \theta_k^p = \phi(\theta_k^p) + \phi^*(\alpha_k^p)$. Substituting these relations into (17) yields

$$\sum_{i=1}^{n} \alpha_i^{tr}\, \theta_i^{tr} = -\sum_{k=1}^{q} \left[\phi(\theta_k^p) + \phi^*(\alpha_k^p)\right] - 2\lambda\|\beta\|_2^2.$$

Therefore, using (21), we obtain

$$\mathcal{L}(\mathcal{D}^{tr}, \beta) = -\sum_{i=1}^{n} \phi^*(\alpha_i^{tr}) - 2\lambda\|\beta\|_2^2 - \sum_{k=1}^{q} \left[\phi(\theta_k^p) + \phi^*(\alpha_k^p)\right]. \tag{22}$$

The right-hand side of (22) is concave because $-\phi$, $-\phi^*$, and $-\|\cdot\|_2^2$ are concave.

3.3. **Eliminating the Poisoning Features.** The stationarity condition (16) contains the products $x_k^p\, \alpha_k^p$, which are bilinear in the variables x^p and α. We now eliminate x_k^p by a change of variables that renders the stationarity condition linear.

To this end, we define

$$h_k := \alpha_k^p x_k^p \in \mathbb{R}^d, \quad k = 1, \ldots, q. \tag{23}$$

For the intercept component $j = 1$, since $x_{k,1}^p = 1$ holds, (23) gives $h_{k,1} = \alpha_k^p$, so the bound (24) right below holds with equality for $j = 1$.

At every point satisfying (12), we have $\alpha_k^p \in (0, 1)$. Therefore, $x_k^p \in [0, 1]^d$ is equivalent to

$$0 \le h_{k,j} \le \alpha_k^p, \quad j = 1, \ldots, d. \tag{24}$$

Moreover, since $\alpha_k^p > 0$ holds at every feasible point, multiplying

$$\theta_k^p = s_k^p \beta^\top x_k^p$$

by α_k^p and using (23) yields the equivalent relation

$$\alpha_k^p \theta_k^p = s_k^p \beta^\top h_k, \quad k = 1, \ldots, q. \tag{25}$$

The poisoning part of stationarity becomes linear:

$$s_k^p x_k^p \alpha_k^p = s_k^p h_k.$$

Hence, the complete stationarity condition (16) is given by

$$\sum_{i=1}^{n} s_i^{tr} x_i^{tr} \alpha_i^{tr} + \sum_{k=1}^{q} s_k^p h_k + 2\lambda\beta = 0. \tag{26}$$

3.4. **Valid Bounds.** The remaining nonconvexities in the reformulation will be the bilinear terms in (25). Spatial branch-and-bound therefore requires finite bounds on β, θ^p, α^p, and h. We next derive such bounds.

Let $N := n + q$ and define

$$\sigma(z) := \phi'(z) = \frac{1}{1 + e^{-z}}.$$

3.4.1. *A Global Bound on the Classifier Coefficients.* We first derive a stronger bound than the one obtained merely by comparing the regularized objective value at β^* and at 0. For this, we need an auxiliary lemma.

Lemma 1. *Let β^* be the unique minimizer of the lower-level problem for an arbitrary admissible poisoning set. Define*

$$H(r) := 2\lambda r^2 + \sum_{i=1}^{n} \phi\big(-r\|x_i^{tr}\|_2\big) + q\,\phi(-r\sqrt{d}), \quad r \geq 0. \tag{27}$$

Then, $H(\|\beta^*\|_2) \leq N \log 2$ holds.

Proof. Let

$$F(\beta) := \mathcal{L}(\mathcal{D}^{tr} \cup \mathcal{D}^p, \beta)$$

denote the unregularized contaminated loss. At the lower-level minimizer β^*, since $\nabla\mathcal{L}(\beta^*) = 0$ holds, we have

$$\nabla F(\beta^*) = -2\lambda\beta^*.$$

By (6)–(8), it follows

$$F(\beta) = \sum_{i=1}^{n} \phi(s_i^{tr}\beta^\top x_i^{tr}) + \sum_{k=1}^{q} \phi(s_k^p \beta^\top x_k^p)$$

holds. Since $\phi(0) = \log 2$, setting $\beta = 0$ gives $F(0) = N \log 2$. Because F is convex,

$$N\log 2 = F(0) \geq F(\beta^*) + \nabla F(\beta^*)^\top(0 - \beta^*) = F(\beta^*) + 2\lambda\|\beta^*\|_2^2.$$

Set $\rho = \|\beta^*\|_2$. By (6), we get $|s_i^{tr}| = 1$. The Cauchy–Schwarz inequality gives, for every clean observation,

$$s_i^{tr}\beta^{*\top}x_i^{tr} \geq -\rho\,\|x_i^{tr}\|_2,$$

and, because ϕ is increasing, we obtain

$$\phi\big(s_i^{tr}\beta^{*\top}x_i^{tr}\big) \geq \phi\big(-\rho\,\|x_i^{tr}\|_2\big)\,.$$

Similarly, for every poisoning point $x_k^p \in [0,1]^d$,

$$s_k^p\beta^{*\top}x_k^p \geq -\rho\,\|x_k^p\|_2 \geq -\rho\sqrt{d},$$

so that $\phi(s_k^p\beta^{*\top}x_k^p) \geq \phi(-\rho\sqrt{d})$. Summing over all data points gives

$$\begin{aligned} F(\beta^*) &= \sum_{i=1}^{n} \phi\big(s_i^{tr}\beta^{*\top}x_i^{tr}\big) + \sum_{k=1}^{q} \phi\big(s_k^p\beta^{*\top}x_k^p\big) \\ &\geq \sum_{i=1}^{n} \phi\big(-\rho\,\|x_i^{tr}\|_2\big) + q\,\phi(-\rho\sqrt{d}) \\ &= H(\rho) - 2\lambda\rho^2, \end{aligned}$$

where the last equality follows from the definition (27) of H. Therefore, we finally get

$$N\log 2 \geq F(\beta^*) + 2\lambda\rho^2 \geq H(\rho). \qquad \square$$

Proposition 3. *Let x^p be an arbitrary admissible poisoning set, i.e., one satisfying (P1b), and let β^* be the corresponding unique minimizer of the lower-level problem (P1c). Consider H as in (27) and let R denote the largest nonnegative solution of*

$$H(R) = N\log 2. \tag{28}$$

Then

$$\|\beta^*\|_2 \leq R. \tag{29}$$

Moreover, $R < \sqrt{N\log 2/(2\lambda)}$.

Proof. From $\phi(0) = \log 2$, we obtain

$$H(0) = 0 + \sum_{i=1}^{n} \phi(0) + q\,\phi(0) = N \log 2.$$

Differentiating (27) at $r = 0$ gives

$$H'(0) = -\phi'(0)\left(\sum_{i=1}^{n} \|x_i^{tr}\|_2 + q\sqrt{d}\right) < 0,$$

so H decreases below $N \log 2$ for small positive r.

Since H is the sum of the strictly convex term $2\lambda r^2$ and convex terms of the form $\phi(-r\|x\|)$, it is strictly convex. Moreover $H(r) \to +\infty$ as $r \to \infty$. Hence, the equation $H(R) = N \log 2$ has a unique positive solution R. The sublevel set

$$\{r \geq 0 : H(r) \leq N \log 2\}$$

is therefore the interval $[0, R]$. By Lemma 1, $H(\|\beta^*\|_2) \leq N \log 2$, so $\|\beta^*\|_2 \leq R$. For the final claim, since $\phi(\theta) = \log(1 + e^{\theta}) > 0$ for all $\theta \in \mathbb{R}$, the definition of H in (27) gives

$$H(R) = 2\lambda R^2 + \underbrace{\sum_{i=1}^{n} \phi\big(-R\|x_i^{tr}\|_2\big) + q\,\phi(-R\sqrt{d})}_{>\,0} > 2\lambda R^2.$$

Combining this with $H(R) = N \log 2$ finally leads to $2\lambda R^2 < N \log 2$, i.e., $R < \sqrt{N \log 2/(2\lambda)}$. □

Proposition 3 is a statement about the lower-level minimizer and does not depend on how the bilevel problem is reformulated. Every feasible point (x^p, β) of (P2) satisfies the stationarity condition (P2c), which, by Proposition 1, characterizes the unique lower-level minimizer; hence $\beta = \beta^*$. Consequently, the constraints

$$\beta \in [-R, R]^d, \quad |\beta^\top x| \leq R\,\|x\|_2 \leq R\sqrt{d} \quad \text{for } x \in [0,1]^d,$$

can be added to (P2) without excluding any feasible point. More generally, all bounds on β, θ^{tr}, and $\beta^\top x_k^p$ derived in this section are properties of the lower-level minimizer and are therefore also valid for (P2). In our numerical experiments, both formulations are tested with the bounds of Proposition 3 alone and with all bounds of Section 3.4; see Section 4.2.

3.4.2. *Componentwise Bounds.* The stationarity condition provides additional componentwise bounds. Let

$$q_0 := \left|\{k\colon y_k^p = 0\}\right|, \quad q_1 := \left|\{k\colon y_k^p = 1\}\right|.$$

For $j = 1, \ldots, d$, define

$$C_j^0 := \sum_{\substack{i=1 \\ y_i^{tr}=0}}^{n} x_{i,j}^{tr}\sigma\big(R\|x_i^{tr}\|_2\big) + q_0\,\sigma(R\sqrt{d}), \tag{30}$$

$$C_j^1 := \sum_{\substack{i=1 \\ y_i^{tr}=1}}^{n} x_{i,j}^{tr}\sigma\big(R\|x_i^{tr}\|_2\big) + q_1\,\sigma(R\sqrt{d}). \tag{31}$$

Proposition 4. *Every lower-level minimizer β^* satisfies*

$$\underline{\beta}_j^0 \leq \beta_j^* \leq \overline{\beta}_j^0, \quad j = 1, \ldots, d, \tag{32}$$

where

$$\underline{\beta}_j^0 := -\min\left\{R, \frac{C_j^0}{2\lambda}\right\}, \quad \overline{\beta}_j^0 := \min\left\{R, \frac{C_j^1}{2\lambda}\right\}. \tag{33}$$

Proof. The jth component of the stationarity condition (P2c) can be written as

$$\begin{aligned} 2\lambda\beta_j^* = &\sum_{\substack{i=1\\ y_i^{tr}=1}}^{n} x_{i,j}^{tr}\big(1 - m_{\beta^*}(x_i^{tr})\big) + \sum_{\substack{k=1\\ y_k^{p}=1}}^{q} x_{k,j}^{p}\big(1 - m_{\beta^*}(x_k^{p})\big) \\ &- \sum_{\substack{i=1\\ y_i^{tr}=0}}^{n} x_{i,j}^{tr}\, m_{\beta^*}(x_i^{tr}) - \sum_{\substack{k=1\\ y_k^{p}=0}}^{q} x_{k,j}^{p}\, m_{\beta^*}(x_k^{p}). \end{aligned} \tag{34}$$

Hence,

$$\begin{aligned} 2\lambda\beta_j^* &\geq -\sum_{\substack{i=1\\ y_i^{tr}=0}}^{n} x_{i,j}^{tr}\, m_{\beta^*}(x_i^{tr}) - \sum_{\substack{k=1\\ y_k^{p}=0}}^{q} x_{k,j}^{p}\, m_{\beta^*}(x_k^{p}), \\ 2\lambda\beta_j^* &\leq \sum_{\substack{i=1\\ y_i^{tr}=1}}^{n} x_{i,j}^{tr}\big(1 - m_{\beta^*}(x_i^{tr})\big) + \sum_{\substack{k=1\\ y_k^{p}=1}}^{q} x_{k,j}^{p}\big(1 - m_{\beta^*}(x_k^{p})\big). \end{aligned} \tag{35}$$

By Proposition 3, $|\beta^{*\top} x_i^{tr}| \leq R\|x_i^{tr}\|_2$ holds. Since σ is increasing, $m_{\beta^*}(x_i^{tr}) = \sigma(\beta^{*\top} x_i^{tr}) \leq \sigma(R\|x_i^{tr}\|_2)$. Moreover, $1 - m_{\beta^*}(x_i^{tr}) = \sigma(-\beta^{*\top} x_i^{tr}) \leq \sigma(R\|x_i^{tr}\|_2)$. Analogously, for a poisoning point,

$$|\beta^{*\top} x_k^{p}| \leq R\sqrt{d},$$

so both corresponding probabilities are bounded above by $\sigma(R\sqrt{d})$. Substituting into the upper and lower bound of (35) and using $x_{k,j}^{p} \leq 1$, gives

$$\begin{aligned} 2\lambda\beta_j^* &\leq \sum_{\substack{i=1\\ y_i^{tr}=1}}^{n} x_{i,j}^{tr}\,\sigma\big(R\|x_i^{tr}\|_2\big) + \sum_{\substack{k=1\\ y_k^{p}=1}}^{q} 1\cdot\sigma(R\sqrt{d}) = C_j^1, \\ 2\lambda\beta_j^* &\geq -\sum_{\substack{i=1\\ y_i^{tr}=0}}^{n} x_{i,j}^{tr}\,\sigma\big(R\|x_i^{tr}\|_2\big) - q_0\,\sigma(R\sqrt{d}) = -C_j^0. \end{aligned}$$

Therefore, we get

$$-\frac{C_j^0}{2\lambda} \leq \beta_j^* \leq \frac{C_j^1}{2\lambda}.$$

Intersecting these bounds with $|\beta_j^*| \leq \|\beta^*\|_2 \leq R$ gives (33). □

The bounds in Proposition 4 are closed-form bounds. They can be tightened further by solving small convex preprocessing problems. In fact, the proof of Lemma 1 shows that

$$N\log 2 \geq F(\beta^*) + 2\lambda\|\beta^*\|_2^2.$$

Splitting $F(\beta^*)$ into its clean and poisoning contributions and using the lower bound on the poisoning part established in that proof gives

$$\begin{aligned} F(\beta^*) &= \sum_{i=1}^{n} \phi(s_i^{tr}\beta^{*\top} x_i^{tr}) + \sum_{k=1}^{q} \phi(s_k^{p}\beta^{*\top} x_k^{p}) \\ &\geq \mathcal{L}(\mathcal{D}^{tr}, \beta^*) + q\phi(-\rho\sqrt{d}) \\ &\geq \mathcal{L}(\mathcal{D}^{tr}, \beta^*) + q\,\phi(-R\sqrt{d}) \geq q\,\phi(-R\sqrt{d}). \end{aligned}$$

Substituting yields

$$\mathcal{L}(\mathcal{D}^{tr}, \beta^*) + 2\lambda\|\beta^*\|_2^2 \leq N \log 2 - q\,\phi(-R\sqrt{d}). \tag{36}$$

Hence, every feasible classifier coefficient vector belongs to the compact convex set

$$\mathcal{B} := \left\{ \beta \in \mathbb{R}^d \;\middle|\; \begin{array}{l} \|\beta\|_2 \leq R, \\ \underline{\beta}_j^0 \leq \beta_j \leq \overline{\beta}_j^0, \quad j = 1, \ldots, d, \\ \mathcal{L}(\mathcal{D}^{tr}, \beta) + 2\lambda\|\beta\|_2^2 \leq N \log 2 - q\,\phi(-R\sqrt{d}) \end{array} \right\}. \tag{37}$$

We may therefore define the tighter bounds

$$\underline{\beta}_j := \min_{\beta \in \mathcal{B}} \beta_j, \quad \overline{\beta}_j := \max_{\beta \in \mathcal{B}} \beta_j. \tag{38}$$

Both problems in (38) are convex optimization problems. The resulting bounds satisfy $\underline{\beta}_j^0 \leq \underline{\beta}_j$ and $\overline{\beta}_j \leq \overline{\beta}_j^0$, that is, they are at least as tight as the closed-form bounds (32). When solving (38) is not desired, one may simply set

$$\underline{\beta}_j = \underline{\beta}_j^0, \quad \overline{\beta}_j = \overline{\beta}_j^0.$$

3.4.3. *Bounds on Signed Margins and Sigmoid Variables.* We now derive bounds on the variables θ, α, and h via propagating the bounds on β obtained above.

First, for a clean observation, the signed margin $\theta_i^{tr} = s_i^{tr} \beta^\top x_i^{tr}$ is determined by β. Since every feasible β belongs to $\mathcal{B}$, the range of θ_i^{tr} is contained in the interval $[\underline{\theta}_i^{tr}, \overline{\theta}_i^{tr}]$ given by

$$\underline{\theta}_i^{tr} := \min_{\beta \in \mathcal{B}} s_i^{tr} \beta^\top x_i^{tr}, \quad \overline{\theta}_i^{tr} := \max_{\beta \in \mathcal{B}} s_i^{tr} \beta^\top x_i^{tr}. \tag{39}$$

For a poisoning point, the signed margin $\theta_k^p = s_k^p \beta^\top x_k^p$ depends on the decision variables β and x_k^p. Consider $\xi^p = \beta^\top x^p$. By the Cauchy–Schwarz inequality, $\|\beta\|_2 \leq R$ and $\|x^p\|_2 \leq \sqrt{d}$, we have that $|\xi^p| \leq R\sqrt{d}$.

A tighter bound can be obtained componentwise. Since we have $x_j^p \in [0, 1]$ and $\beta_j \in [\underline{\beta}_j, \overline{\beta}_j]$, each product $\beta_j x_j^p$ satisfies

$$\beta_j x_j^p \in \left[\min\{0, \underline{\beta}_j\},\ \max\{0, \overline{\beta}_j\}\right].$$

Summing over j and intersecting with the global bound gives $\xi^p \in [\underline{\xi}^p, \overline{\xi}^p]$ with

$$\underline{\xi}^p := \max\left\{ -R\sqrt{d},\ \sum_{j=1}^{d} \min\{0, \underline{\beta}_j\} \right\}, \tag{40}$$

$$\overline{\xi}^p := \min\left\{ R\sqrt{d},\ \sum_{j=1}^{d} \max\{0, \overline{\beta}_j\} \right\}. \tag{41}$$

The signed margin is $\theta_k^p = s_k^p \xi^p$. When $s_k^p = 1$, the bounds on ξ^p apply directly. When $s_k^p = -1$, multiplying by -1 reverses the interval. Both cases are captured by

$$\underline{\theta}_k^p := \min\left\{ s_k^p \underline{\xi}^p,\ s_k^p \overline{\xi}^p \right\}, \quad \overline{\theta}_k^p := \max\left\{ s_k^p \underline{\xi}^p,\ s_k^p \overline{\xi}^p \right\}. \tag{42}$$

If the tightest poisoning-margin bounds implied by $\mathcal{B}$ are desired, they can instead be obtained from

$$\underline{\xi}_{\mathcal{B}}^p := \min_{v \in \{1\} \times \{0,1\}^{d-1}} \min_{\beta \in \mathcal{B}} \beta^\top v, \tag{43}$$

$$\overline{\xi}_{\mathcal{B}}^p := \max_{v \in \{1\} \times \{0,1\}^{d-1}} \max_{\beta \in \mathcal{B}} \beta^\top v. \tag{44}$$

Indeed, for fixed β, a linear function of $x \in \{1\} \times [0,1]^{d-1}$ attains its minimum and maximum at a vertex of this admissible box. Each inner problem in (43)–(44) is convex. The price for these exact bounds is the possible enumeration of 2^{d-1} admissible vertices; hence (40)–(41) provide a substantially cheaper alternative.

Next, bounds on θ propagate to bounds on α. Since $\alpha_i = \phi'(\theta_i)$ at every feasible point and ϕ' is strictly increasing, $\theta \in [\underline{\theta}, \overline{\theta}]$ implies $\alpha \in [\phi'(\underline{\theta}), \phi'(\overline{\theta})]$. Explicitly,

$$\underline{\alpha}_i^{tr} := \phi'(\underline{\theta}_i^{tr}) \le \alpha_i^{tr} \le \phi'(\overline{\theta}_i^{tr}) := \overline{\alpha}_i^{tr}, \tag{45}$$

$$\underline{\alpha}_k^{p} := \phi'(\underline{\theta}_k^{p}) \le \alpha_k^{p} \le \phi'(\overline{\theta}_k^{p}) := \overline{\alpha}_k^{p}. \tag{46}$$

Since $\phi'(\theta) \in (0,1)$ for every finite θ, in particular

$$0 \le \underline{\alpha}_k^{p} \le \alpha_k^{p} \le \overline{\alpha}_k^{p} \le 1.$$

Finally, the bounds on h follow directly from (24) and the upper bound on α_k^p:

$$0 \le h_{k,j} \le \alpha_k^{p} \le \overline{\alpha}_k^{p}. \tag{47}$$

The bounds (38), (42), (46), and (47) provide finite intervals for every variable that appears in a bilinear term. These bounds are derived from the problem data and the regularization parameter λ. Thus, no arbitrary big-M constants are needed.

3.5. **Equivalent Reformulated Problem.** Collecting the preceding results yields the alternative single-level reformulation

$$\max_{\beta,\theta^{tr},\theta^{p},\alpha^{tr},\alpha^{p},h} \; -\sum_{i=1}^{n} \phi^*(\alpha_i^{tr}) - \sum_{k=1}^{q} \big[\phi(\theta_k^{p}) + \phi^*(\alpha_k^{p})\big] - 2\lambda\|\beta\|_2^2 \tag{P3a}$$

$$\text{s.t.} \quad \theta_i^{tr} = s_i^{tr}\beta^\top x_i^{tr}, \quad i = 1,\dots,n, \tag{P3b}$$

$$0 \le h_{k,j} \le \alpha_k^{p}, \quad k = 1,\dots,q,\ j = 2,\dots,d, h_{k,1} = \alpha_k^{p}, \tag{P3c}$$

$$\alpha_k^{p}\,\theta_k^{p} = s_k^{p}\beta^\top h_k, \quad k = 1,\dots,q, \tag{P3d}$$

$$\sum_{i=1}^{n} s_i^{tr} x_{i,j}^{tr} \alpha_i^{tr} + \sum_{k=1}^{q} s_k^{p} h_{k,j} + 2\lambda\beta_j = 0, \quad j = 1,\dots,d, \tag{P3e}$$

$$\sum_{i=1}^{n} \big[\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr})\big] + \sum_{k=1}^{q} \big[\phi(\theta_k^{p}) + \phi^*(\alpha_k^{p})\big] + 2\lambda\|\beta\|_2^2 \le 0, \tag{P3f}$$

$$\underline{\beta}_j \le \beta_j \le \overline{\beta}_j, \quad j = 1,\dots,d, \tag{P3g}$$

$$\underline{\theta}_i^{tr} \le \theta_i^{tr} \le \overline{\theta}_i^{tr}, \quad i = 1,\dots,n, \tag{P3h}$$

$$\underline{\theta}_k^{p} \le \theta_k^{p} \le \overline{\theta}_k^{p}, \quad k = 1,\dots,q, \tag{P3i}$$

$$\underline{\alpha}_i^{tr} \le \alpha_i^{tr} \le \overline{\alpha}_i^{tr}, \quad i = 1,\dots,n, \tag{P3j}$$

$$\underline{\alpha}_k^{p} \le \alpha_k^{p} \le \overline{\alpha}_k^{p}, \quad k = 1,\dots,q. \tag{P3k}$$

Proposition 5. *Formulations (P2) and (P3) are equivalent. More precisely, every feasible point of (P2) can be extended to a feasible point of (P3) with the same objective value, and every feasible point of (P3) determines feasible poisoning features for (P2) with the same objective value.*

Proof. First, let (x^p, β) be feasible for (P2). Define

$$\theta_i^{tr} = s_i^{tr}\beta^\top x_i^{tr}, \quad \theta_k^{p} = s_k^{p}\beta^\top x_k^{p}, \quad \alpha_i^{tr} = \sigma(\theta_i^{tr}), \quad \alpha_k^{p} = \sigma(\theta_k^{p}), \quad h_k = \alpha_k^{p} x_k^{p},$$

with $h_{k,1} = \alpha_k^p x_{k,1}^p = \alpha_k^p$ for every $k = 1,\dots,q$. Then, (P3b)–(P3d) hold. Moreover, $\alpha_i = \phi'(\theta_i)$ at every feasible point, which has two consequences. First, by (8),

$$s_i^{tr}\alpha_i^{tr} = m_\beta(x_i^{tr}) - y_i^{tr}, \quad s_k^{p}\alpha_k^{p} = m_\beta(x_k^{p}) - y_k^{p},$$

so (P3e) is precisely (P2c). Second, Proposition 2 guarantees that every Fenchel gap (11) vanishes, so (P3f) holds with equality. The bound constraints (P3g)–(P3k) are satisfied because β is a lower-level minimizer and the bounds in Section 3.4 were derived for every such minimizer. Finally, the objective value of (P3) equals $\mathcal{L}(\mathcal{D}^{tr}, \beta)$ by (22), which is the objective of (P2).

Conversely, let $(\beta, \theta^{tr}, \theta^p, \alpha^{tr}, \alpha^p, h)$ be feasible for (P3). Taking the scalar product of (P3e) with β and using (P3b) and (P3d) gives

$$2\lambda\|\beta\|_2^2 = -\sum_{i=1}^{n} \alpha_i^{tr}\theta_i^{tr} - \sum_{k=1}^{q} \alpha_k^p \theta_k^p.$$

Substituting this into the left-hand side of (P3f) yields the sum of all individual Fenchel gaps:

$$\sum_{i=1}^{n}\left[\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr}) - \alpha_i^{tr}\theta_i^{tr}\right] + \sum_{k=1}^{q}\left[\phi(\theta_k^p) + \phi^*(\alpha_k^p) - \alpha_k^p\theta_k^p\right] \le 0. \tag{48}$$

By Proposition 2, each Fenchel gap satisfies $\phi(\theta) + \phi^*(\alpha) - \alpha\theta \ge 0$ for any $\alpha \in [0,1]$ and $\theta \in \mathbb{R}$. Hence, every summand is nonnegative, which means the sum is also ≥ 0. Combined with ≤ 0 in (48), the sum equals zero. Since each summand is nonnegative and their sum is zero, each must vanish:

$$\begin{aligned}\phi(\theta_i^{tr}) + \phi^*(\alpha_i^{tr}) - \alpha_i^{tr}\theta_i^{tr} = 0, \quad i = 1,\dots,n,\\ \phi(\theta_k^p) + \phi^*(\alpha_k^p) - \alpha_k^p\theta_k^p = 0, \quad k = 1,\dots,q.\end{aligned}$$

By Proposition 2, equality holds in (11) if and only if (12) is satisfied. Since every gap vanishes, we conclude

$$\alpha_i^{tr} = \phi'(\theta_i^{tr}), \quad \alpha_k^p = \phi'(\theta_k^p). \tag{49}$$

In particular, $\alpha_k^p > 0$ holds, so we recover the poisoning features by setting

$$x_{k,j}^p := \frac{h_{k,j}}{\alpha_k^p}. \tag{50}$$

By (P3c), $x_k^p \in [0,1]^d$, with $x_{k,1}^p = 1$ for every $k = 1,\dots,q$.

Dividing (P3d) by α_k^p and using (50) gives

$$\theta_k^p = s_k^p\,\beta^\top \frac{h_k}{\alpha_k^p} = s_k^p\,\beta^\top x_k^p.$$

It remains to show that (x^p, β) is feasible for (P2) and that the objective values coincide. As shown in the forward direction, using (49), (8) implies

$$s_i^{tr}\alpha_i^{tr} = m_\beta(x_i^{tr}) - y_i^{tr}, \quad s_k^p\alpha_k^p = m_\beta(x_k^p) - y_k^p.$$

Substituting into (P3e) recovers (P2c), so (x^p, β) is feasible for (P2).

Finally, since all Fenchel gaps vanish, (P3f) holds with equality. By (22), the objective of (P3) equals $\mathcal{L}(\mathcal{D}^{tr}, \beta)$, which is the objective of (P2). □

3.6. **Convexity Structure and Spatial Branching.** The reformulation (P3) isolates the nonconvexity of the poisoning problem in constraint (P3d). More precisely, we have the following:

- The maximization objective (P3a) is concave.
- Constraints (P3b), (P3c), and (P3e) are affine.
- Constraint (P3f) is convex, since ϕ, ϕ^*, and $\|\cdot\|_2^2$ are convex.
- The only nonconvex constraint family is (P3d). It contains the products $\alpha_k^p\theta_k^p$ and $\beta_j h_{k,j}$.

Thus, if (P3d) were removed, the remaining problem would be convex. In particular, neither ϕ nor ϕ^* requires spatial branching and both can be represented using exponential-cone constraints if supported by the solver.

For spatial branch-and-bound, it is convenient to lift the remaining bilinearities. To this end, we introduce

$$r_k = \alpha_k^p \theta_k^p, \quad w_{k,j} = \beta_j h_{k,j}. \tag{51}$$

Then (P3d) becomes the linear equation

$$r_k = s_k^p \sum_{j=1}^{d} w_{k,j}, \quad k = 1, \dots, q, \tag{52}$$

and the only nonconvexities are ones in (51).

For a generic bilinear constraint

$$w = uv, \quad u \in [\underline{u}, \overline{u}], \quad v \in [\underline{v}, \overline{v}],$$

its convex hull over the bounding box is given by the McCormick inequalities (McCormick 1976)

$$w \geq \underline{u}\, v + \underline{v}\, u - \underline{u}\, \underline{v}, \tag{53}$$

$$w \geq \overline{u}\, v + \overline{v}\, u - \overline{u}\, \overline{v}, \tag{54}$$

$$w \leq \overline{u}\, v + \underline{v}\, u - \overline{u}\, \underline{v}, \tag{55}$$

$$w \leq \underline{u}\, v + \overline{v}\, u - \underline{u}\, \overline{v}. \tag{56}$$

For the product $r_k = \alpha_k^p \theta_k^p$, the required bounds are

$$\alpha_k^p \in [\underline{\alpha}_k^p, \overline{\alpha}_k^p], \quad \theta_k^p \in [\underline{\theta}_k^p, \overline{\theta}_k^p],$$

and for the product $w_{k,j} = \beta_j h_{k,j}$, they are

$$\beta_j \in [\underline{\beta}_j, \overline{\beta}_j], \quad h_{k,j} \in [0, \overline{\alpha}_k^p],$$

together with the stronger linear relation $h_{k,j} \leq \alpha_k^p$.

If explicit bounds on the lifted variables are desired, they follow directly and are given by

$$r_k \in \left[\min\left\{\underline{\alpha}_k^p \underline{\theta}_k^p,\, \underline{\alpha}_k^p \overline{\theta}_k^p,\, \overline{\alpha}_k^p \underline{\theta}_k^p,\, \overline{\alpha}_k^p \overline{\theta}_k^p\right\}, \max\left\{\underline{\alpha}_k^p \underline{\theta}_k^p,\, \underline{\alpha}_k^p \overline{\theta}_k^p,\, \overline{\alpha}_k^p \underline{\theta}_k^p,\, \overline{\alpha}_k^p \overline{\theta}_k^p\right\}\right]$$

as well as

$$w_{k,j} \in \left[\min\{0, \underline{\beta}_j \overline{\alpha}_k^p\}, \max\{0, \overline{\beta}_j \overline{\alpha}_k^p\}\right].$$

4. Numerical Results

In this section, we evaluate the poisoning attack on logistic regression and assess the computational benefits of the convex-analytic reformulation (P3). Section 4.1 describes the benchmark datasets, Section 4.2 the computational setup, Section 4.3 the evaluation criteria, and Section 4.4 the results.

4.1. **Datasets.** We evaluate the attack on eight binary classification datasets drawn from the RDatasets collection (Arel-Bundock 2026), namely from the MASS and ISLR packages (James et al. 2021), and from the UCI Machine Learning Repository (Kelly et al. 2023). The datasets are chosen to cover application domains where data poisoning poses a realistic threat: clinical decision support (Biopsy, breast cancer diagnosis), financial prediction (Smarket, S&P 500 direction), counterfeit detection (Banknote), radar surveillance (Ionosphere), and spam filtering (Spambase (Hopkins et al. 1999)). Spam filters are a classical target of poisoning, since they are retrained on data that an adversary can partly control; Spambase was

TABLE 1. Benchmark datasets after preprocessing. m is the total number of observations, d the number of numeric features (including the intercept column), and $P(y{=}1)$ the proportion of positive labels.

Instance	Source	m	d	$P(y{=}1)$	Domain
Biopsy	MASS	683	10	0.35	Healthcare (breast cancer)
Smarket	ISLR	1250	7	0.52	Finance (S&P 500)
fgl	MASS	214	10	0.68	Forensics (glass)
College	ISLR	777	18	0.73	Education
OJ	ISLR	1070	11	0.61	Retail (brand choice)
Banknote	UCI	1372	5	0.44	Counterfeit detection
Ionosphere	UCI	351	34	0.64	Radar (surveillance)
Spambase	UCI	4601	58	0.39	Spam filtering

also used by Mei and Zhu (2015) to evaluate poisoning attacks on logistic regression. The remaining instances (fgl, forensic glass identification; College, university classification; OJ, brand choice) complement the benchmark with varying sample sizes and feature dimensions. Table 1 lists the final instances.

We apply the following preprocessing to every dataset. First, observations with at least one missing value are removed. Second, the target variable is encoded as a binary label $y \in \{0, 1\}$; for fgl, whose original target distinguishes several glass types, we merge the window-glass categories (WinF, WinNF) into the positive class ($y = 1$) and the remaining types into the negative class ($y = 0$). Third, since the attack model requires features in a continuous box $[0, 1]^d$, only numeric features are retained. Columns with zero variance are removed because they carry no discriminative information and would produce a degenerate scaling. For Smarket, we use the predictors of the standard logistic regression example of (James et al. 2021) (Lag1–Lag5, Volume); in particular, Today is excluded, since the label Direction is its sign. For OJ, and only there, numeric columns with at most ten distinct values are removed: store identifiers and promotion indicators encode categories, for which intermediate values have no meaning, and list prices are linear combinations of retained features. Ordinal scores, such as the predictors of biopsy (integers from 1 to 10), are retained and treated as continuous. Poisoning points may thus take values that are not physically realizable; since this relaxation only enlarges the attacker's feasible set, the optimal training-loss objective value is an upper bound on that attainable by realizable attacks.

All features are scaled to $[0, 1]$. For each coordinate j, the minimum $l_j = \min_i x_j^i$ and maximum $u_j = \max_i x_j^i$ are computed on the training partition only, and the transformation

$$\tilde{x}_j^i = (x_j^i - l_j)/(u_j - l_j)$$

is applied. If $u_j = l_j$, the denominator is set to 1. Test points are transformed with the same parameters l_j, u_j and clipped to $[0, 1]$, so that all points lie in the same box as the admissible poisoning features. After scaling, a constant column of ones is prepended to encode the intercept, so that $x_{i,1} = 1$ for all training, test, and poisoning points, as required by the model in Section 2.

4.2. **Computational Setup.** For each dataset, we perform a stratified split that assigns 30 % of the observations of each class to the training set and the remaining

observations to the test set, so that the training set size is

$$n = \sum_{c \in \{0,1\}} \lfloor 0.3\, n_c \rceil\,,$$

where n_c denotes the number of observations with label c and $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, with ties rounded to the nearest even integer. We consider poisoning budgets $q = \max\{1, \lfloor \rho\, n \rceil\}$ with $\rho \in \{5\,\%, 10\,\%, 15\,\%, 20\,\%\}$. The labels of the poisoned points are fixed in advance and alternate between the two classes, so that $\lceil q/2 \rceil$ points have $y = 0$ and $\lfloor q/2 \rfloor$ have $y = 1$; the attacker only chooses their features. The regularization parameter is set to $\lambda = 1$ throughout.

The experiments are organized into two phases, each addressing a different question.

Phase 1: Attack Effectiveness: We first ask whether inserting poisoned points substantially degrades the classifier. Answering this requires an effective attack rather than an optimal one, so a local solver suffices, which allows us to repeat the experiment over 5 data partitions. For each partition, we compare:

- LR: the regularized logistic regression (2), solved with Ipopt (Wächter and Biegler 2006). Since (2) is strongly convex, a single run yields its unique global solution.
- PA-L: the KKT reformulation (P2) with the bounds of Proposition 3, solved with Ipopt from 5 random starting points, with $\beta \sim \mathcal{N}(0, 0.1^2 I)$ and the poisoned features drawn uniformly in $[0, 1]$. Each run is limited to 300 s, and the solution with the largest upper-level objective value is retained.

Phase 2: Global Optimization: The second question is how much the convex-analytic reformulation helps global optimization. To answer it, we use one partition per dataset and compare, with a time limit of 3600 s per instance:

- PA-G: SCIP applied to the KKT reformulation (P2), with only the bounds of Proposition 3.
- CAR: SCIP applied to the convex-analytic reformulation (P3), with only the bounds of Proposition 3.
- PA-G+: SCIP applied to (P2), with all bounds of Section 3.4.
- CAR+: SCIP applied to (P3), with all bounds of Section 3.4.

The convex problems (38) and (39) that provide the bounds of PA-G+ and CAR+ are solved with Ipopt (tolerance 10^{-9}), and each computed bound v is relaxed by $10^{-6}(1 + |v|)$. These bounds are therefore numerical: they are valid up to the accuracy of Ipopt, but not certified. The bounds of Proposition 3, used by PA-G and CAR, are computed by scalar bisection.

Finally, to avoid evaluating $\log 0$ in ϕ^*, the bounds on α in CAR and CAR+ are clipped to $[10^{-8}, 1 - 10^{-8}]$. All models are implemented in Julia 1.12.3 (Bezanson et al. 2017) using JuMP 1.29.3 (Dunning et al. 2017), with SCIP 10.0.3 (Hojny et al. 2025) (relative gap 10^{-4}, feasibility tolerance 10^{-6}) and Ipopt 3.14.19 (tolerance 10^{-6} for PA-L). Six runs are executed in parallel, with one thread each, on an Apple M3 Pro (12 cores: 6 performance and 6 efficiency) with 36 GB unified memory. Phase 1 uses five random partitions (seeds 42–46); Phase 2 uses the first of them (seed 42), and the five starting points of PA-L are drawn with seeds $1000p + s$, where p is the partition seed and $s = 1, \ldots, 5$ indexes the starts. The code is publicly available at https://github.com/mariaepinheiro/Poisoning_Attacks_Logistic_Regression.

In preliminary experiments on a subset of the instances, Gurobi 13.0 (Gurobi Optimization, LLC 2025) obtained substantially weaker dual bounds than SCIP for both formulations. We therefore report results only for SCIP, which has the

additional advantage of being open source, so that all results can be reproduced without a commercial license.

4.3. **Evaluation Criteria.**

4.3.1. *Attack Effectiveness.* A test point x is assigned to the positive class whenever $m_\beta(x) \geq 0.5$ and to the negative one otherwise. Knowing the true labels, we categorize predictions as true positive (TP), true negative (TN), false positive (FP), or false negative (FN), and compute two metrics for which a higher value indicates a better classification. From the resulting true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), we compute the accuracy

$$\mathrm{AC} := \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \in [0, 1], \tag{57}$$

and the Matthews correlation coefficient (Chicco and Jurman 2023)

$$\mathrm{MCC} := \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \in [-1, 1], \tag{58}$$

with $\mathrm{MCC} = 0$ when the denominator vanishes. For both metrics, higher values indicate better classification. Let $\mathrm{AC}_{\mathsf{LR}}$ and $\mathrm{MCC}_{\mathsf{LR}}$ denote the test metrics of the clean model LR, and $\mathrm{AC}_{\mathsf{PA\text{-}L}}$ and $\mathrm{MCC}_{\mathsf{PA\text{-}L}}$ those of the poisoned model returned by PA-L. The effect of the attack is measured by the drops

$$\Delta\mathrm{AC} = \mathrm{AC}_{\mathsf{LR}} - \mathrm{AC}_{\mathsf{PA\text{-}L}}, \quad \Delta\mathrm{MCC} = \mathrm{MCC}_{\mathsf{LR}} - \mathrm{MCC}_{\mathsf{PA\text{-}L}}, \tag{59}$$

computed for each partition; a positive drop indicates that the attack degraded the classifier. To assess the computational cost of the attack, we also record the wall-clock times of LR and PA-L in seconds. All quantities are reported as mean ± standard deviation over the 5 partitions.

4.3.2. *Global Optimization.* By Proposition 5, Problems (P2) and (P3) are equivalent maximization problems with optimal value z^*. The bounds of Proposition 3 do not exclude any feasible point, and those of Section 3.4 are valid up to the accuracy of Ipopt. For CAR and CAR+, the clipping of α to $[10^{-8}, 1-10^{-8}]$ is needed because SCIP cannot evaluate log at values below its tolerance. However, it restricts the feasible set, so their dual bounds refer to this restricted problem.

To assess the bounds on the variables, we report, for each set of bounds, the average width of the intervals obtained for β_j, θ_i^{tr}, ξ^p, and α_k^p, where the average is taken over all components and over the four budgets ρ. We also report the time needed to compute all bounds of Section 3.4, which is spent before SCIP is started and is not included in its time limit.

For each method f, SCIP reports a *primal bound* $\underline{z}_f$, the objective value of the best feasible solution found, and a *dual bound* $\overline{z}_f$, obtained from relaxations, so that

$$\underline{z}_f \leq z^* \leq \overline{z}_f. \tag{60}$$

In addition, (36) yields the analytic bound $z^* \leq \bar{U} := N \log 2 - q\,\phi(-R\sqrt{d})$, which requires no optimization. Since $\min\{\overline{z}_f, \bar{U}\}$ is also an upper bound on z^*, we use it in place of $\overline{z}_f$ below.

As a method may fail to find a feasible solution within the time limit, we use the best primal bound of all four methods,

$$z_{\mathrm{best}} = \max\{\underline{z}_{\mathsf{PA\text{-}G}}, \underline{z}_{\mathsf{CAR}}, \underline{z}_{\mathsf{PA\text{-}G+}}, \underline{z}_{\mathsf{CAR+}}\},$$

as a common reference, and measure the distance of each dual bound from it by

$$\mathrm{gap}_f = \frac{\min\{\overline{z}_f, \bar{U}\} - z_{\mathrm{best}}}{|z_{\mathrm{best}}|}. \tag{61}$$

By (60), a gap of 0 certifies global optimality, and a gap of, say, 0.2 guarantees that z^* exceeds the best attack found by at most 20 %. Since all methods share the same reference, their gaps are directly comparable.

To compare the dual bounds obtained for (P2) and (P3) independently of the primal bounds, we also report, for each set of bounds, the ratio

$$r = \frac{\min\{\overline{z}_{\mathsf{PA\text{-}G}}, \bar{U}\}}{\min\{\overline{z}_{\mathsf{CAR}}, \bar{U}\}}, \quad r_+ = \frac{\min\{\overline{z}_{\mathsf{PA\text{-}G+}}, \bar{U}\}}{\min\{\overline{z}_{\mathsf{CAR+}}, \bar{U}\}}. \tag{62}$$

A ratio greater than 1 means that the corresponding convex-analytic method provides the tighter upper bound on z^* within the time limit. For each method, we report $\underline{z}_f$, $\overline{z}_f$, and gap_f in percent, together with the analytic bound $\bar{U}$ and the ratios r and r_+.

4.4. **Discussion of the Results.**

4.4.1. *Attack Effectiveness.* Figures 1 and 2 show the test accuracy and MCC of the LR model and of the poisoned models returned by PA-L for each budget ρ. Table 2 reports the corresponding drops and solve times.

The attack degrades the mean accuracy in all datasets, although the mean drops are small for Smarket. The degradation generally grows with the budget ρ, reaching drops of up to 0.29 (Banknote). For Smarket, the clean model has mean accuracy AC = 0.512, and the observed mean accuracy drops range from 0.003 to 0.012. As with accuracy, MCC generally decreases with the budget ρ in all datasets except Smarket, but its drops are larger and reveal effects that accuracy hides. For College ($\rho \geq 15\,\%$) and Spambase ($\rho \geq 10\,\%$), the mean MCC of the poisoned model is close to zero.

The response to increasing ρ differs substantially between datasets. At $\rho = 5\,\%$, the mean accuracy drop is only 0.014 for Biopsy, compared with 0.132 for Banknote and 0.260 for Spambase. For Banknote, the drop continues to grow across the four budgets, whereas most of the accuracy degradation on Spambase already occurs at $\rho = 5\,\%$. The low accuracy on Smarket, however, is already present in the clean model and does not demonstrate greater susceptibility to poisoning.

A plateau in test accuracy does not imply a plateau in the attacker's objective. For Spambase, the mean objective value returned by PA-L increases from approximately 908 at $\rho = 5\,\%$ to 998 at $\rho = 20\,\%$. The attacker maximizes logistic loss on the clean training data, whereas accuracy and MCC evaluate thresholded predictions on the test data. Predicted probabilities can change without changing the predicted classes, and the training and test losses need not respond in the same way.

Possible explanations for the differences between datasets include the magnitude of the clean decision scores and the sensitivity of the fitted coefficients to poisoning. Points that are correctly classified with larger absolute decision scores (i.e., more negative θ under the convention of Section 3.1) require larger changes before their predicted labels change. The coefficient sensitivity depends on the curvature of the lower-level objective and the gradient contributions of admissible poisoning points; this relationship underlies influence-function analysis (Koh and Liang 2017). Moreover, the box constraint permits feature combinations beyond those observed in the clean data, which may give poisoning points substantial influence. These mechanisms provide hypotheses for further investigation. The present experiments do not measure their contributions and therefore do not establish a general rule relating dataset properties to attack susceptibility. Such comparisons also depend on scaling and regularization: with the summed loss used here, a fixed λ corresponds to a penalty weight $\lambda/(n+q)$ relative to the average loss on the contaminated training set.

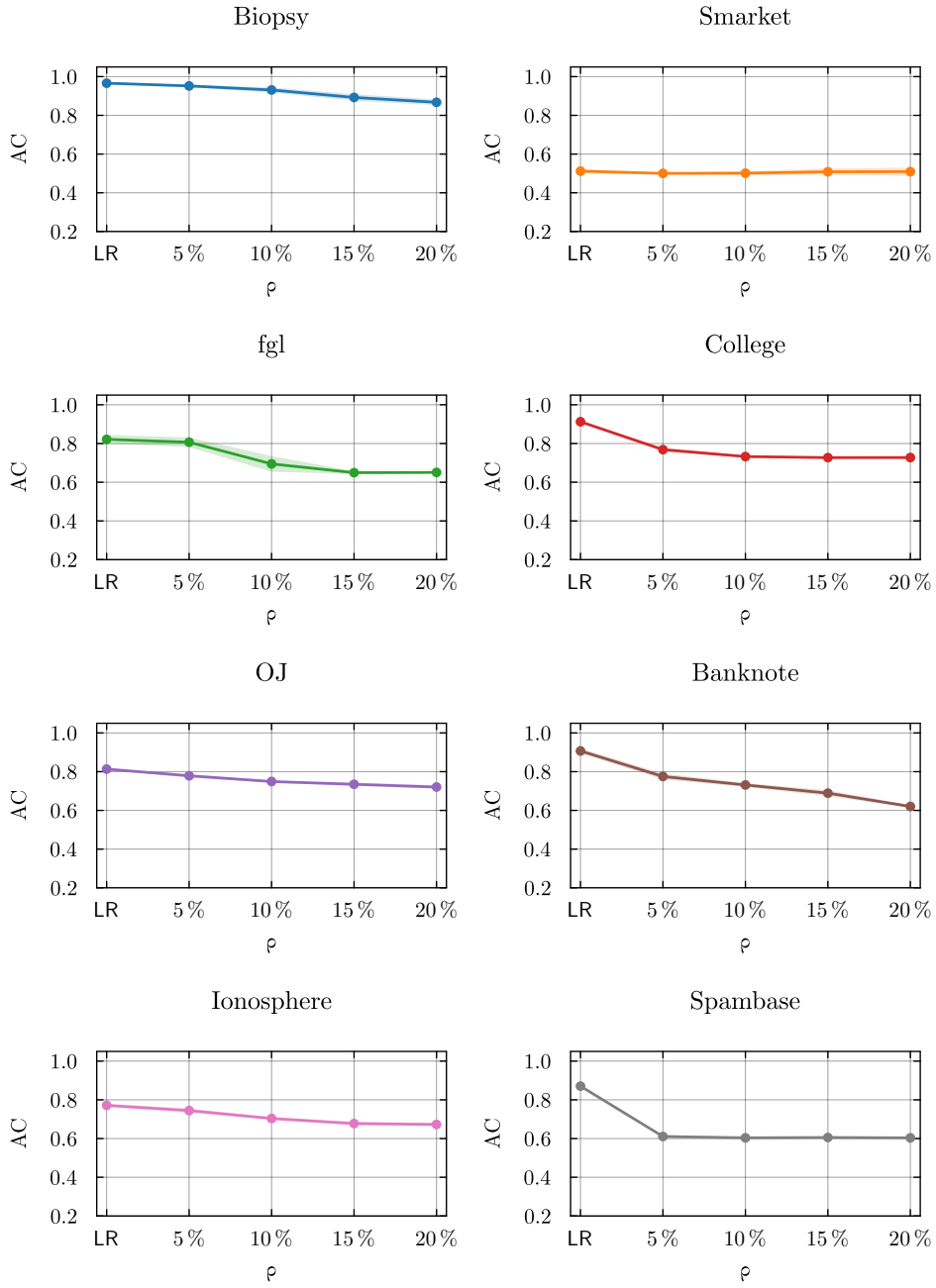


FIGURE 1. Test accuracy of the clean model (LR) and of the poisoned models returned by PA-L for each budget ρ. Markers: mean over 5 partitions; shaded areas: $\pm$ one standard deviation.

The attack is also cheap to compute. LR takes at most 0.1 s, and PA-L, including all 5 starts, at most 0.73 s on average, except for Spambase, where it takes between 8 and 10 s, and about 68 s for $\rho = 5\,\%$. We stress, however, that these results come from local solutions of (P2), with no guarantee of global optimality.

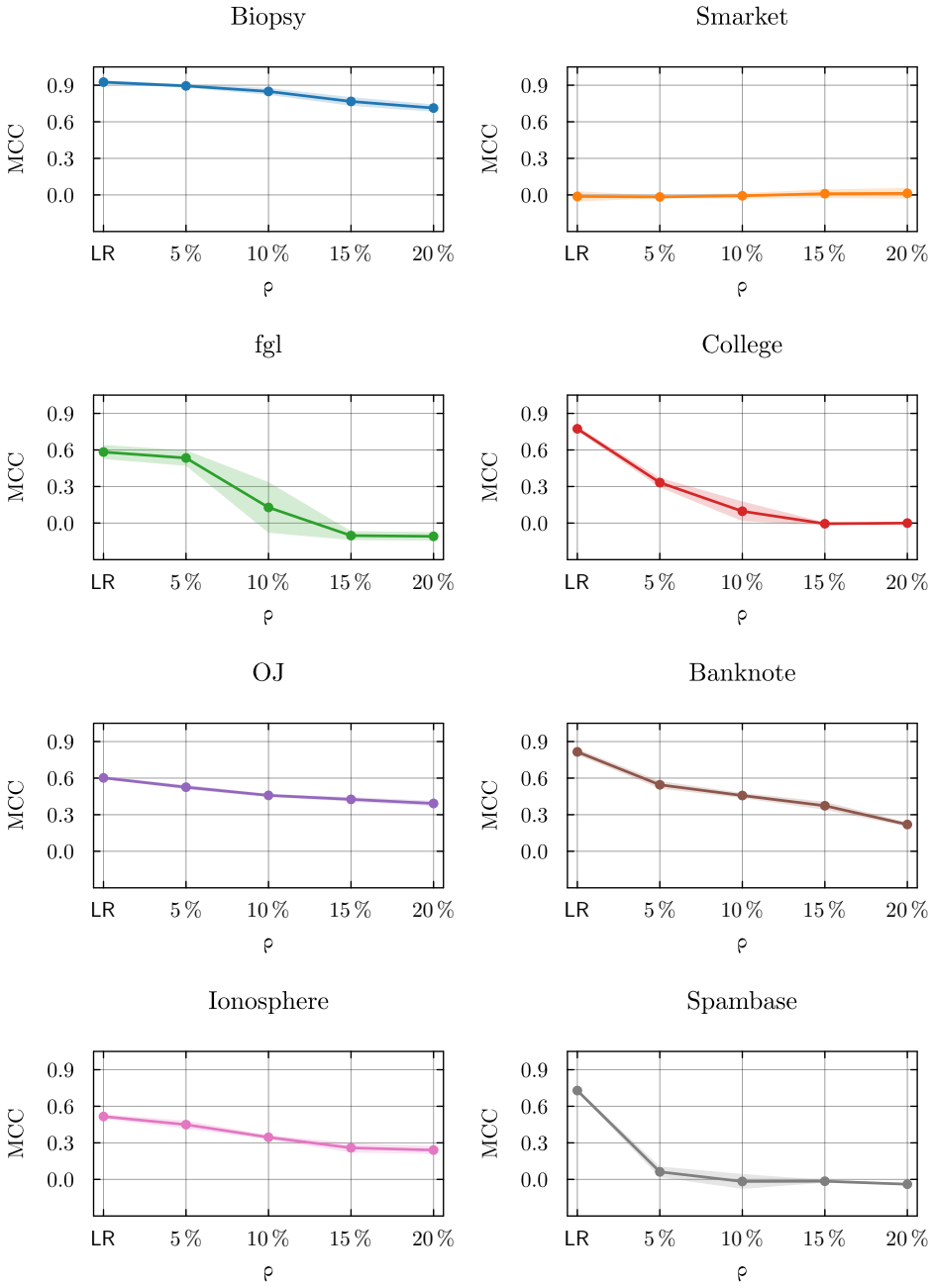


FIGURE 2. Test MCC of the clean model (LR) and of the poisoned models returned by PA-L for each budget ρ. Markers: mean over 5 partitions; shaded areas: $\pm$ one standard deviation.

4.4.2. *Global Optimization.* Table 3 reports the average width of the bounds obtained from Proposition 3 alone and from all bounds of Section 3.4. Computing all bounds takes between 0.34 and 434.88 s on average over the four budgets (last

TABLE 2. Attack effectiveness: drops ΔAC and ΔMCC as in (59) induced by PA-L, computed per partition, and solve times in seconds; PA-L times include all 5 starts. All values: mean ± standard deviation over 5 partitions.

	LR		PA-L		
Dataset	Time (s)	ρ	ΔAC	ΔMCC	Time (s)
Biopsy	0.003 ± 0.001	5 %	0.014 ± 0.003	0.031 ± 0.008	0.179 ± 0.027
		10 %	0.035 ± 0.010	0.076 ± 0.021	0.183 ± 0.022
		15 %	0.074 ± 0.015	0.158 ± 0.031	0.165 ± 0.039
		20 %	0.099 ± 0.013	0.212 ± 0.028	0.152 ± 0.019
Smarket	0.003 ± 0.001	5 %	0.012 ± 0.011	0.004 ± 0.039	0.152 ± 0.011
		10 %	0.011 ± 0.020	-0.005 ± 0.059	0.166 ± 0.021
		15 %	0.003 ± 0.025	-0.022 ± 0.073	0.197 ± 0.035
		20 %	0.003 ± 0.028	-0.025 ± 0.080	0.201 ± 0.022
fgl	0.002 ± 0.001	5 %	0.015 ± 0.019	0.049 ± 0.050	0.039 ± 0.007
		10 %	0.127 ± 0.035	0.455 ± 0.188	0.042 ± 0.007
		15 %	0.172 ± 0.029	0.685 ± 0.085	0.046 ± 0.004
		20 %	0.171 ± 0.027	0.692 ± 0.083	0.056 ± 0.006
College	0.011 ± 0.002	5 %	0.144 ± 0.014	0.441 ± 0.049	0.575 ± 0.062
		10 %	0.180 ± 0.010	0.677 ± 0.090	0.662 ± 0.072
		15 %	0.185 ± 0.006	0.779 ± 0.019	0.712 ± 0.113
		20 %	0.185 ± 0.007	0.774 ± 0.018	0.732 ± 0.081
OJ	0.006 ± 0.002	5 %	0.035 ± 0.012	0.077 ± 0.028	0.293 ± 0.022
		10 %	0.064 ± 0.009	0.144 ± 0.021	0.347 ± 0.027
		15 %	0.079 ± 0.009	0.176 ± 0.019	0.351 ± 0.020
		20 %	0.093 ± 0.010	0.210 ± 0.024	0.367 ± 0.028
Banknote	0.003 ± 0.001	5 %	0.132 ± 0.007	0.270 ± 0.015	0.207 ± 0.014
		10 %	0.176 ± 0.007	0.358 ± 0.015	0.165 ± 0.011
		15 %	0.218 ± 0.006	0.442 ± 0.016	0.137 ± 0.006
		20 %	0.287 ± 0.011	0.596 ± 0.022	0.129 ± 0.009
Ionosphere	0.011 ± 0.000	5 %	0.027 ± 0.009	0.067 ± 0.026	0.200 ± 0.014
		10 %	0.068 ± 0.013	0.171 ± 0.033	0.235 ± 0.028
		15 %	0.094 ± 0.017	0.258 ± 0.055	0.293 ± 0.026
		20 %	0.099 ± 0.014	0.276 ± 0.045	0.337 ± 0.024
Spambase	0.095 ± 0.012	5 %	0.260 ± 0.012	0.667 ± 0.052	68.059 ± 15.025
		10 %	0.267 ± 0.010	0.746 ± 0.069	8.226 ± 0.145
		15 %	0.265 ± 0.006	0.744 ± 0.010	8.520 ± 0.318
		20 %	0.267 ± 0.007	0.769 ± 0.018	10.072 ± 0.225

column of Table 3); except for Spambase, it stays below 8 s. This time comes on top of the time limit of SCIP.

The additional bounds substantially tighten the quantities associated with the training data: across the 32 individual combinations of datasets and poisoning budgets, the width of the bounds on β is reduced to between 18 % and 67 %, and that of the bounds on the margins θ^{tr} to between 5 % and 38 % of the width obtained from Proposition 3. In contrast, the quantities associated with the poisoning points are hardly affected. The bounds on ξ^p, and hence on θ^p, are unchanged in 24 of the 32 instances and only moderately tighter in the others. Since the sigmoid saturates quickly, the bounds on α^p, and therefore on h, remain essentially $[0, 1]$ in all instances. Table 4 reports the primal bounds, dual bounds, and gaps (61) of the four methods. PA-G+ finds a feasible solution in every instance, and its primal bound is within approximately 0.02 of the maximum z_{best} over the four methods, which we use as a common reference; it therefore does not depend on the solutions

TABLE 3. Average bound widths with the bounds of Proposition 3 only (Prop. 3) and with all bounds of Section 3.4 (All), and time to compute all bounds. Mean over the four budgets ρ.

	β		θ^{tr}		ξ^p		α^p		Time
Dataset	Prop. 3	All	Prop. 3	All	Prop. 3	All	Prop. 3	All	(s)
Biopsy	17.87	11.69	26.03	9.69	56.51	56.51	1.000	1.000	1.46
Smarket	24.18	6.33	38.22	2.56	63.97	44.29	1.000	1.000	4.80
fgl	9.98	4.64	15.76	3.75	31.57	31.57	1.000	1.000	0.34
College	19.10	11.75	37.88	7.92	81.02	81.02	1.000	1.000	7.34
OJ	22.36	10.63	42.49	7.23	74.17	74.17	1.000	1.000	5.39
Banknote	25.34	10.03	38.74	4.69	56.66	49.88	1.000	1.000	3.63
Ionosphere	12.85	6.77	51.90	8.18	74.95	74.95	1.000	1.000	4.59
Spambase	46.38	25.41	53.37	11.80	353.20	353.20	1.000	1.000	434.88

found by CAR+. None of the 128 runs certifies global optimality: all of them stop at the time limit of 3600 s.

On the primal side, PA-G+ finds a feasible solution in all 32 instances, PA-G in 29, CAR in 26, and CAR+ in 24; the primal bounds of CAR and CAR+ are up to 6 % below z_{best}. The methods therefore differ mainly in their dual bounds. Within the time limit, the dual bounds of PA-G and PA-G+ never improve on $\bar{U}$, so both attain the gaps given by the analytic bound, between 4.2 % and 227 %. CAR and CAR+ improve on $\bar{U}$ in all 32 instances; their gaps range from 4.2 % to 78.3 %. CAR+ and CAR attain the same displayed gap in 24 instances, and the largest difference is 7.5 percentage points, in favour of CAR (biopsy, $\rho = 15\,\%$). Compared with PA-G+, CAR+ reduces the gap by a factor of up to 4.35 (median 1.6); the improvement decreases with ρ and is negligible on Smarket.

The additional bounds lower the dual bounds of PA-G+ by about one order of magnitude, although not below $\bar{U}$, whereas those of CAR+ barely change. This difference reflects the structure of the two formulations. In (P2), the training points enter through the sigmoid and the softplus, whose relaxations improve with the much narrower bounds on θ^{tr} (Table 3). In (P3), they enter only through affine constraints and the convex functions ϕ and ϕ^*, which are relaxed well regardless of the bounds; the only nonconvexity, (P3d), involves the poisoning points, whose bounds are hardly tightened.

Figure 3 shows the dual-bound ratios (62). Both are at least 1 in every instance, so CAR+ never yields a weaker bound than PA-G+. They lie between 1.00 and 2.02 and decrease with ρ in every dataset except Smarket, where they remain at 1.00; on Spambase, they are at most 1.04. r and r_+ differ by at most 0.05.

5. CONCLUSION

We have developed a convex-analytic reformulation for poisoning attacks against ℓ_2-regularized logistic regression. By exploiting convex conjugacy and a change of variables, we obtain a concave maximization objective and affine stationarity conditions, with nonconvexity isolated in bounded bilinear constraints. We derive valid variable bounds and convex preprocessing problems for tightening them.

Our experiments on eight binary classification datasets show that attacks obtained by local optimization degrade predictive performance on most datasets. The convex-analytic formulation substantially improves the dual bounds reported by SCIP compared with the direct stationarity formulation, while the reported primal bounds are similar. Nevertheless, no run certifies global optimality within the time limit of 3600 s.

TABLE 4. Primal bounds $\underline{z}_f$, dual bounds $\overline{z}_f$ (SCIP, time limit 3600 s), analytic bound $\bar{U}$, and gaps (61) in %. Smallest gap of each instance in bold; "–" = no feasible solution found.

		Biopsy	Smarket	fgl	College	OJ	Banknote	Ionosphere	Spambase
$\rho = 5\,\%$									
$\underline{z}_f$	PA-G	45.55	261.98	33.37	102.44	150.43	209.98	52.84	912.97
	CAR	43.83	260.37	33.37	102.45	150.44	207.11	52.85	–
	PA-G+	45.55	261.98	33.37	102.44	150.43	209.98	52.84	912.97
	CAR+	45.55	261.98	33.37	102.45	150.44	210.00	52.85	–
$\overline{z}_f$	PA-G	3016.9	9190.6	475.4	6411.4	10223.3	9856.9	3843.3	48918.4
	CAR	73.8	272.9	36.5	126.1	180.4	248.6	62.5	967.4
	PA-G+	661.5	330.5	63.7	867.1	967.9	607.3	373.2	6728.7
	CAR+	73.8	272.9	36.4	126.1	180.4	248.8	62.5	967.4
	$\bar{U}$	149.0	273.1	46.4	170.5	233.6	300.1	76.9	1004.4
gap (%)	PA-G	227.1	**4.2**	39.2	66.4	55.3	42.9	45.6	10.0
	CAR	62.1	**4.2**	9.5	**23.1**	**19.9**	**18.4**	**18.3**	**6.0**
	PA-G+	227.1	**4.2**	39.2	66.4	55.3	42.9	45.6	10.0
	CAR+	**62.0**	**4.2**	**9.0**	**23.1**	**19.9**	18.5	**18.3**	**6.0**
$\rho = 10\,\%$									
$\underline{z}_f$	PA-G	57.66	264.21	35.17	117.35	166.57	233.66	60.64	–
	CAR	57.67	260.40	35.17	117.36	166.58	233.68	60.65	–
	PA-G+	57.66	264.21	35.17	117.35	166.57	233.66	60.64	951.82
	CAR+	57.67	264.22	35.17	117.36	166.58	233.68	60.65	–
$\overline{z}_f$	PA-G	3606.1	9534.6	571.0	6446.8	10516.6	10227.6	3907.6	50069.5
	CAR	101.3	286.1	40.6	146.6	207.7	286.1	72.3	1024.0
	PA-G+	886.7	395.3	88.5	916.3	974.2	697.1	425.5	7424.0
	CAR+	98.9	286.1	40.4	146.6	207.7	285.0	72.3	1024.0
	$\bar{U}$	156.0	286.3	48.5	178.1	244.7	314.0	81.1	1052.2
gap (%)	PA-G	170.4	**8.3**	38.0	51.8	46.9	34.4	33.7	10.5
	CAR	75.7	**8.3**	15.3	**24.9**	**24.7**	22.4	**19.2**	**7.6**
	PA-G+	170.4	**8.3**	38.0	51.8	46.9	34.4	33.7	10.5
	CAR+	**71.5**	**8.3**	**14.8**	**24.9**	**24.7**	**22.0**	**19.2**	**7.6**
$\rho = 15\,\%$									
$\underline{z}_f$	PA-G	69.63	266.07	37.04	128.34	178.98	253.10	64.65	–
	CAR	65.42	265.68	37.05	128.35	–	253.12	64.65	–
	PA-G+	69.63	266.07	37.04	128.34	178.98	253.10	64.65	977.53
	CAR+	69.49	265.98	37.05	–	–	252.29	64.65	–
$\overline{z}_f$	PA-G	3722.0	10055.4	627.1	7028.3	10521.7	10554.1	4074.1	51194.8
	CAR	119.0	298.6	45.7	162.2	230.1	316.6	78.2	1072.9
	PA-G+	937.4	456.8	108.5	972.4	1063.7	753.5	473.9	8239.4
	CAR+	124.2	298.5	45.7	162.2	230.1	316.7	78.2	1072.9
	$\bar{U}$	163.6	298.7	51.3	186.5	255.8	328.6	84.6	1100.0
gap (%)	PA-G	134.9	12.3	38.5	45.3	42.9	29.8	30.8	12.5
	CAR	**70.8**	**12.2**	**23.4**	**26.3**	**28.6**	**25.1**	**20.9**	**9.8**
	PA-G+	134.9	12.3	38.5	45.3	42.9	29.8	30.8	12.5
	CAR+	78.3	**12.2**	**23.4**	**26.3**	**28.6**	**25.1**	**20.9**	**9.8**
$\rho = 20\,\%$									
$\underline{z}_f$	PA-G	78.77	267.85	38.34	136.93	189.30	268.10	67.49	–
	CAR	78.78	261.53	38.34	136.94	–	268.12	67.50	–
	PA-G+	78.77	267.85	38.34	136.93	189.30	268.10	67.49	998.70
	CAR+	78.78	–	38.34	135.25	–	260.29	67.50	–
$\overline{z}_f$	PA-G	3814.9	10424.3	613.8	7235.6	11051.3	10867.7	4152.8	52295.9
	CAR	134.1	311.8	49.0	174.2	249.1	337.1	83.3	1141.2
	PA-G+	981.4	517.3	119.8	1038.8	1156.5	811.6	499.4	8848.3
	CAR+	134.1	311.8	49.0	174.2	249.1	337.1	83.3	1122.1
	$\bar{U}$	170.5	311.9	53.4	194.8	266.9	342.4	88.0	1147.9
gap (%)	PA-G	116.4	16.5	39.2	42.2	41.0	27.7	30.4	14.9
	CAR	**70.3**	**16.4**	**27.8**	**27.2**	**31.6**	**25.7**	**23.5**	14.3
	PA-G+	116.4	16.5	39.2	42.2	41.0	27.7	30.4	14.9
	CAR+	**70.3**	**16.4**	**27.8**	**27.2**	**31.6**	**25.7**	**23.5**	**12.4**

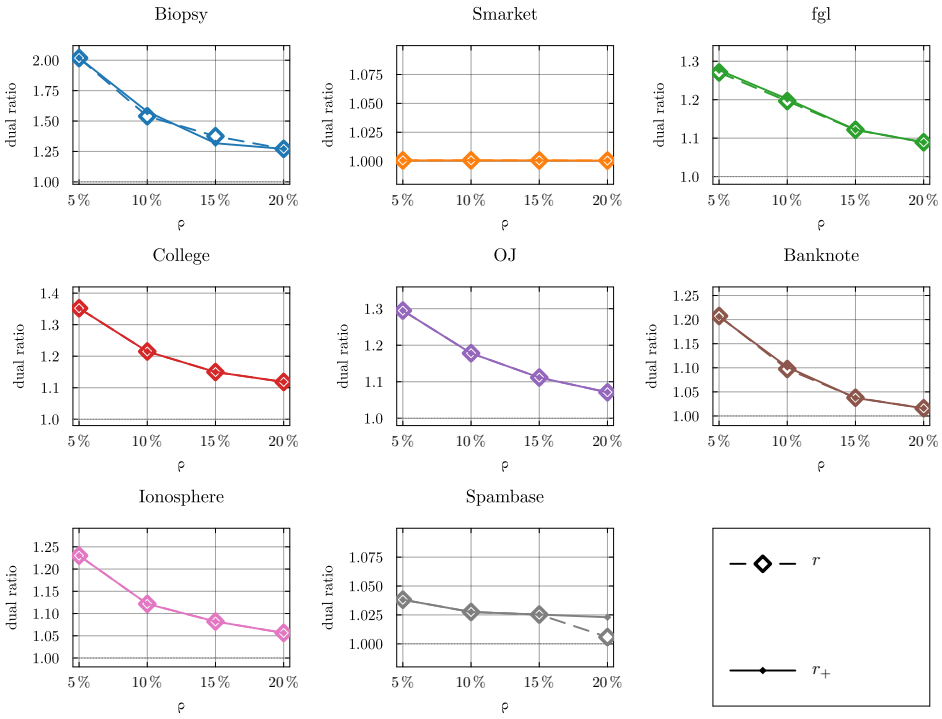


FIGURE 3. Dual-bound ratios (62) for each budget ρ: r (hollow markers) and r_+ (filled markers) as in (62). Note the different scales of the vertical axes.

Future research includes deriving tighter bounds on the poisoning variables and strengthening the relaxations of the remaining bilinear terms to reduce the optimality gaps. The reformulation relies only on the convex conjugate of the loss and on the stationarity conditions of the lower-level problem, so it may extend to other classifiers trained by regularized convex loss minimization, such as multinomial logistic regression or other generalized linear models. Finally, global bounds on the damage an attacker can cause are a natural starting point for defenses against poisoning. They can certify the worst-case effect of a given budget and guide the design and evaluation of robust training methods.

USE OF AI

OpenAI's Codex was used to assist with mathematical and editorial review, consistency checks between the manuscript and the accompanying code, and the drafting of text. Anthropic's Claude Code was used to help with the implementation of the models. The authors retain full responsibility for verifying all AI-assisted content and for the final manuscript.

REFERENCES

Arel-Bundock, V. (2026). *Rdatasets: A collection of datasets originally distributed in various R packages*. URL: https://vincentarelbundock.github.io/Rdatasets/ (visited on 10/29/2026).

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2017). "Julia: A Fresh Approach to Numerical Computing." In: *SIAM Review* 59.1, pp. 65–98. DOI: 10.1137/141000671.

Biggio, B., B. Nelson, and P. Laskov (2012). "Poisoning Attacks against Support Vector Machines." In: *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, pp. 1467–1474. URL: https://icml.cc/2012/papers/880.pdf.

Caballero, W. N., M. LaRosa, A. A. Fisher, and V. Tarokh (2025). "Indiscriminate disruption of conditional inference on multivariate Gaussians." In: *European Journal of Operational Research* 327.1, pp. 191–202. DOI: 10.1016/j.ejor.2025.06.011.

Carnerero-Cano, J., L. Muñoz-González, P. Spencer, and E. C. Lupu (2021). "Regularization Can Help Mitigate Poisoning Attacks... with the Right Hyperparameters." In: *ICLR Workshop on Security and Safety in Machine Learning Systems*. URL: https://aisecure-workshop.github.io/aml-iclr2021/papers/20.pdf.

Chicco, D. and G. Jurman (2023). "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification." In: *BioData Mining* 16.1, 4. DOI: 10.1186/s13040-023-00322-4.

Dunning, I., J. Huchette, and M. Lubin (2017). "JuMP: A Modeling Language for Mathematical Optimization." In: *SIAM Review* 59.2, pp. 295–320. DOI: 10.1137/15M1020575.

Gambella, C., B. Ghaddar, and J. Naoum-Sawaya (2021). "Optimization problems for machine learning: A survey." In: *European Journal of Operational Research* 290.3, pp. 807–828. DOI: 10.1016/j.ejor.2020.08.045.

Guedes-Ayala, M., L. Schewe, Z. Şuvak, and M. Anjos (2025). *Generating Poisoning Attacks against Ridge Regression Models with Categorical Features*. arXiv preprint arXiv:2501.07275. DOI: 10.48550/arXiv.2501.07275.

Gurobi Optimization, LLC (2025). *Gurobi Optimizer Reference Manual*. Version 13.0. URL: https://docs.gurobi.com/projects/optimizer/en/13.0/index.html.

Hojny, C., M. Besançon, K. Bestuzheva, S. Borst, A. Chmiela, J. Dionísio, J. Ehls, L. Eifler, M. Ghannam, A. Gleixner, A. Göß, A. Hoen, J. von Holly-Ponientzietz, R. van der Hulst, D. Kamp, T. Koch, K. Kofler, J. Lentz, M. Lübbecke, S. J. Maher, P. M. Meinhold, G. Mexi, T. Mohr, E. Mühmer, K. K. Patel, M. E. Pfetsch, S. Pokutta, C. Reinartz Groba, F. Serrano, Y. Shinano, M. Turner, S. Vigerske, M. Walter, D. Weninger, and L. Xu (2025). *The SCIP Optimization Suite 10.0*. arXiv preprint arXiv:2511.18580. DOI: 10.48550/arXiv.2511.18580.

Hopkins, M., E. Reeber, G. Forman, and J. Suermondt (1999). *Spambase*. UCI Machine Learning Repository. DOI: 10.24432/C53G6X.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). *An Introduction to Statistical Learning: with Applications in R*. 2nd ed. New York: Springer. DOI: 10.1007/978-1-0716-1418-1.

Kelly, M., R. Longjohn, and K. Nottingham (2023). *The UCI Machine Learning Repository*. URL: https://archive.ics.uci.edu/ (visited on 10/29/2026).

Koh, P. W. and P. Liang (2017). "Understanding Black-box Predictions via Influence Functions." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1885–1894. URL: https://proceedings.mlr.press/v70/koh17a.html.

Koh, P. W., J. Steinhardt, and P. Liang (2022). "Stronger Data Poisoning Attacks Break Data Sanitization Defenses." In: *Machine Learning* 111.1, pp. 1–47. DOI: 10.1007/s10994-021-06119-y.

McCormick, G. P. (1976). "Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems." In: *Mathematical Programming* 10.1, pp. 147–175. DOI: 10.1007/BF01580665.

Mei, S. and X. Zhu (2015). "Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1, pp. 2871–2877. DOI: 10.1609/aaai.v29i1.9569.

Steinhardt, J., P. W. Koh, and P. Liang (2017). "Certified Defenses for Data Poisoning Attacks." In: *Advances in Neural Information Processing Systems*. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html.

Şuvak, Z., M. F. Anjos, L. Brotcorne, and D. Cattaruzza (2026). "Design of Poisoning Attacks on Linear Regression Using Bilevel Optimization." In: *Data Science and Optimization*. Ed. by S. Pokutta. Vol. 91. Fields Institute Communications. Cham: Springer, pp. 41–65. DOI: 10.1007/978-3-032-03844-9_2.

Suya, F., S. Mahloujifar, A. Suri, D. Evans, and Y. Tian (2021). "Model-Targeted Poisoning Attacks with Provable Convergence." In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 10000–10010. URL: https://proceedings.mlr.press/v139/suya21a.html.

Wächter, A. and L. T. Biegler (2006). "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming." In: *Mathematical Programming* 106.1, pp. 25–57. DOI: 10.1007/s10107-004-0559-y.

Xiao, H., B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli (2015). "Is Feature Selection Secure against Training Data Poisoning?" In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. PMLR, pp. 1689–1698. URL: https://proceedings.mlr.press/v37/xiao15.html.

(C. M. Costa) State University of Maringá, Department of Mathematics, Maringá, Brazil
Email address: cmcosta@uem.br

(M. E. Pinheiro) Federal University of Santa Catarina, Department of Mathematics, Blumenau Campus, Blumenau, Brazil
Email address: maria.eduarda.pinheiro@ufsc.br

(M. Schmidt) Trier University, Department of Mathematics, Universitätsring 15, 54296 Trier, Germany
Email address: martin.schmidt@uni-trier.de