Proceedings of Machine Learning Research 299:1–27, 2025 Conference on Applied Machine Learning for Information Security

Adversarial Machine Learning Attacks on Financial Reporting via Maximum Violated Multi-Objective Attack

Edward Raff EDWARD.RAFF@CROWDSTRIKE.COM
CrowdStrike
University of Maryland, Baltimore County

Karen Kukla KAKUKLA@SYR.EDU
Syracuse University

Michel Benaroch MBENAROC@SYR.EDU
Syracuse University

Joseph Comprix JJCOMPRI@SYR.EDU
Syracuse University

Editor: Edward Raff and Ethan M. Rudd

Abstract

Bad actors, primarily distressed firms, have the incentive and desire to manipulate their financial reports to hide their distress and derive personal gains. As attackers, these firms are motivated by potentially millions of dollars and the availability of many publicly disclosed and used financial modeling frameworks. Existing attack methods do not work on this data due to anti-correlated objectives that must both be satisfied for the attacker to succeed. We introduce Maximum Violated Multi-Objective (MVMO) attacks that adapt the attacker's search direction to find 20× more satisfying attacks compared to standard attacks. The result is that in ≈ 50% of cases, a company could inflate their earnings by 100-200%, while simultaneously reducing their fraud scores by 15%. By working with lawyers and professional accountants, we ensure our threat model is realistic to how such frauds are performed in practice.

1. Introduction

Given a target function $f(\cdot)$ that takes an input $\boldsymbol{x}$ and output $\hat{y}$, it has been found that it is surprisingly easy for an adversary $\mathcal{A}$ to produce a perturbed $\tilde{\boldsymbol{x}}$ that is trivially similar to input x, such that $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \leq \epsilon$, and yet $f(\tilde{\boldsymbol{x}}) \neq \hat{y}$. This has been achieved primarily for deep learning algorithms in classification tasks, such as computer vision (predict the label) and natural language processing (predict the next token).

In this work, we seek to study a new problem space for adversarial machine learning (AML): *fraudulent financial reporting*. This task has many differences from standard applications that make it interesting from a purely machine-learning perspective while simultaneously being highly relevant to real-world use (i.e., regulators need to know the extent of what is possible). Financial reporting fraud already occurs and is motivated by millions-to-billions of dollars in potential impact. Modeling frameworks used in this space are primarily linear in nature, making them a significantly different surface than most prior work, and the outputs y are usually real-valued regression problems instead of classification.

The decisions of investors, banks, and regulators can have significant impacts on a company in receiving new capital, loans, and scrutiny respectively. Each party makes these

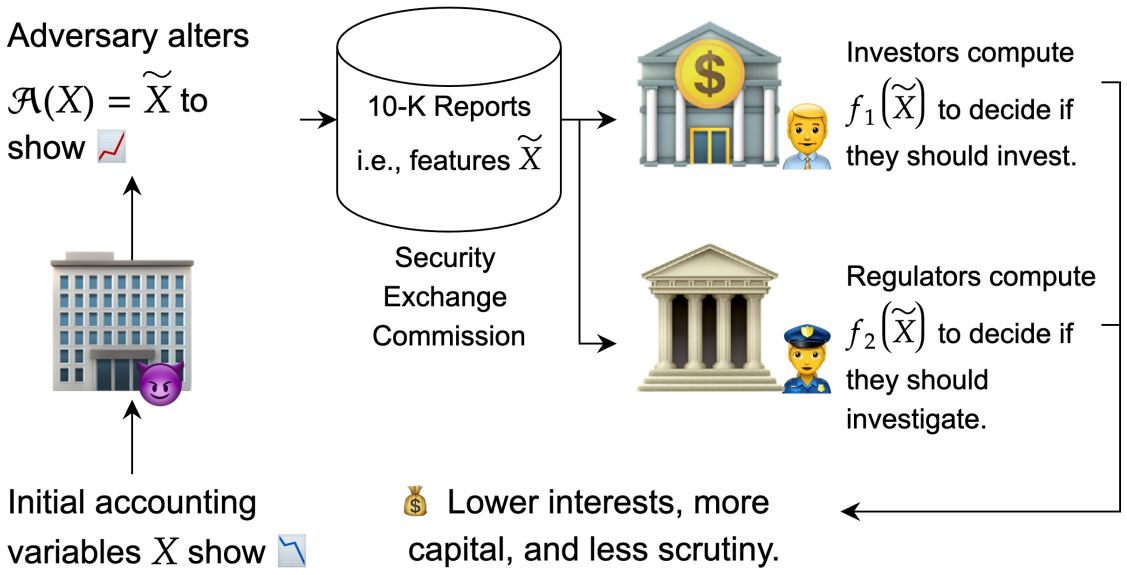

Figure 1: The overall threat model and scenario of this work. The adversary $\mathcal{A}$ is the (distressed) firm, which must file reports at a central database (run by the SEC). There are two models that need to be manipulated to achieve the adversarial firm's goals. These reports are used by investors to compute $f_1(\cdot)$ to decide if they should purchase shares or offer better loan terms (i.e., firm gains capital). Regulators also use these reports to run a fraud model $f_2(\cdot)$, to determine if investigation or other regulatory scrutiny of the firm is necessary.

decisions in large part using models based on financial ratios (Altman, 1968), which are computed from a financial statement called a 10-K. This 10-K is filed with the Security and Exchange Commission (SEC) every year and is produced *by the company itself*. Thus, the company has the motivation and means to manipulate its financial reporting.

To wit, our work is the first to study AML as applied to this problem. It is critical from a social perspective to understand the strengths and limitations of the current regulatory frameworks, i.e., Generally Accepted Accounting Principals (GAAP), with respect to potential risks as machine learning techniques become more widespread. Working with adversarial ML experts, two Certified Public Accountants (CPAs), and a lawyer, we build accurate and realistic threat models to apply AML methods to. Ultimately, we show positive and negative results for current GAAP-based accounting: in half of cases, adversaries can have extraordinary success in evading detection while inflating earnings, but in the other half of cases, they can only achieve one of these goals.

We consider two sets of goals the adversary wishes to achieve: avoiding existing fraud models and inflating their reported earnings. Both goals are numeric targets y with differing scales over orders of magnitude, making existing attack approaches ineffective. To demon-

strate that an attack is still possible, we develop a novel paired numeric optimization that seeks to obtain an equal relative improvement in both goals. Working with domain expert accountants and lawyers, we develop a realistic threat model that defines an action-space for the threat model.

Our work uses two correlated metrics as exemplars: the earnings per share and a fraud model called M-score, which are prevalent in practical use. From a practical perspective, we show that the current environment is at risk of successful adversarial attacks where earnings are inflated by over 100% while simultaneously reducing the fraud risk scores by 15%. From an AML perspective, current techniques for adversarial regression do not succeed in this task, and we show how a multi-task regression attack can be constructed. Critically, it abides by the constraint that both earnings must go up, and the fraud score must go down, for the attack to be viable.

This article is organized as follows. First, we provide a brief primer on the relevant financial and accounting details needed to understand this work with related work in § 1.1. Given this understanding, we can detail the threat model and its goals, attack strategy, and action space in § 2. Our novel MVMO attack is defined in §3. The results of our attacks are demonstrated on real-world financial data in § 4, with ablation of adding a third objective to demonstrate MVMO's ability to handle additional targets. Finally, we conclude in § 6.

This work is intrinsically interdisciplinary and thus covers a wide scope of material. We will first take a moment to review the context of the accounting background needed to understand the work and the accounting terminology that is necessary.

1.1. Accounting Context

The set of players and goals in our work is outlined in Figure 1. There are two sets of actors in this work. *First is the company (a.k.a. the firm)* that may engage in fraudulent reporting of its financial status, which tends to occur when a firm's financial health degrades (Stolowy and Breton, 2004; Rosner, 2003). All publicly traded companies (in the United States, though most nations have a counterpart) must file a set of financial documents every fiscal year called a *10-K*. A 10-K is freely available to anyone online and contains a balance sheet and income statement. These documents summarize the current net value of the firm and how the firm's net value changed (via revenue and costs) over the past year. *Second are the investors (and regulators/lenders)*, who use the raw features from the 10-K to compute their own models, comparisons, and thus decisions based on the 10-K reports.

From each 10-K there are several variables that are reported, with Generally Accepted Accounting Principles (GAAP) that dictate how companies should allocate expenses and revenues, both cash and debts, into the 10-K so that investors, banks, analysts, and others have a uniform and repeatable interface to judge the financial health of a company and compare it with competing firms. All variables involved in this work are provided in Table 1 with industry abbreviations that will be used.

As the name *balance* sheet implies, it is important that a number of the variables balance out to the same total value. Thus, one can not arbitrarily set every variable in a 10-K, because the numbers would not add up to the same total value. For the purposes of this work, a complete list of all variables that must balance, and how they interact, is given in Figure 2.

Table 1: The set of all common accounting variables extracted from a 10-K that are used in this work. A complete understanding of all variables is not necessary and is provided as a reference so that the common accounting abbreviation can be expanded to its full definition.

Abreviation	Variable
AT	Total Assets
ACT	Current Assets - Total
RECT	Net Receivables
INVT	Inventory
PPENT	Net Property, Plant, & Equipment
PPEGT	Gross Total Property, Plant, & Equipment
LT	Liabilities - Total
LCT	Current Liabilities
DLTT	Long-term debt - Total
NI	Net Income
SALE	Net Sales
COGS	Cost of Goods Sold
DP	Depreciation
AM	Amortization
XSGA	Selling, General and Administrative expenses
OANCF	Operating Activities - Net Cash
DVP	Dividends, Preferred
XSTF	Staff Expenses - Total
CSHO	Common Shares Oustanding
XAGT	Administrative and General Expenses - Total
XEQO	Equipment and occupancy Expense
XOPR	Operating Expense

Every leaf node of the tree in Figure 2 we shall term an *atom*, in that it can be directly manipulated. However, manipulating any atom will alter all remaining variables further up in the hierarchy. For example, increasing the DLTT will also increase the LT by the same amount, because DLTT is a descendant of LT.

A company self-reports all values in the 10-K to the appropriate regulators, and so the 10-K variables are the raw features that investors use to make decisions. This is commonly done by creating *ratios* constructed from the various variables. For example, the *current ratio* is defined as the current assets (AT) divided by current liabilities (LT). While non-linearities can be used in computing a "ratio", it is common for all terms to be addition, subtraction, multiplication, and division (Piotroski, 2000).

The use of these ratios constitutes effective feature engineering on the part of the investors. Linear and logistic regressions, or the ratio itself, are used by investors/regulators as they are interpretable and common machine learning has not produced sufficiently superior results (Höglund, 2012; Sanad, 2021). Thus, there is an actor (the firm) that has a

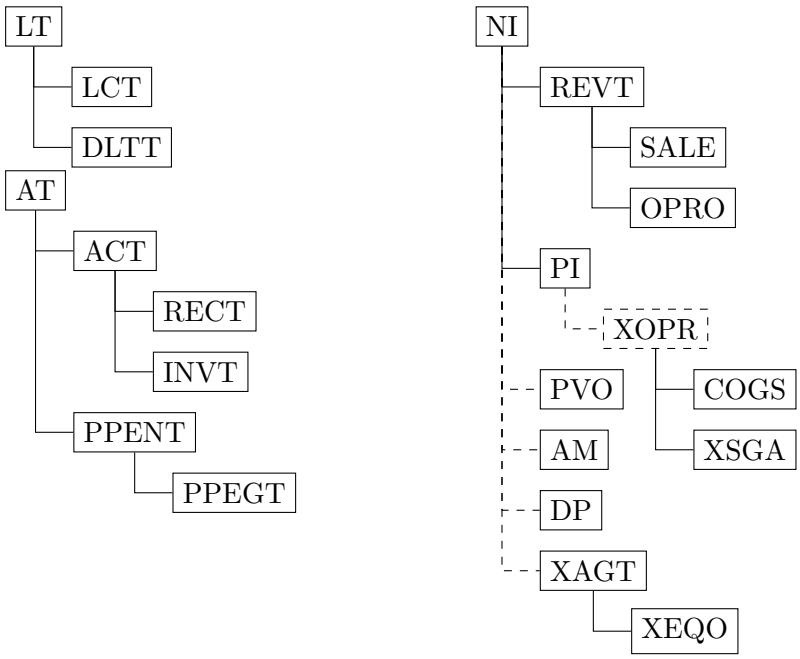

Figure 2: Solid line indicates a value-added into the parent value, dashed lines a value subtracted from the parent value. This is not a complete diagram, and some of the leaf nodes may actually have more children. For this work, we ignore the children when it is an unnecessary detail.

method to alter the features (financial variables) used by other parties (the investors) with significant motivation.

1.2. Related Work

Since the seminal work of (Healy, 1985) in 1985, fraud models in accounting have been predominantly based around linear models on manually engineered features of accounting variables, termed accounting ratios in practice. The vast majority of fraud models in use are thus followers of this overall style of linear models trained on a small set of known fraud cases against a similarly sized reference population and then applied to future years/companies (Jones, 1991; DeFond and Jiambalvo, 1994; Dechow et al., 1995; Dechow and Dichev, 2002; Hribar and Collins, 2002; Rosner, 2003; Spathis, 2002). The primary fraud model we will study is the M-score (Beneish, 1997). Despite its age, it is one of the most widely used fraud models. It has been used reliably over decades for financial modeling/portfolio management (Beneish and Nichols, 2007; Beneish et al., 2012), detected real-world incidents of fraud (noa, 1998; Ramírez-Orellana et al., 2017), and is still actively modified/used as a benchmark in current accounting literature (Narsa et al., 2023; Lu and Zhao, 2020). As such, we use it as our target model to evade as a major representative of the models in use and one of the most prevalent types of models, if not individual models, used in practice. While linear models have been widely used in machine learning for

high-dimensional problems and cases where privacy/provable guarantees are needed Khanna et al. (2024); Wu et al. (2024); Lu et al. (2022); Khanna et al. (2025); Raff et al. (2023b,b); Khanna et al. (2023); Ng (2004); Zou and Hastie (2005); Fan et al. (2008); Baracaldo et al. (2017); Liu et al. (2017), their use in accounting is unique and interesting in the dependence on domain expert feature engineering that is not as prevalent in machine-learning use cases. Similarly, Bayesian probabilistic programming is often used to encode domain knowledge into the modeling of the algorithm, rather than the feature stage Klein et al. (2024); Rubin (1984); Gelman et al. (2013); Gelman (2006). The results we are are relatively robust compared to current computer vision Biggio and Roli (2018); Doldo et al. (2025); Cinà et al. (2024); Floris et al. (2023); Bryniarski et al. (2021); Rahnama et al. (2020); Carlini and Wagner (2017) and natural language processing Mehrotra et al. (2024); Das et al. (2025); Carlini et al. (2021, 2018) susceptibility to adversarial attacks, despite not being developed for such attacks, is a strong indicator of the intrinsic value in accounting's focus on such feature engineering via accounting ratios and the importance of considering real-world or "problem space" attacks Apruzzese et al. (2023); Raff et al. (2023a); Pierazzi et al. (2020).

From an AML perspective, our work deals with regression and multiple objectives. The use of AML in regression has received minimal study, as noted in (Nguyen and Raff, 2019; Gupta et al., 2021). This poses issues when the gradient can change by orders of magnitude because the response changes by orders of magnitudes (Kong and Ge, 2023), but is uniquely different in that the magnitude is an artifact of the information of the domain rather than an "obfuscated gradient" that confounds many defensive works in classification problems (Athalye et al., 2018).

In addition, little work considers multiple attack objectives simultaneously. Prior works either use a simple weighted average of loss terms (Qin et al., 2019) and are designed for related/correlated objectives (Williams and Li, 2023; Bui et al., 2023; Yang et al., 2024), where our work has anti-correlated objectives that must be satisfied.

2. Threat Model

Following (Biggio et al., 2014) we will now specify the threat model of our study. First is the high-level inputs, followed by the attacker's optimization targets. This includes a novel formulation of the adversarial regression problem to optimize two variables that must both be attacked successfully for the attack to be viable in practice. Then we will review the strategies that can be used to perturb the variables from the 10-K in a realistic manner.

For the discussion, we will be using ratios that compare the current year to a previous year (denoted by a subscript of $_{t-1}$). There will thus be T different years of 10-K values and $T-1$ computed scores (excluding the first year). Each company has a different value of T based on data availability. Three matrices are used to encode the data and attack. $X \in \mathbb{R}^{T,D}$ is the history of T different financial reports of D different variables. $P \in \mathbb{R}^{T,L}$ is the perturbation amounts applied to each of the T different financial reports over time. There are L options for each point in time that indicate which of L different perturbation strategies will be used. $M \in \mathbb{R}^{L,D}$ is the sparse matrix of L perturbation strategies mapped to the D different variables that may be impacted.

2.1. Attack Optimization Targets

Here we detail what the target function to attack, $f(\cdot)$ is. Broadly, many possible targets could be the function of an adversarial attack with the goal of performing fraud. In financial literature, these may often be found by referring to *earnings management* (EM). EM is not synonymous with fraud but covers a broad spectrum of financial decisions and accounting decisions that may be made with an eye toward outcomes more appealing to investors and other parties (e.g., a bank determining loan risk/rates). It is worth noting thought that EM can become fraud if pushed too far, and the scope of all possible EM that could be performed using adversarial ML is not the scope of this article.

Our focus is on the matter of two representative targets that a bad actor may target in performing financial fraud. We will detail each individually, which can be represented by Equation 1. As we will demonstrate later in section 4, the straightforward attack of either objective does not result in a holistic strategy that an adversary is likely to employ. For this reason, we will develop a novel strategy to combine both numeric objects into a single optimization target.

$$\arg\min_{P} f(X + X \odot PM) \quad s.t. \quad |P| \leq \epsilon \tag{1}$$

2.1.1. Financial Target

The first, and intuitive target, is the earnings per share (EPS) of the company, defined by as the Net Income (NI) minus cash paid out to investors in the form of preferred dividends (DVP), divided by the number of common shares outstanding (CSHO).

$$\text{Earnings Per Share}(EPS) = \frac{NI - DVP}{CSHO} \tag{2}$$

Generally, the higher the EPS, the more attractive the stock looks to investors. This has multiple macro-incentives at the institution level: being able to raise more capital via the sale of shares, more leverage for the purchase of other companies, and more favorable loan terms from banks. On an individual level the potential bad actor, any actor who has meaningful compensation in the form of stock grants will stand to directly benefit from a higher EPS calculation.

2.1.2. Evasion Target: M-score

The high incentives to better earnings reports by means of manual human effort in creating fraudulent 10-Ks has been a long-standing issue, and it can take years for fraud to be detected (if at all). For this reason, many have developed models for attempting to detect the risk of fraud or earnings manipulation. We select the M-Score (Beneish, 1997, 1999) as a representative model to try and evade because the M-score has been in use for several decades (Beneish and Nichols, 2007, 2009; Beneish et al., 2012) and was even used by business students to identify Enron's collapse a year in advance (noa, 1998). The model calculates an M-score based on a number of known and custom financial ratios (i.e., manual feature engineering) computed from the 10-K statements on a year-over-year basis. The M-score indicates the degree of manipulation in earnings by a company. For example, a score of

-2.50 suggests a low likelihood of manipulation, whereas a score exceeding -1.78 suggests that the company is likely to be a manipulator. Like most models developed in the financial literature (Piotroski, 2000), it is a linear model over the ratios. In this particular case, there are eight ratios with corresponding covariates (and bias term) defined as Equation 3.

$$\begin{aligned}\text{M-Score} = &-4.84 + 0.92 \cdot DSRI + 0.528 \cdot GMI + \\ &0.404 \cdot AQI + 0.892 \cdot SGI + 0.115 \cdot DEPI \\ &- 0.172 \cdot SGAI + 4.679 \cdot TATA - 0.327 \cdot LVGI\end{aligned} \tag{3}$$

The eight ratios of the M-score are defined below. Due to space limitations, we will not review each individually.

In each equation, $t-1$ is used to represent the previous year's value of the variable, and no subscript is used for the current year.

$$\text{Days Sales In Receivables Index } (DSRI) = \frac{\frac{RECT}{SALE}}{\frac{RECT_{t-1}}{SALE_{t-1}}}$$

$$\text{Gross Margin Index } (GMI) = \frac{\frac{SALE_{t-1} - COGS_{t-1}}{SALE_{t-1}}}{\frac{SALE - COGS}{SALE}}$$

$$\text{Asset Quality Index } (AQI) = \frac{1 - \left(\frac{ACT + PPENT}{AT}\right)}{1 - \left(\frac{ACT_{t-1} + PPENT_{t-1}}{AT_{t-1}}\right)}$$

$$\text{Sales Growth Index } (SGI) = \frac{SALE}{SALE_{t-1}}$$

$$\text{Depreciation Index } (DEPI) = \frac{\frac{DP_{t-1} + AM_{t-1}}{DP_{t-1} + AM_{t-1} + PPENT_{t-1}}}{\frac{DP + AM}{DP + AM + PPENT}}$$

$$\text{SGA Index } (SGAI) = \frac{\frac{XSGA}{SALE}}{\frac{XSGA_{t-1}}{SALE_{t-1}}}$$

$$\text{Leverage Index } (LVGI) = \frac{\frac{DLTT + LCT}{AT}}{\frac{DLTT_{t-1} + LCT_{t-1}}{AT_{t-1}}}$$

$$\text{Total Accruals to Total Assets } (TATA) = \frac{NI - ONCAF}{AT}$$

Note that there are both basic variables that are sub-dividable into further components (e.g., the Net Income (NI) used in TATA is constructed from most other variables), and other variables that are final atoms that can not be subdivided. For this reason, a naive strategy of attempting to minimize/maximize each individual ratio may not be effective because it would alter other variables in the hierarchy. Second, most of the ratios are relative year-over-year comparisons, and so greedily altering one year may impact a previous year's calculation. This is a critical consideration in the performance of earnings management because it may create the appearance of volatility in the company's performance, which is itself a negative signal used by analysts (Simko et al., 2020).

2.1.3. Secondary Fraud Score

The Earnings Per Share (EPS) and Benish's M-Score represent two anti-correlated targets to be minimized simultaneously. Though these are the primary targets of analysis, as EPS is a fundamental business metric and the M-Score is one of the most widely used fraud models, we will include a second fraud detection model as a third objective to be optimized in extended testing. This will allow us to demonstrate that our MVMO algorithm can succeed with more targets simultaneously.

In particular, we will use another well-used accounting fraud model proposed by Charalambos T. Spathis (Spathis, 2002), which we will term as the *S-Score* going forward. The S-Score model was selected as the coefficients are publicly available, and its relative contrast with the M-score is that it has only three covariates in the model specification and thus far has fewer accounting ratios compared to many alternative models. It's compact specification is given by $\mathrm{S} = 1.250 + 2.252 \cdot INVT/SALE - 33.029 \cdot NI/AT - 6.878 \cdot WC/AT$.

One could hypothesize our success against the M-Score is achieved only due to a large number of terms and thus complex interactions between them, as each accounting variable in Equation 3 is composed of the sum of many atoms and thus creates more of the non-linearities that are exploited in neural networks for evasive purposes. By showing that the S-Score also succumbs to our MVMO algorithms, we show a three-way joint optimization and that simpler fraud models are equally at risk.

3. Maximum Violated Multi-Objective Attack

Notably, all of the targets of our adversarial attack are real-valued functions. Most literature on AML is performed on classification tasks, with relatively little work done on regression problems (Nguyen and Raff, 2019; Gupta et al., 2021). In our case, the range of EPS and M-score calculations differ by up to $1000\times$, making the scale matching to optimize both together difficult. Also problematic is that the EPS and M-score both range from $-\infty$ to ∞, and in all cases, we wish to maximize EPS and minimize the M-score.

Our results will show that simply optimizing for the average of multiple least-squares objectives will not produce a satisfying outcome. Thus, our strategy will focus on the nature of the threat model: *all objectives must be satisfied for the result to satisfice*. If EPS increases (good) but the M-score also increases (bad), that means the fraud is more likely to be detected and so undesirable. Similarly, if M-score decreases but the EPS decreases, then there is a low desire to perpetrate the fraud due to negative outcomes. In our threat model, there is no amount by which EPS may increase that will make up for an increase in the M-score. For this reason we will focus on a gating strategy that will depress any objective that is satisfied in favor of other objectives that are unsatisfied.

Second, since we have objectives of infinite support, there is the real-world consideration that they may have different sensitivities. That is to say, it may be easier to change one objective by a large magnitude than another objective. To balance this, we will instead focus on the magnitude of the changes made by taking the transformation $g(\Delta) = \mathrm{sign}(\Delta)\log(|\Delta| + 1)$, where Δ is the change in the objective compared to its original value. This transformation will maintain the sign of the change that has occurred (hence the $+1$ so that the origin remains the same pre/post-transformation) but only in-

crease logarithmically. In this way, a more sensitive function will not gain undue favor in the numerical optimization.

This leads to our numeric optimization (shown for two variables for illustration purposes) in Equation 4, where we compute the transformation of EPS and M-score (M) as α and β, respectively. A larger EPS is better, so the minimization goal is $-\alpha$. A smaller M-score is better, so β can be optimized directly. This is converted to a final optimization score by taking the weighted average of $-\alpha$ and β via the softmax function.

$$\begin{aligned} \alpha &= g(EPS(X) - EPS(X + X \odot PM)) \\ \beta &= g(M(X) - M(X + X \odot PM)) \\ \arg\min_{P} [-\alpha, \beta]^\top \text{Softmax}([-\alpha, \beta] \cdot C) \quad & s.t. \quad |P| \leq \epsilon \end{aligned} \tag{4}$$

The softmax is computed so that more weight is placed on the metric that is currently performing worse by our goals. e.g., if $\alpha = 100$, and $\beta = 5$, we have already succeeded in making the EPS significantly higher, but have done so at increased risk of detection. So the Softmax($[-100, 5]$) $\approx [0, 1]$ will result in $-100 \cdot 0 + 5 \cdot 1 = 5$, so that the optimization effort is spent on reducing the M-score.

Algorithm 1 Maximum Violated Multi-Objective Attack

Require: Feature set x, with K victim functions $f_1(\cdot), \ldots, f_K(\cdot)$ to minimize. Maximum number of iterations T, and constraint projection $\pi(\cdot)$. Exaggeration constant C (initialized to 1 in our tests).

$x' \leftarrow x$

$d_0 \leftarrow [f_1(x'), f_2(x'), \ldots, f_K(x')]$ {Reference objectives needed to ensure we do not over-optimize one at the expense of others.}

for $t = 1$ **to** T **do**

$d_t \leftarrow [f_1(x'), f_2(x'), \ldots, f_K(x')]$ {d holds the directions that each value takes to make both a minimization goal.}

$\chi \leftarrow \text{sign}(d_t - d_0) \odot \log(|d_t - d_0| + 1)$ {Transform the targets to be insensitive to order-of-magnitude differences.}

$\mathcal{L} \leftarrow \chi^\top \text{Softmax}(\chi \cdot C)$ {Final optimization gibes the least-improved the most weight.}

$x' \leftarrow \pi(x' + \nabla_{x'} \mathcal{L})$

end for

return $\delta = x - x'$ {Adversarial perturbation found}

This forms the inspiration and method of our Maximum Violated Multi-Objective (MVMO), which focuses optimization on only the objectives that are currently unsatisfied relative to the others, as shown in algorithm 1. The projection function π keeps the solution constrained to the feasible attack space, and significantly improves upon naive PGD attacks. Note that in the case of perfectly correlated victim functions where $f_i(x) = f_j(x)$, our method will reduce to standard PGD as the softmax will compute the simple average in such cases. Our approach is simple to implement, and as we will show, significantly more effective in practice.

3.0.1. Attacker Constraint

We have now enumerated the goals of the attacker. The constraints are simply to specify a limit on ϵ, which is interpretable as the maximum relative change allowed to any atom in the original 10-K report. A relative change in a variable by over 70% occurs in real-life fraud (Ramírez-Orellana et al., 2017). For this reason, we will consider a maximum limit of $\epsilon \leq 40\%$ so that our results are well within the scope of perturbations committed in real-world situations.

Note that some atom/leaf variables in our model, like DLTT, actually have more children in the complete specification of the financial variables. By treating DLLT and others as an atom instead, *we effectively reduce the total strength of a true theoretical attack.* Consider that DLTT may be 0, when in actuality it has two large values Z and $-Z$ that contribute to its calculation. In this case, the net score is 0, and any percent change will be zero, so no valid perturbation is possible in our threat model. In actuality, one of the two sub-components could have been modified, thus increasing the total set of valid attacks. Our simplification is done as a matter of exposition (we do not wish to teach the reader all of the accounting) and practicality (the sub-accounts are not always used by all companies, making the code more complex).

We note that it may seem uncessary for an iterative gradient attack due to the linear nature of accounting variables (addition, subtraction, multiplication, and division). In fact, to keep the attack feasible, it is key to constraint the year-to-year perturbatinos to be consistent, which makes the problem non-convex.

Theorem 1 *Given a function* $f(x) = \sum_{t=1}^{T} \alpha_t \frac{\beta_{t-1}}{\beta_t}$ *where* β_t *is the* t*'th year's accounting variable, making* β_{t-1}/β_t *the accounting ratio, and* α_t *the weight of the variable. The objective* $f(x)$ *is non-context over the sum of years* T*.*

Proof The diagonal of the hessian is $\frac{\partial^2}{\partial \beta_i^2}\left(\sum_{t=1}^{T} \alpha_t \frac{\beta_{t-1}}{\beta_t}\right) = \frac{2\beta_{i-1}\alpha_{t-1}}{\beta_i^3}$, for which $\frac{2\beta_{i-1}\alpha_{t-1}}{\beta_i^3} < 0$ whenever $\beta_i < 0$. We thus fail to satisfy the necessary condition that $f''(x) > 0 \forall x$ for $f(x)$ to be a convex function. ■

3.1. Perturbation Strategies

We will detail the strategies that we encode into the threat model below. In each case, we have two corresponding variables that are added to the matrix M, with each row representing one pair of variables mentioned below. For example, if the relationship of the i'th variable is an anti-correlated impact on the j'th variable, the relationship will be encoded as $M[r, i] = 1$ and $M[r, j] = -1$, where r is an arbitrary row that is holding the current relationship. Thus when the matrix product is taken with P, the corresponding i, j variables will be altered in a one-to-one relationship.

By the nature of how P and M are represented, the strategies used in this section are single-year strategies, actions that can be taken in one year to influence the final reported 10-K, and thus the ratios that will be used to evaluate the company. This is a realistic threat model in that a bad actor could apply these techniques to optimize for the current year's 10-K. While more complex multi-year strategies are possible, they become more dependent

on the degree of long-term financial planning conducted at each firm, which can be industry and Chief Financial Officer (CFO) specific.

To ensure that our strategies are plausible to a real-world threat model, they were constructed with guidance and collaboration with domain experts: including processional certified public accountants (CPAs) and lawyers. It is critical that we emphasize that these are not legal for someone to perform, but reflect strategies observed in real-world financial fraud.

3.1.1. Cost of Goods Sold (COGS) based Strategies

COGS is, in general, expenses incurred that are directly traceable back to the production of a product (e.g., physical materials the product is made from). It, along with Selling, General, & Administrative (XSGA) expenses can have significant overlap in the nature of their accounting: both at a high level are tracking expenses related to the production of revenue, the difference is where the expense is traceable to. However, this similarity means an expense can be easily moved between the two accounts. Both COGS and XSGA are used in the Beneish-M score, impacting different variables, and so incentivize allocation of the expenses to the minimally penalized variable. For the same logical reasons, a third account for Staff Expenses (XSTF) can also be used to reallocate funds, while the XSTF is less frequently used in standard financial ratios. Thus, for a $1 increase in COGS, the adversary can alter XSGA or XSTF by -$1 in their model (and in the reverse as well).

A second variable that is not obviously connected to COGS is the Inventory (INVT). This is because, on a firm's balance sheet, INVT and COGS are expected to balance via a third factor as: $COGS = \text{Beginning Inventory} + \text{Purchases} - \text{Ending Inventory}$.

Where the inventory sold during the year produces the revenue (and corresponding tracked expense) that goes into the calculation of COGS. However, tracking inventory is notoriously error-prone. This thus makes it a convenient place for an adversary to intentionally "poorly track" their inventory to inflate/deflate it as desired, and thus alter the COGS. Thus, a $1 increase in COGS can also be achieved by a -$1 change to INVT, and vice versa.

3.1.2. Debt Term Alteration

Debt is categorized into Current Liabilities (LCT) that are due within the year, and Lonterm debt (DLTT) that is due in more than a year's time. DLTT becomes LCT as the debt nears its due date. For a bad actor, so long as they pay all debts due, they can missreport DLTT as LCT by claiming in the following year that the LCT was paid off, but new DLTTs were acquired after the year. Similarly, LCT can be reported as DLTTs by that are explained away the following year as being paid off early. In each case, the debtors are paid on time, but the accounting is done fraudulently. Thus DLTT and LCT have a $1/-$1 invertible relationship.

3.1.3. Phantom Sales

A firm's revenue (SALE) when reported on a 10-K does not reflect *just* literal cash received in the fiscal year. Obligations from other parties to pay the firm in the near future, but that have not yet been actually paid, also contribute to SALE. This is accounted for with the

Accounts Receivable (RECV), and so fraudulent SALE volume can be achieved by adding a corresponding amount of RECV. The strategy thus has a \$1/\$1 relationship in M. This could be explained away the following year by claiming that the purchaser fell through and failed to make payment.

3.1.4. Hiding Property Expenses

The previous strategies mentioned are all contained within the balance sheet of a 10-K, which is a reflection of the company's net value from inception, accumulated to the current fiscal year. To alter the income sheet of a 10-K, we can use the property-related reports. Equipment and Occupancy Expense (XEQO, on the income statement) that accounts for maintenance of Property, Plant, and Equipment (PPEGT, on the balance sheet) provides a direct connection. Altering XEQO by -\$1 (and thus, increasing net income) by claiming it instead as the purchase of new PPEGT (and thus increasing it by \$1) creates a double-dipping effect on the total finances.

3.1.5. Depreciation Tax Shield

Depreciation (DP) in accounting is the accumulation of "negative" value to an asset already purchased. This is valuable financially because DP has a "tax shield" effect, where DP is removed from the total income for tax calculation purposes, even though DP does not correspond with a cash transfer in the current year. DP calculations are highly subjective, alterable, and can be challenging to track. A company can fraudulently claim accelerated or non-existent DP to obtain an undue tax credit. This is encoded as just a single $M[r, DP] = 1$ in the perturbation strategy matrix.

4. Results

To produce our results, we implement our model in the JAX (Bradbury et al., 2018) framework. JAXOPT (Blondel et al., 2021) is used to perform the adversarial search using projected gradient descent (PGD) (Madry et al., 2018) in order to respect the ϵ limits. We will evaluate values of $\epsilon \in \{5\%, 10\%, 20\%, 40\%\}$. Each variable is encoded with respect to Figure 2 such that any perturbation to one atom/leaf node automatically propagates to all other calculations. This allows implementation via automatic differentiation through the financial ratios. Our data is sourced from (BAO et al., 2020), which contains 402 companies that have confirmed years of actual financial fraud occurring. There are a cumulative 979 years of 10-Ks in the data to be altered. Each company has an average of 7.9 years of available data. We obtain the original 10-K data in a standard format via the Wharton Research Data Services (Services) database. This gives us ideal data to evaluate if our attacks would reduce the M-score commonly used for the identification of companies that are at risk of earnings management.

In each experiment, we will examine the percentage of years that an attacker could apply the MVO (or alternative) algorithm and satisfy the attacker's goal under the threat model: reducing M-Score and increasing EPS simultaneously. Because sign changes occur, we use the Relative Percent Difference $RPD(a, b) = \frac{a-b}{(|a|+|b|)/2}$, to compute the relative change in

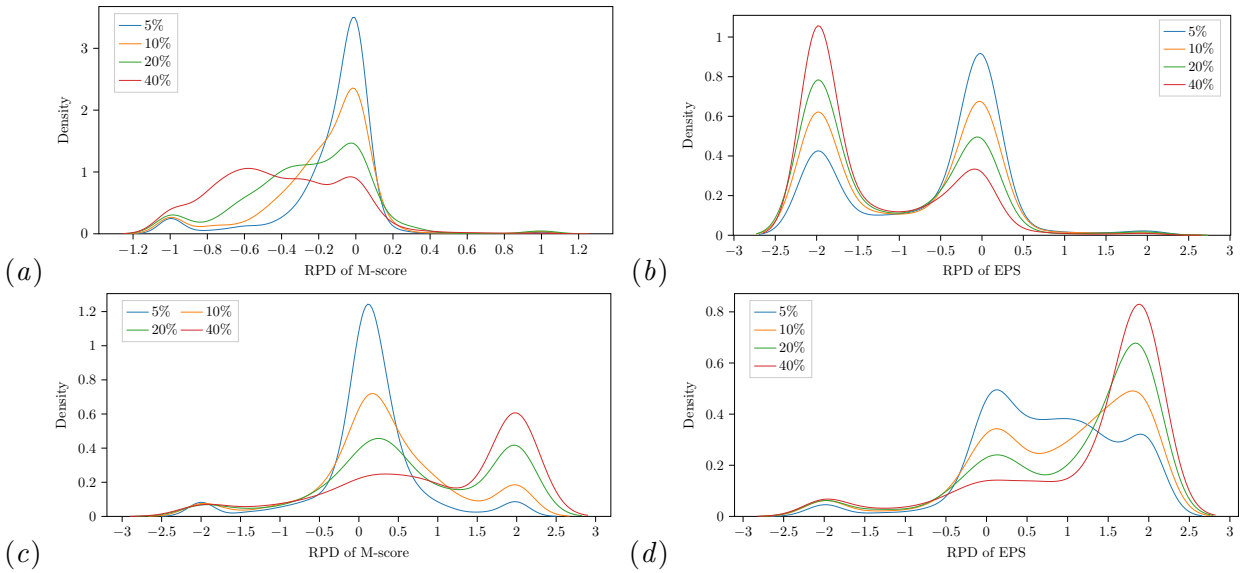

Figure 3: Kernel Density plot results when using the simple strategies of attacking just the M-score (top row) or just EPS (bottom row) for differing values of ϵ. The x-axis in each case is the RPD ($1 = 100\%$) showing the effect on the dataset. The EPS goes down when targeting M-score (top-right), which defeats the purpose of the adversary committing the fraud, making it non-viable. Attacking EPS is easily detected by raising the M-score (bottom-left).

EPS and M-score, which would be problematic for the standard relative change calculation of $(a - b)/a$.

As baselines, we will consider the SA-MOO multi-object approach of Williams and Li (2023), which we find fails due to sensitivity to order-of-magnitude change in the objectives and over-focusses on the easier metric of EPS because it allows driving the cumulative reward up by orders of magnitude. Second, we consider the Manifold ELBO of Yang et al. (2024), which we find fails due to sampling noise in its stochastic estimation of the ELBO. Finally, we test three forms of PGD: attacking the M-score, EPS, and the average of M-score and EPS. In each case, PGD alone is insufficient.

The simple attack baseline strategy from section 5 that numerically evaluates one of either M-score or EPS are shown in Figure 3 (i.e., target only one of the two outcomes). As can be seen, the two strategies are naturally correlated. Decreasing the M-score to evade detection also decreases the EPS. Since the purpose of the fraud is to falsely bolster the financial performance, the attack becomes undesirable. Similarly, attacking the EPS also raises the M-score, increasing the risk of fraud being detected. Thus, the naive strategy of attacking one single metric is inadvisable.

We also tested a number of simple ablations for merging M-score and EPS targets. This includes taking a weighted average, testing increments of 10% for the weighting, and taking the difference in the mean decrease in scores (measured over all the years of a company's

Table 2: Based on the attack method and ϵ (1st & 2nd column), the percent that satisfy both (3rd) a lower M-score and higher mean EPS. The MVMO method success rate goes down slightly as ϵ increases, but the size of the average attack's influence increases (4th and 5th columns). Only the MVMO method of Equation 4 satisfies a large number of cases.

Attack Method	ϵ	Satisfy Goals (%)	Avg. M-score RPD (%)	Avg. EPS RPD(%)
SA-MOO	5%	6.30	13.71	53.78
SA-MOO	10%	5.72	35.40	63.95
SA-MOO	20%	5.17	72.89	76.29
SA-MOO	40%	4.15	104.86	86.77
Manifold ELBO	5%	1.61	2.47	3.19
Manifold ELBO	10%	1.96	0.66	6.61
Manifold ELBO	20%	0.41	9.61	12.59
Manifold ELBO	40%	0.77	4.20	12.67
PGD-M	5%	4.60	-12.04	-60.31
PGD-M	10%	3.88	-18.04	-88.91
PGD-M	20%	2.66	-26.43	-110.42
PGD-M	40%	1.84	-40.6	-135.29
PGD-EPS	5%	13.28	12.17	74.69
PGD-EPS	10%	12.05	31.43	88.76
PGD-EPS	20%	10.93	64.72	105.93
PGD-EPS	40%	8.89	93.11	120.49
PGD-Avg	5%	2.66	-5.42	-42.93
PGD-Avg	10%	3.47	-20.92	-49.26
PGD-Avg	20%	3.88	-43.95	-58.72
PGD-Avg	40%	2.25	-55.83	-62.5
MVMO	5%	**49.13**	-8.83	32.79
MVMO	10%	**61.18**	-11.52	43.35
MVMO	20%	**65.99**	-14.01	51.53
MVMO	40%	**63.84**	-15.54	57.62

data). In all cases, we observed one of the two scenarios outlined in Figure 3: M-score and EPS both decreasing or both increasing.

Our novel MVMO attack strategy that works to equalize the order-of-magnitude change in both goals performed significantly better. The results for all methods are shown in Table 2, where the third column indicates the percentage of years that satisfy the goal of decreasing RPD while also increasing EPS. The MVMO optimization we propose in Equation 4 is the only strategy that has any meaningful amount of success. PGD applied to just M-score (PGD-M) and EPS (PGD-EPS) each only achieve their target goal, at the cost of the other. Attempting to average the losses (PGD-Avg) performs even worse than naively ignoring a goal. Similarly, normalizing gradients, rescaling, and clipping did not

improve performance. Our MVMO is the only mechanism able to satisfy the two anti-correlated targets simultaneously.

Notably, the *average* relative changes can be large, by up to an absolute 135%. The EPS changes are large in magnitude in every case, likely because the EPS is one ratio where the M-score is a regression of multiple ratios, and so is more robust. Our MVMO attack does a better job of satisfying both goals by reducing the gap in their absolute magnitudes, which is the intended effect. A key takeaway from these results is that *significant EPS manipulation is possible with no detectable impact on fraudulent activity.*

That we can reach a 66% success is also notable in the much higher success rates often seen in the deep-learning literature (Biggio and Roli, 2018). This result is more congruent with older adversarial machine learning literature that studied simpler linear and kernel methods with higher intrinsic robustness (Biggio et al., 2012).

A key baseline in Table 2 is PGD-Avg, which takes the average of the EPS and M-Score objective. This had a lower satisfying rate than naively optimizing just the EPS, at 4% for PGD-Avg compared to 13% for PGD-EPS. This demonstrates how the scale and individual target sensitivity can lead to unintuitive outcomes when not factored into the attack strategy accordingly.

The "price" to pay for MVMO's success in this context is a reduction in the average RPD of the target metrics. Though this tradeoff is to be expected, the relative cost of the tradeoff is inconsequential to the attacker's perspective. Any increase in EPS at no increased risk is a win for the attacker, and the EPS change is only a factor of two smaller compared to the naive PGD-EPS optimizer. Similarly, the M-Score RPD reduction is approximately 2.5× smaller for each budget of ϵ.

One might also ask about the distribution of outcomes beyond a binary success/failure criterion. For example, if there are many M-Score RPDs that are positive but essentially zero, the risk of attacker success in practice may be greater still. This hypothesis is confirmed by a Kernel Density Estimate (KDE) plot of both objectives in Figure 4. A large density near zero confirms the hypothesis and indicates that a large population of companies is challenging to optimize both EPS and M-score concurrently; however, many EPS increases can be obtained with minor M-score increases. A key takeaway from this initial result is that *financial ratios in use today are reasonably effective against adversarial attacks.* This result is largely explained by the theoretical work of (Ribeiro and Schön, 2023), who showed that the risk of linear regression models is primarily a function of the number of features in the input. Because most financial models use a relatively small number of ratios, the total size of the attack space is constrained. 66% is a much lower attack success rate compared to machine learning literature at large. This is in many ways a positive indicator of the tools that have evolved over time in accounting, which makes sense given the adversarial environment that had already formed between duplicitous firms and regulators/the market. The results are far from allowing accounting research or defensive AML research to ignore this application area though.

At the same time, this amplifies the fact that the high-average impact on EPS is obtained from a second population that routinely obtains up to 200% increases in EPS. This is a significant effect size that would have a material impact on all agents involved. This poses a non-trivial risk to investors, regulators, and the population at large whose investments may be held with these firms. The magnitude of the attack successes may inform regulators of

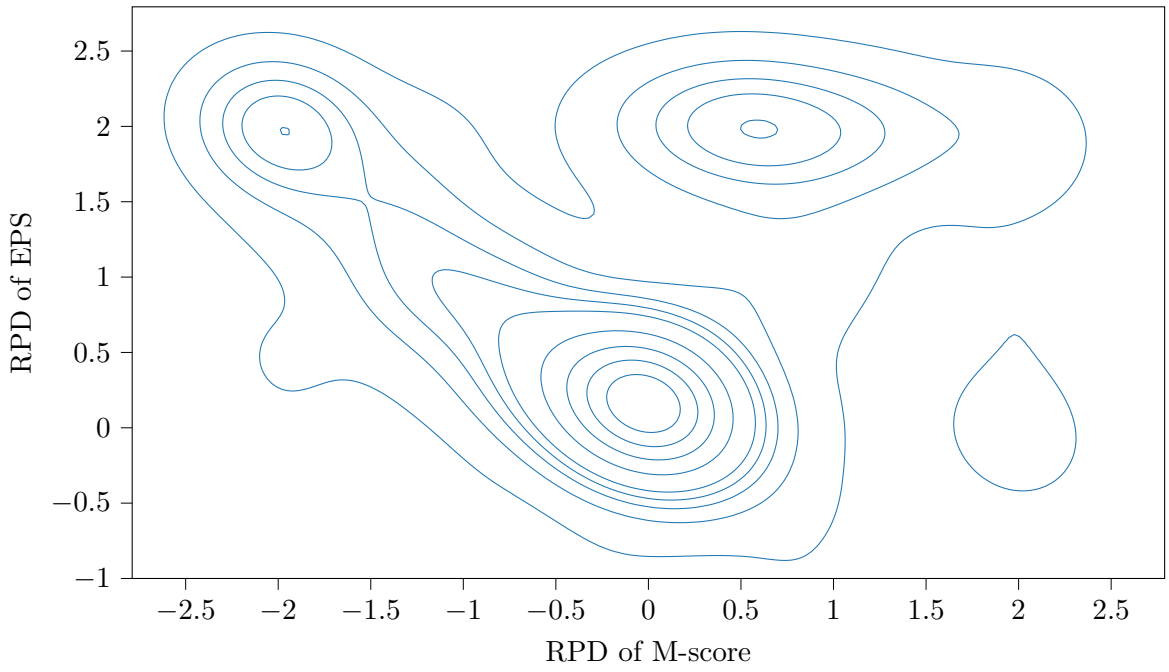

Figure 4: Two-dimensional Kernel Density Estimate (KDE) of the RPD scores for the M-score (x-axis, lower is better) and the EPS (y-axis, higher is better). 49% of the density exists in the satisficing zone of the top-left quadrant, which are successful attacks satisfying both objectives. The KDE shows that the highest density region is at the center with no relative change to either variable, indicating a large number of points that are difficult to optimize.

the scope of risk, and the degree of attack success against different firms may help inform where to most judiciously apply their regulator actions. The results in simulation do not account for additiona actions a bad actor may take beyond simple 10-K reporting fraud if they had the tools available to perform these more advanced manipulation strategies, and could further re-structure their corporate operations to make AML more amenable to future year's reporting.

4.1. Ablating Additional Objective

To evaluate our model when a second fraud model is in consideration, we use the S-Score (Spathis, 2002). EPS and M-Score are anti-correlated objectives, while M-Score and S-Score should be correlated as they are both fraud detection models. This answers two questions: 1) How well does an evasive 10-K report generalize to fraud models not previously seen? 2) Can MVMO successfully optimize three simultaneous objectives?

The results of this experiment are shown in Table 3, where MVMO-M is the same model from Table 2 that optimized just two objectives: EPS and M-Score; MVMO-MS optimizes both M and S-scores. MVMI-MS has almost equivalent satisfaction rate against

Table 3: Results of algorithm 1 when optimizing EPS and M-Score (MVMO-M), and optimizing an additional third objective for the M-score (MVMO-MS). Columns show the percentage of times that the respective objective was satisfied. The S-Score is correlated with M-Score as another fraud model.

Algo	ϵ	EPS%	M-score%	S-Score%	3-way Joint%
MVMO-M	5%	76.71	70.38	53.22	27.17
MVMO-M	10%	80.39	78.96	54.34	34.22
MVMO-M	20%	84.27	80.39	56.38	37.18
MVMO-M	40%	84.88	77.53	55.67	34.01
MVMO-MS	5%	77.63	68.64	87.64	**42.59**
MVMO-MS	10%	80.18	78.24	89.58	**54.14**
MVMO-MS	20%	84.27	81.31	92.34	**62.31**
MVMO-MS	40%	85.70	78.86	93.46	**62.72**

three targets as MVMO did to just two, showing more objectives can be satisfied. As can be seen in the final "3-way Joint" column, MVMO-M has a non-trivial satisfaction rate of 34% against all three desiderata, even though it was never previously exposed to the S-Score. This result alone is encouraging from the attacker's perspective in that there is evidence they can evade detection from models that they have not previously considered. We remind the reader that the S-Score is a fraud detection model, so even a 50% satisfaction rate against just the S-Score would be better than random guessing as the S-Score is supposed to detect fraud, which is actively being performed.

5. Ablated Attack Objectives

We note in this section prior attempts at implementing an attack, which necessitated our eventual design. First, we note the obvious strategy is to perform mean squared error (MSE) loss on the $-$EPS (i.e., maximize) and M-score (i.e., minimize) objectives. When restricted to just the M-score alone we found MSE based optimization ineffective due to over-optimizing just one year's M-score. This was exacerbated when attempting to add EPS, due to its different scale, and resulted in zero years where both EPS went up and M-score went down. This was repeated with the L_1 loss and the same issue was observed.

Our second attempt to implement the attack was informed by noticing that the M-score for one particular year may become the target of optimization because it is so much lower than all other values, causing the optimization process to over-fit the year. This occurs because of the use of ratios in the M-score calculation, and if one year has a particulary small numerator/denominator in one year, it may become easier to leverage a unit change in the input for an out-sized impact on the average M-score across years. Thus, we used a maximum-violator policy and took only the maximum M-score over all T years in each call to the objective function so that the worst violating year was optimized down. This requires proportionally T more PGD iterations to influence all potential years, however, we

found it the second-best strategy in our testing. Our baseline results in section 4 all use this strategy for M-score optimization, and use standard MSE-based loss for EPS optimization.

Other similar strategies for the M-score were tested as well. This includes clipping M-scores to a minimum value that was iteratively lowered as a function of the larger M-score, using a power function to squash the larger magnitude M-scores into smaller ranges, and top-k M-score selection. All resulted in comparatively equivalent performance in practice, and so we keep the max selection as our initial baseline for ease of implementation.

6. Conclusion

In this work, we have demonstrated the applicability of AML to financial fraud modeling. Using a novel multi-goal optimization strategy that is designed to handle anti-correlated objectives, we are able to demonstrate significant $\approx 50\%$ of years a company can inflate earnings by as much as 200% while having an average 15% decrease in fraud risk scores.

In consideration of publishing this work, we note that our approach has no impact on the ability to evade financial auditing, a regular requirement of large companies. While auditing is not flawless, and there are concerns about reduced auditing effectiveness as the market concentrates on the "big four" accounting firms[1], it remains a secondary check on malicious behavior. Financial ratios are also not the only source by which fraud is detected. The role of the press is non-trivial in detecting financial fraud (Miller, 2006). Executives are also expected to engage with investors in annual meetings, which can be revealing of fraud risk (Larcker and Zakolyukina, 2012), and non-participation would pose an even greater red-flag signal. Given these other checks on fraudulent behavior and the skills gap for CFOs and accountants to perform our attack, we believe the interest in disclosing the capability to encourage study and remediation outweighs the risk of it being used without eventual detection.

In making this discovery public, it is worth noting some of the ways in which we hope our results will lead to positive outcomes. First, we note that this interdisciplinary study required collaboration across multiple disciplines to achieve its goals. Encouraging this collaboration, particularly between government, industry, and academia, to study this problem is an immediate goal. In particular, the use of linear models increases the possibility of developing provable bounds on attack risk and success rates, which can lead to better fiscal oversight and the allocation of regulatory resources. In particular, if regulators can focus their time on cases with the greatest risk of successful fraud, then all parties can ultimately benefit.

References

Cornell Research Report On Enron 1998 | PDF | Enron | Discounted Cash Flow, 1998. URL https://www.scribd.com/doc/66581069/Cornell-Research-Report-on-Enron-1998.

Edward I. Altman. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968.

1. https://www.ifiar.org/?wpdmdl=15294

ISSN 1540-6261. doi: 10.1111/j.1540-6261.1968.tb00843.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1968.tb00843.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1968.tb00843.x.

Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 339–364, 2023. doi: 10.1109/SaTML54575.2023.00031.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *International Conference on Machine Learning (ICML)*, 2018. URL http://arxiv.org/abs/1802.00420. arXiv: 1802.00420.

YANG BAO, BIN KE, BIN LI, Y. JULIA YU, and JIE ZHANG. Detecting accounting fraud in publicly traded u.s. firms using a machine learning approach. *Journal of Accounting Research*, 58(1):199–235, 2020. doi: https://doi.org/10.1111/1475-679X.12292. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12292.

Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, and Jaehoon Amir Safavi. Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 103–110, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5202-4. doi: 10.1145/3128572.3140450. URL http://doi.acm.org/10.1145/3128572.3140450. Series Title: AISec '17.

Messod D. Beneish. Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy*, 16(3):271–309, September 1997. ISSN 0278-4254. doi: 10.1016/S0278-4254(97)00023-9. URL https://www.sciencedirect.com/science/article/pii/S0278425497000239.

Messod D. Beneish. The Detection of Earnings Manipulation. *Financial Analysts Journal*, 55(5):24–36, September 1999. ISSN 0015-198X. doi: 10.2469/faj.v55.n5.2296. URL https://doi.org/10.2469/faj.v55.n5.2296. Publisher: Routledge _eprint: https://doi.org/10.2469/faj.v55.n5.2296.

Messod D. Beneish and Craig Nichols. The Predictable Cost of Earnings Manipulation, August 2007. URL https://papers.ssrn.com/abstract=1006840.

Messod D. Beneish and Craig Nichols. Identifying Overvalued Equity, June 2009. URL https://papers.ssrn.com/abstract=1134818.

Messod D. Beneish, Charles M. C. Lee, and D. Craig Nichols. Fraud Detection and Expected Returns, February 2012. URL https://papers.ssrn.com/abstract=1998387.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, December 2018. ISSN

00313203. doi: 10.1016/j.patcog.2018.07.023. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320318302565. arXiv: 1712.03141.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning Attacks Against Support Vector Machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1467–1474, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1. URL http://dl.acm.org/citation.cfm?id=3042573.3042761. Series Title: ICML'12.

Battista Biggio, Giorgio Fumera, and Fabio Roli. Security evaluation of pattern classifiers under attack. *IEEE Transactions on Knowledge and Data Engineering*, 26(4):984–996, 2014. ISSN 10414347. doi: 10.1109/TKDE.2013.57.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, and Nicholas Carlini. Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent, June 2021. URL http://arxiv.org/abs/2106.15023. arXiv:2106.15023 [cs].

Anh Tuan Bui, Trung Le, He Zhao, Quan Hung Tran, Paul Montague, and Dinh Phung. Generating adversarial examples with task oriented multi-objective optimization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2f81Q622ww.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, May 2017. ISBN 978-1-5090-5533-3. doi: 10.1109/SP.2017.49. URL http://ieeexplore.ieee.org/document/7958570/.

Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets. 2018. URL http://arxiv.org/abs/1802.08232. arXiv: 1802.08232.

Nicholas Carlini, Ariel Herbert-voss, Dawn Song, Florian Tramèr, Katherine Lee, Eric Wallace, Adam Roberts, Alina Oprea, Matthew Jagielski, Tom Brown, Colin Raffel, and Peter W. Extracting Training Data from Large Language Models. In *USENIX Security*, 2021. URL https://arxiv.org/abs/2012.07805. arXiv: 2012.07805v1.

Antonio Emanuele Cinà, Francesco Villani, Maura Pintor, Lea Schönherr, Battista Biggio, and Marcello Pelillo. σ-zero: Gradient-based Optimization of ℓ_0-norm Adversarial Examples, February 2024. URL http://arxiv.org/abs/2402.01879. arXiv:2402.01879 [cs].

Nilanjana Das, Edward Raff, Aman Chadha, and Manas Gaur. Human-readable adversarial prompts: An investigation into llm vulnerabilities using situational context, 2025. URL https://arxiv.org/abs/2412.16359.

Patricia M. Dechow and Ilia D. Dichev. The Quality of Accruals and Earnings: The Role of Accrual Estimation Errors. *The Accounting Review*, 77(s-1):35–59, March 2002. ISSN 0001-4826. doi: 10.2308/accr.2002.77.s-1.35. URL https://doi.org/10.2308/accr.2002.77.s-1.35.

Patricia M. Dechow, Richard G. Sloan, and Amy P. Sweeney. Detecting Earnings Management. *The Accounting Review*, 70(2):193–225, 1995. ISSN 0001-4826. URL https://www.jstor.org/stable/248303. Publisher: American Accounting Association.

Mark L. DeFond and James Jiambalvo. Debt covenant violation and manipulation of accruals. *Journal of Accounting and Economics*, 17(1):145–176, January 1994. ISSN 0165-4101. doi: 10.1016/0165-4101(94)90008-6. URL https://www.sciencedirect.com/science/article/pii/0165410194900086.

Philip Doldo, Derek Everett, Amol Khanna, Andre T Nguyen, and Edward Raff. Stop walking in circles! bailing out early in projected gradient descent. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 6373–6382, June 2025.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

Giuseppe Floris, Raffaele Mura, Luca Scionis, Giorgio Piras, Maura Pintor, Ambra Demontis, and Battista Biggio. Improving Fast Minimum-Norm Attacks with Hyperparameter Optimization. In *ESANN 2023 proceesdings*, pages 127–132, 2023. doi: 10.14428/esann/2023.ES2023-164. URL http://arxiv.org/abs/2310.08177. arXiv:2310.08177 [cs].

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, September 2006. ISSN 1936-0975, 1931-6690. doi: 10.1214/06-BA117A. URL https://projecteuclid.org/journals/bayesian-analysis/volume-1/issue-3/Prior-distributions-for-variance-parameters-in-hierarchical-models-comment-on/10.1214/06-BA117A.full. Publisher: International Society for Bayesian Analysis.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian Data Analysis Third edition (with errors fixed as of 13 February 2020). (February):677, 2013. ISBN: 978-1439840955.

Kavya Gupta, Beatrice Pesquet-Popescu, Fateh Kaakai, Jean-Christophe Pesquet, Fragkiskos D Malliaros, and Universite Paris-Saclay. An adversarial attacker for neural networks in regression problems. *Proceedings of the Workshop on Artificial Intelligence Safety 2021 co-located with the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*, 2021.

Paul M. Healy. The effect of bonus schemes on accounting decisions. *Journal of Accounting and Economics*, 7(1):85–107, April 1985. ISSN 0165-4101. doi: 10.1016/0165-4101(85)90029-1. URL https://www.sciencedirect.com/science/article/pii/0165410185900291.

Paul Hribar and Daniel W. Collins. Errors in Estimating Accruals: Implications for Empirical Research. *Journal of Accounting Research*, 40(1):105–134, 2002. ISSN 1475-679X. doi: 10.1111/1475-679X.00041. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.00041. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.00041.

Henrik Höglund. Detecting earnings management with neural networks. *Expert Systems with Applications*, 39(10):9564–9570, August 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2012.02.096. URL https://www.sciencedirect.com/science/article/pii/S0957417412003594.

Jennifer J. Jones. Earnings Management During Import Relief Investigations. *Journal of Accounting Research*, 29(2):193–228, 1991. ISSN 0021-8456. doi: 10.2307/2491047. URL https://www.jstor.org/stable/2491047. Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].

Amol Khanna, Fred Lu, Edward Raff, and Brian Testa. Differentially private logistic regression with sparse solutions. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 1–9, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702600. doi: 10.1145/3605764.3623910. URL https://doi.org/10.1145/3605764.3623910.

Amol Khanna, Edward Raff, and Nathan Inkawhich. Sok: A review of differentially private linear models for high-dimensional data. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 57–77, April 2024. doi: 10.1109/SaTML59370.2024.00012. URL https://arxiv.org/abs/2404.01141.

Amol Khanna, Fred Lu, and Edward Raff. Differentially private iterative screening rules for linear regression. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, CODASPY '25, page 72–83, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714764. doi: 10.1145/3714393.3726507. URL https://doi.org/10.1145/3714393.3726507.

Ashley Klein, Edward Raff, Elisabeth Seamon, Lily Foley, and Timothy Bussert. More Options for Prelabor Rupture of Membranes, A Bayesian Analysis. In *11th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2024*, 2024. URL https://www.arxiv.org/abs/2408.10876. Best Paper Award.

Xiangyin Kong and Zhiqiang Ge. Adversarial attacks on regression systems via gradient optimization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(12):7827–7839, 2023. doi: 10.1109/TSMC.2023.3302838.

David F. Larcker and Anastasia A. Zakolyukina. Detecting Deceptive Discussions in Conference Calls. *Journal of Accounting Research*, 50(2):495–540,

2012. ISSN 1475-679X. doi: 10.1111/j.1475-679X.2012.00450.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-679X.2012.00450.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-679X.2012.00450.x.

Chang Liu, Bo Li, Yevgeniy Vorobeychik, and Alina Oprea. Robust Linear Regression Against Training Data Poisoning. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 91–102, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5202-4. doi: 10.1145/3128572.3140447. URL http://doi.acm.org/10.1145/3128572.3140447. Series Title: AISec '17.

Fred Lu, Francis Ferraro, and Edward Raff. Continuously Generalized Ordinal Regression for Linear and Deep Models. In *SIAM International Conference on Data Mining (SDM22)*, 2022. URL http://arxiv.org/abs/2202.07005. arXiv: 2202.07005.

Wanting Lu and Xiaokang Zhao. Research and improvement of fraud identification model of Chinese A-share listed companies based on M-score. *Journal of Financial Crime*, 28(2): 566–579, January 2020. ISSN 1359-0790. doi: 10.1108/JFC-12-2019-0164. URL https://doi.org/10.1108/JFC-12-2019-0164. Publisher: Emerald Publishing Limited.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb. arXiv: 1802.10217.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 61065–61105. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/70702e8cbb4890b4a467b984ae59828a-Paper-Conference.pdf.

Gregory S. Miller. The Press as a Watchdog for Accounting Fraud. *Journal of Accounting Research*, 44(5):1001–1033, 2006. ISSN 1475-679X. doi: 10.1111/j.1475-679X.2006.00224.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-679X.2006.00224.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-679X.2006.00224.x.

Niluh Putu Dian Rosalina Handayani Narsa, Lesta Mega Evi Afifa, and Oktaviani Ari Wardhaningrum. Fraud triangle and earnings management based on the modified M-score: A study on manufacturing company in Indonesia. *Heliyon*, 9(2):e13649, February 2023. ISSN 2405-8440. doi: 10.1016/j.heliyon.2023.e13649. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9970902/.

Andrew Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. *Twenty-first international conference on Machine learning - ICML '04*, page 78, 2004. doi: 10.1145/1015330.1015435. URL http://portal.acm.org/citation.cfm?doid=1015330.1015435. Publisher: ACM Press Place: New York, New York, USA ISBN: 1581138285.

Andre T. Nguyen and Edward Raff. Adversarial attacks, regression, and numerical stability regularization. In *The AAAI 2019 Workshop on Engineering Dependable and Secure Machine Learning Systems*. arXiv, 2019. doi: 10.48550/arXiv.1812.02885. URL http://arxiv.org/abs/1812.02885. arXiv:1812.02885 [cs, stat].

Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. Intriguing Properties of Adversarial ML Attacks in the Problem Space. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1332–1349. IEEE, May 2020. ISBN 978-1-72813-497-0. doi: 10.1109/SP40000.2020.00073. URL https://ieeexplore.ieee.org/document/9152781/.

Joseph D. Piotroski. Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers. *Journal of Accounting Research*, 38:1–41, 2000. ISSN 0021-8456. doi: 10.2307/2672906. URL https://www.jstor.org/stable/2672906. Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5231–5240. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/qin19a.html.

Edward Raff, Michel Benaroch, and Andrew L. Farris. You don't need robust machine learning to manage adversarial attack risks, 2023a. URL https://arxiv.org/abs/2306.09951.

Edward Raff, Amol Khanna, and Fred Lu. Scaling up differentially private lasso regularized logistic regression via faster frank-wolfe iterations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36349–36363. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/72235260ae8d57ac42638a26d3b7d089-Paper-Conference.pdf.

Arash Rahnama, Andre T. Nguyen, and Edward Raff. Robust Design of Deep Neural Networks against Adversarial Attacks based on Lyapunov Theory. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8178–8187, 2020. URL http://arxiv.org/abs/1911.04636. arXiv: 1911.04636.

Alicia Ramírez-Orellana, María J. Martínez-Romero, and Teresa Mariño-Garrido. Measuring fraud and earnings management by a case of study: Evidence from an international family business. *European Journal of Family Business*, 7(1):41–53, Jan 2017. ISSN 2444-877X. doi: 10.1016/j.ejfb.2017.10.001.

Antônio H. Ribeiro and Thomas B. Schön. Overparameterized linear regression under adversarial attacks. *IEEE Transactions on Signal Processing*, 71:601–614, 2023. doi: 10.1109/TSP.2023.3246228.

Rebecca L. Rosner. Earnings Manipulation in Failing Firms. *Contemporary Accounting Research*, 20(2):361–408, 2003. ISSN 1911-3846. doi: 10.1506/8EVN-9KRB-3AE4-EE81. URL https://onlinelibrary.wiley.com/doi/abs/10.1506/8EVN-9KRB-3AE4-EE81. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1506/8EVN-9KRB-3AE4-EE81.

Donald B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4):1151–1172, December 1984. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346785. URL https://projecteuclid.org/journals/annals-of-statistics/volume-12/issue-4/Bayesianly-Justifiable-and-Relevant-Frequency-Calculations-for-the-Applied-Statistician/10.1214/aos/1176346785.full. Publisher: Institute of Mathematical Statistics.

Zakeya Sanad. Machine Learning and Earnings Management Detection. In Abdalmuttaleb M. A. Musleh Al-Sartawi, editor, *The Big Data-Driven Digital Economy: Artificial and Computational Intelligence*, Studies in Computational Intelligence, pages 77–83. Springer International Publishing, Cham, 2021. ISBN 978-3-030-73057-4. doi: 10.1007/978-3-030-73057-4_6. URL https://doi.org/10.1007/978-3-030-73057-4_6.

Wharton Research Data Services. Wharton research data services. URL https://wrds-www.wharton.upenn.edu/pages/.

P.J. Simko, J.S. Wallace, and J. Comprix. *Financial accounting for executives and MBAs*. Cambridge Business Publishers, 2020. ISBN 978-1-61853-366-1. URL https://books.google.com/books?id=u-2NzQEACAAJ.

Charalambos T. Spathis. Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*, 17(4):179–191, January 2002. ISSN 0268-6902. doi: 10.1108/02686900210424321. URL https://doi.org/10.1108/02686900210424321. Publisher: MCB UP Ltd.

Hervé Stolowy and Gaétan Breton. Accounts Manipulation: A Literature Review and Proposed Conceptual Framework. *Review of Accounting and Finance*, 3(1):5–92, January 2004. ISSN 1475-7702. doi: 10.1108/eb043395. URL https://doi.org/10.1108/eb043395. Publisher: Emerald Group Publishing Limited.

Phoenix Neale Williams and Ke Li. Black-box sparse adversarial attack via multi-objective optimisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12291–12301, 2023.

Skyler Wu, Fred Lu, Edward Raff, and James Holt. Stabilizing linear passive-aggressive online learning with weighted reservoir sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, December 2024. URL https://openreview.net/forum?id=FNOBf6JM7r.

Zhaoyuan Yang, Zhiwei Xu, Jing Zhang, Richard Hartley, and Peter Tu. Adversarial purification with the manifold hypothesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16379–16387, Mar. 2024. doi: 10.1609/aaai.v38i15.29574. URL https://ojs.aaai.org/index.php/AAAI/article/view/29574.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, April 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x.