Data Poisoning Attacks against Conformal Prediction

Yangyi Li [*1] Aobo Chen [*1] Wei Qian [1] Chenxu Zhao [1] Divya Lidder [1] Mengdi Huai [1]

Abstract

The efficient and theoretically sound uncertainty quantification is crucial for building trust in deep learning models. This has spurred a growing interest in conformal prediction (CP), a powerful technique that provides a model-agnostic and distribution-free method for obtaining conformal prediction sets with theoretical guarantees. However, the vulnerabilities of such CP methods with regard to dedicated data poisoning attacks have not been studied previously. To bridge this gap, for the first time, we in this paper propose a new class of black-box data poisoning attacks against CP, where the adversary aims to cause the desired manipulations of some specific examples' prediction uncertainty results (instead of misclassifications). Additionally, we design novel optimization frameworks for our proposed attacks. Further, we conduct extensive experiments to validate the effectiveness of our attacks on various settings (e.g., the full and split CP settings). Notably, our extensive experiments show that our attacks are more effective in manipulating uncertainty results than traditional poisoning attacks that aim at inducing misclassifications, and existing defenses against conventional attacks are ineffective against our proposed attacks.

1. Introduction

Deep Neural Networks (DNNs) have achieved remarkable success in recent years. Although deep learning models work well in numerous fields, deploying such models in real-world applications often requires to appropriately quantify the uncertainty of their predictions. To tackle uncertainty issues, people have developed different uncertainty quantification techniques, including Bayesian neural networks (Trinh et al., 2022; Hobbhahn et al., 2022).

Among different uncertainty quantification techniques, conformal prediction (CP), pioneered by Vovk et al. (2005), has become a popular distribution-free technique to perform uncertainty quantification (Ndiaye, 2022; Fisch et al., 2022; Stutz et al., 2021; Fisch et al., 2021; Qian et al., 2024). The model-agnostic and distribution-free nature of CP makes it particularly suitable for large neural networks. Concretely, we are mainly interested in a conformal set prediction setting where we are given n examples $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \cdots, n$ as calibration data, that are drawn exchangeably from some underlying distribution $\mathcal{P}$ (Humbert et al., 2023; Fisch et al., 2022; Lin et al., 2022; Teng et al., 2022). Let $X_{n+1} \in \mathcal{X}$ be a new exchangeable test example for which we would like to predict $Y^*_{n+1} = f(X_{n+1}; \theta) \in \mathcal{Y}$, where $\theta \in \Theta$ is a well-trained model. CP aims to construct a conformal prediction set, i.e., $\mathcal{C}_\varepsilon(x_{n+1}; \theta)$, that contains Y^*_{n+1} with marginal coverage at a significance level $\varepsilon \in (0, 1)$, i.e.,

$$\mathbb{P}(Y^*_{n+1} \in \mathcal{C}_\varepsilon(X_{n+1}; \theta)) \geq 1 - \varepsilon. \tag{1}$$

A conformal model is considered to be valid if the frequency of error, $Y^*_{n+1} \notin \mathcal{C}_\varepsilon(X_{n+1}; \theta)$, remains below the threshold ε. CP offers straightforward uncertainty estimates, where larger conformal sets $\mathcal{C}$ generally convey higher uncertainty.

Although CP is being increasingly used in safety-critical and security related applications, there's still a gap in understanding the effects of poisoning attacks on CP, an area that remains largely unexplored. In practice, the risk of data poisoning attacks (Yang et al., 2023; Jagielski et al., 2021; Qian et al., 2023) intensifies in DNNs, since they rely on large and diverse datasets and their size makes it difficult to guarantee the trustworthiness of the training data. As a result, models trained on such datasets are susceptible to data poisoning attacks, wherein an adversary places specifically constructed poisoned examples into the training data with harmful intentions (e.g., leading to unequalized and unfair coverage outcomes). In Schwarzschild et al. (2021), industry practitioners have identified data poisoning as the most significant concern among various threats.

In this work, we perform the first study on data poisoning attacks against CP, where the adversary aims to undermine the use of CP techniques by manipulating conformal prediction sets while ensuring label correctness. Consider a scenario

[*] Equal contribution [1] Department of Computer Science, Iowa State University, United States. Correspondence to: Mengdi Huai <mdhuai@iastate.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

where a doctor relies on conformal prediction to distinguish if a model prediction is reliable enough or requires more attention from the doctor. An adversary can compromise this process through poisoning attacks. Such attacks affect the model's ability to accurately estimate uncertainty, leading to potential risks in medical decision-making. Traditional data poisoning attacks (Zhao et al., 2024; Jagielski et al., 2021; Li et al., 2021; Geiping et al., 2021b; Peri et al., 2020; Foret et al., 2020; Qian et al., 2023) mainly focus on inducing misclassifications, whereas we focus on vulnerabilities related to the model's prediction uncertainties. Data poisoning attacks against CP could be more subtle and harder to detect compared to traditional poisoning attacks, since these attacks manipulate the model's prediction uncertainties rather than directly altering label predictions and might bypass existing defenses that are dependent on label changes.

While there are a few existing works (Ghosh et al., 2023; Gendler et al., 2021; Zhao et al., 2023) addressing test-time adversarial attacks on CP, they do not consider the risks of data poisoning attacks during the training process. Compared with these existing adversarial attacks, performing data poisoning attacks in the above discussed CP setting. is more stealthy, due to the preservation of the data exchangeability assumption in such scenarios. Notably, among the limited existing works on adversarial attacks against CP, they usually focus on how to ensure the validity (i.e., the coverage guarantee in Eq. (1)) under the violations of the data exchangeability, and fail to consider the maliciously manipulated efficiency, where the adversary targets prediction confidence. For example, the adversary might deliberately craft the poisoning training data to cause unequalized coverage probabilities that fail for specific sub-populations. Therefore, these adversarial robust CP methods are not equipped to counteract our proposed data poisoning attacks.

Motivated by the above, we thus believe that studying poisoning attacks targeting the prediction uncertainty is essential for safety applications of CP. In this work, we move the first step towards this direction, i.e., understanding the effects of data poisoning attacks on CP. To this end, we design a novel bi-level poisoning attack framework to craft effective poisoning points in the black-box setting. In the proposed framework, we first design approximate relaxation to handle the discrete conformal sets and the non-differential quantile. We also present a new worst-case adversarial loss to maximize the poisoning effect on the worst-case model for a strong poisoning effect. Further, we present novel efficient optimization methods by rigorously refining our attacks of generating effective poisoning points through the closed-form updates, thus eliminating the need for extensive model retraining or full access to training data. We conduct thorough experiments to verify the effectiveness of our attacks in various scenarios, including both full and split settings. Our detailed analysis reveals that these attacks are more successful at manipulating uncertainty outcomes than conventional poisoning attacks. Moreover, we found that current defenses against traditional poisoning attacks do not effectively counter our proposed attacks, underscoring the need for new strategies to address these advanced forms of data poisoning. The findings underscore the potential negative impacts of poisoning attacks on CP, aiming to raise awareness within the research community about this issue.

2. Related Work

Compared with traditional uncertainty estimation techniques, CP (Vovk et al., 2005) is a general framework for constructing conformal confidence sets, with the remarkable properties of being distribution-free, having coverage guarantees, and being able to be adapted to any estimator. However, previous literature on uncertainty estimation (Ren et al., 2023; Ledda et al., 2023; Alarab & Prakoonwit, 2022; Wicker et al., 2020; Yuan et al., 2020; Wang et al., 2018; 2022) has not delved into the vulnerability of CP to data poisoning attacks. On the other hand, data poisoning attacks at training time have emerged as a threat perceived to be of significant potential threat. Traditional poisoning attacks primarily deceive the model into making incorrect predictions. However, the distinct characteristics of both split CP and full CP (e.g., the coverage guarantees) pose challenges in directly applying these conventional poisoning attacks.

On the other hand, existing defenses against data poisoning attacks primarily depend on either anomaly detection based on nearest neighbors, training loss, singular-value decomposition, clustering (Peri et al., 2020; Cretu et al., 2008; Tran et al., 2018; Chen et al., 2018; Steinhardt et al., 2017), or robust training based on randomized smoothing, ensembling, data augmentation, and adversarial training (Weber et al., 2023; Li et al., 2021; Tao et al., 2021; Levine & Feizi, 2020; Ma et al., 2019; Abadi et al., 2016). For example, Peri et al. (2020) filters examples whose class labels differ from those of their nearest neighbors in the feature space. However, our proposed attacks are tailored to exploit vulnerabilities related to prediction uncertainty via CP, a nuance that these existing defenses may not be designed to handle. Additionally, compared with existing works (Jagielski et al., 2021; Geiping et al., 2021b; Koh & Liang, 2017) that do not consider retraining in the end-to-end manner, our proposed attacks maximize the poisoning effects of the worst-case model during optimization, resulting in enhanced stability for strong attack performance. Additionally, our proposed attacks also present a more rigorous derivation of our attack optimization methodology through the closed-form gradient updates between the poisoned and benign models. Importantly, these closed-form updates afford our optimization framework the ability to modulate its precision and computational cost by adjusting the model update precision.

3. Preliminaries

We assume pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ have a joint distribution denoted as $\mathcal{P}$, with the marginal distributions of X and Y and the conditional distribution $Y|X$ denoted as $\mathcal{P}_X$, $\mathcal{P}_Y$, and $\mathcal{P}_{Y|X}$, respectively. Given a new sample X, for every candidate label $Y \in \mathcal{Y}$, CP applies a simple test to either accept or reject the null hypothesis that pair (X, Y) is correct (Fisch et al., 2021). The test statistic for this test is a nonconformity measure, $\mathcal{S}((X, Y); \theta)$, where θ is a model fit to the training data using some learning algorithm. Informally, a lower value of $\mathcal{S}$ reflects (X, Y) conforms to the training data, whereas a higher value of $\mathcal{S}$ reflects that (X, Y) is atypical relative to the training data.

Assumption 3.1 (Exchangeability)**.** Consider the calibration data $Z_1 = (X_1, Y_1), \cdots, Z_n = (X_n, Y_n)$ and the test data $Z_{n+1} = (X_{n+1}, Y_{n+1})$. The examples are exchangeable if any permutation yields the same distribution, i.e.,

$$(Z_1, \cdots, Z_{n+1}) \overset{d}{=} (Z_\tau(1), \cdots, Z_{\tau(n+1)}), \tag{2}$$

with arbitrary permutation τ of the integers $1, \cdots, n + 1$.

Let $\mathcal{D} = \{Z_i = (X_i, Y_i)\}_{i=1}^n$ denote a calibration set of exchangeable (see Assumption 3.1) and correctly labeled examples. To determine the conformal prediction set for a test sample X, the classifier tests the nonconformity score for each potential label Y, against a pre-defined significance level ε, and includes all Y for which the null hypothesis—that the candidate data pair (X, Y) is conformal—is not rejected. This is achieved by comparing the nonconformity score of the test candidate against the nonconformity scores computed over the calibration dataset $\mathcal{D}$. This comparison uses the below quantile

$$Q_{1-\varepsilon}(\mathcal{D}, \theta) := \text{Quantile}(1 - \varepsilon; \{\mathcal{S}((X_i, Y_i); \theta)\}_{i=1}^n \cup \{\infty\}). \tag{3}$$

Note that compared with full CP, split CP is fast and easy to implement and model.

Theorem 3.2 (Vovk et al., 2005)**.** *Assume that examples (X_i, Y_i), $i = 1, \cdots, n + 1$ are exchangeable. For any nonconformity measure $\mathcal{S}$ and ε, define the conformal set (based on the first n examples) at $X_{n+1} \in \mathcal{X}$ as*

$$\mathcal{C}_\varepsilon(X_{n+1}; \theta) = \{Y_{n+1} \in \mathcal{Y} : \mathcal{S}(X_{n+1}, Y_{n+1}) \leq \textit{Quantile}(1 - \varepsilon; \{\mathcal{S}((X_i, Y_i); \theta)\}_{i=1}^n \cup \{\infty\})\}. \tag{4}$$

Then $\mathcal{C}_\varepsilon(X; \theta)$ satisfies Eq. (1).

4. Problem Statement

4.1. Threat Model

We consider a realistic threat model, where the adversary has no knowledge of the internal model parameters and the training process of the victim model, and is unable to alter test data during the model's testing phase. Additionally, the adversary cannot gain knowledge of the data points adopted for training, and can inject a limited number of manipulated new points into the training data. This situation depicts an attack setting where the adversary spreads poisoned data that developers unknowingly compile, along with vast benign data, to form the model's training set. However, we allow the adversary to have the computational capability required to train a pre-trained model $\theta^*(\mathcal{D})$ with a separate auxiliary dataset $\mathcal{D}$, which is comparable to the victim model (Jagielski et al., 2021; Geiping et al., 2021b). $\mathcal{D}$ is similar to the training data owned by the model owner and sampled from the distribution. Note that, for the assumption that the adversary is able to access a pre-trained model trained over an auxiliary dataset, it is reasonable given the widespread availability of public data. It has been a common assumption for black-box attacks in existing literature (Jagielski et al., 2021). Additionally, we also consider the white-box setting (Chen & Gu, 2020; Huai et al., 2020a; Neekhara et al., 2021; Wang et al., 2021; Huai et al., 2022; Gluch & Urbanke, 2021; Liu et al., 2024; Suya et al., 2021; Schwarzschild et al., 2021). In this scenario, the adversary has full access to the threat model's training data and network architecture.

The adversary aims to interfere with conformal predictions either by inducing *overconfidence CP attacks*, leading the model to underestimate prediction uncertainty, or *underconfidence CP attacks*, making the model underconfident by widening its conformal prediction sets. Additionally, to ensure stealth, we also consider maintaining the same coverage guarantees and executing targeted attacks without compromising uncertainty accuracy in benign samples. Note that our proposed data poisoning attacks can also be utilized to cause unequalized coverage subgroups.

4.2. Attack Formulation

Here, we propose our attacks for crafting poisoning samples against CP. As discussed in Section 3, in CP, we first split the auxiliary dataset $\mathcal{D}$ into a training fold $\mathcal{D}^{tr}$ and a calibration fold $\mathcal{D}^{ca}$. Then, based on the learning algorithm $\mathcal{A}$ and the training data $\mathcal{D}^{tr}$, the adversary can train a model f with parameters θ which correctly classifies as many data points as possible, maximizing $\mathbb{E}_{X,Y\sim\mathcal{D}^{tr}}\mathbb{I}(f(X;\theta) = Y)$, where $\mathbb{I}$ is the indicator function. We denote the training loss over the training data as $\mathcal{L}(\theta; \mathcal{D}^{tr}) = \frac{1}{n}\sum_{i=1}^n l(f(X_i; \theta), Y_i)$. We denote the set of victim target samples as $\{(X_v, Y_v)\}_{v=1}^V$. We assume that the adversary selects a subset $\mathcal{D}_p^{tr}$ from $\mathcal{D}^{tr}$ which takes an $\xi_1 \in [0, 1]$ percentage of $\mathcal{D}^{tr}$, and replaces it with a poisoning set $\tilde{\mathcal{D}}_p^{tr}$. We denote the remaining clean data as $\mathcal{D}_c^{tr} = \mathcal{D}^{tr} \setminus \mathcal{D}_p^{tr}$. For simplicity, we will omit the superscripts for $\mathcal{D}_p^{tr}$, $\mathcal{D}_c^{tr}$ and $\tilde{\mathcal{D}}_p^{tr}$ in the following. The effective poisoning points can be obtained by solving the following formulated optimization problem

$$\begin{aligned}\mathcal{D}_p^* \leftarrow & \arg\max_{\tilde{\mathcal{D}}_p} \ell_1(\{X_v\}_{v=1}^V; \theta(\tilde{\mathcal{D}}_p), Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p))) \\ & = \sum_{v=1}^{V} |\mathcal{C}_\varepsilon(X_v; \theta(\tilde{\mathcal{D}}_p))| + \sum_{v=1}^{V} \mathbb{I}(Y_v^* = f(X_v; \theta(\tilde{\mathcal{D}}_p))) \\ & + \sum_{v=1}^{V} \mathbb{I}(Y_v^* \in \mathcal{C}(X_v; \theta(\tilde{\mathcal{D}}_p))),\end{aligned} \tag{5}$$

where $\theta(\tilde{\mathcal{D}}_p)$ is obtained by training on the poisoned data $\tilde{\mathcal{D}}_{tr} = \tilde{\mathcal{D}}_p \cup \mathcal{D}_c$, and $\mathcal{C}_\varepsilon(X_v; \theta(\tilde{\mathcal{D}}_p)) = \{Y \in \mathcal{Y} : \mathcal{S}(X_v, Y; \theta(\tilde{\mathcal{D}}_p)) < Q_{1-\varepsilon}(\mathcal{D}^{ca}, \theta(\tilde{\mathcal{D}}_p))\}$. $Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p))$ is the new quantile calculated from the poisoned model $\theta(\tilde{\mathcal{D}}_p)$. Without loss of generality, we here focus on scenarios where the adversary aims to increase the prediction uncertainty by enlarging the sizes of conformal prediction sets. The second and third loss terms in the above equation are designed to ensure correct label predictions and the inclusion of true labels in post-attack conformal prediction sets, respectively. This enhances attack stealthiness without impacting coverage results and altering label predictions. Note that the above equation is a bi-level optimization problem—the minimization for $\tilde{\mathcal{D}}_p$ involves the model parameters $\theta(\tilde{\mathcal{D}}_p)$, which are themselves the minimizer of the following training problem

$$\theta(\tilde{\mathcal{D}}_p) = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta; \tilde{\mathcal{D}}_{tr} = \tilde{\mathcal{D}}_p \cup \mathcal{D}_c). \tag{6}$$

Note that Eq. (5) and (6) provide a high-level formulation for crafting poisoning examples $\tilde{\mathcal{D}}_p$ to increase the conformal set sizes (i.e., $|\mathcal{C}_\varepsilon(X_v; \theta(\tilde{\mathcal{D}}_p))|$). However, directly solving this framework is infeasible due to the discrete nature of the conformal sets. Recall that the conformal prediction set $\mathcal{C}_\varepsilon(X_v; \theta(\tilde{\mathcal{D}}_p))$ (defined in Eq. (4)) is based on comparing nonconformity scores to a threshold. A straightforward way is to directly adopt the quantile to formulate the relative comparison. However, this is impractical due to the difficulty of expressing the quantile $Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p))$ in a continuous and differential way. To overcome this, we develop a more feasible method, drawing upon the derivation method of $Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p))$ in Eq. (3). Based on this, we can have

$$\begin{aligned}& \min_{\tilde{\mathcal{D}}_p} \ell_2(\{X_v\}_{v=1}^V; \theta(\tilde{\mathcal{D}}_p), Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p))) = \sum_{v=1}^{V} \sum_{i=1}^{n_\varepsilon^{ca}} [\\ & \sum_{Y \in \mathcal{C}_\varepsilon(X_v; \theta(\tilde{\mathcal{D}}_p)) \cup \mathcal{Y}_a} \max(\mathcal{S}(X_v, Y; \theta(\tilde{\mathcal{D}}_p)) - \\ & \mathcal{S}(X_i, Y_i^*; \theta(\tilde{\mathcal{D}}_p)), 0)] + \sum_{v=1}^{V} \max(\max_{Y \neq Y_v^*} f_Y(X_v; \theta(\tilde{\mathcal{D}}_p)) \\ & - f_{Y_v^*}(X_v; \theta(\tilde{\mathcal{D}}_p)), -\beta),\end{aligned} \tag{7}$$

where $n_\varepsilon^{ca} = \lceil (1-\varepsilon) * |\mathcal{D}^{ca}| \rceil$, $\mathcal{Y}_a$ is the set of labels we aim to add into the prediction set, and β is a constant. Since the second and third terms are non-convex and non-differential, we design the surrogate losses to approximate them.

Note that the above optimization in Eq. (7) and (6) is designed to craft effective poisoning samples to fulfill the adversary's objectives, which are then injected into the dataset of the model owner. However, during re-training for optimization, the poisoned model $\theta(\tilde{\mathcal{D}}_p)$ can converge differently due to training uncertainties like model initialization and hyperparameter choice. Consequently, this can diminish the effectiveness of these poisoning samples and reduce their overall poisoning impact. To address this, we propose to focus on the worst-case poisoned model, which is the inner minima in Eq. (6) that has the worst poisoning effect (Andriushchenko & Flammarion, 2022; Wen et al., 2022). Our key idea here is to maximize the poisoning effect of the worst-case model to ensure that a high poisoning effect is preserved for other models. We then can formulate the worst-case poisoned model as $\theta' = \arg\max_{\theta \in \Theta_p} \ell_2(\{X_v\}_{v=1}^V; \theta(\tilde{\mathcal{D}}_p), Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p)))$, where $\Theta_p = \{\theta : \mathcal{L}(\theta; \tilde{\mathcal{D}}_{tr} = \tilde{\mathcal{D}}_p \cup \mathcal{D}_c) \leq \tau_1\}$ is the poisoned model space that is the set of all models that are trained on poisoned dataset and have a small training loss. Then, based on the notion of model sharpness (Foret et al., 2020), we can approximate the worst-case loss $\ell_2(\theta')$ by $\ell_2(\theta') \approx \max_{||\zeta||_q \leq \rho} \ell_2(\{X_v\}_{v=1}^V; \theta(\tilde{\mathcal{D}}_p) + \zeta, Q_{1-\varepsilon}(\theta(\tilde{\mathcal{D}}_p)))$. Therefore, we can obtain

$$\begin{aligned}\mathcal{D}_p^* \leftarrow & \arg\min_{\tilde{\mathcal{D}}_p} \ell_3(\{X_v\}_{v=1}^V; \theta(\tilde{\mathcal{D}}_p), Q_{1-\varepsilon}) \\ & = \arg\min_{\tilde{\mathcal{D}}_p} \max_{\|\zeta\|_p \leq \rho} \ell_2(\{X_v\}_{v=1}^V; \theta(\tilde{\mathcal{D}}_p) + \zeta, Q_{1-\varepsilon}),\end{aligned} \tag{8}$$

where $\theta(\tilde{\mathcal{D}}_p)$ is the minimizer of the training problem in Eq. (6). In the above, we locally maximize the loss by perturbing $\theta(\tilde{\mathcal{D}}_p)$ with a vector ζ (constrained by a norm limit $\|\zeta\|_p \leq \rho$). In this way, the perturbed model $\theta(\tilde{\mathcal{D}}_p) + \zeta$ has a worst poisoning effect compared to $\theta(\tilde{\mathcal{D}}_p)$.

The attack framework described in Eq. (8) and Eq. (6) is a bi-level optimization problem, where the outer optimization in Eq. (8) defines the adversarial attack objective and the inner problem in Eq. (6) specifies the model's learning objective using both the clean and poisoning data. Notably, compared with the original adversarial objective in Eq. (7), for the worst-case optimization in Eq. (8), the perturbations on the inner minima help achieve a strong poisoning effect.

4.3. Optimization

Fundamentally, the formulated bi-level optimization problem in Eq. (8) and Eq. (6) can be computationally expensive especially for DNNs, since we need to fully solve the inner problem in Eq. (6) to update the outer variables. Besides the high computation complexity, the inner optimization also incurs significant storage costs to maintain the entirety of the large training dataset. These raise a critical question: *“Is it possible to craft effective poisoning points without needing to retrain DNNs and accessing the entire training dataset?”* This question underscores the need for more resource-efficient strategies that circumvent the extensive computational and storage requirements typically associated with such end-to-end poisoning attacks (Foret et al., 2020).

To address the above challenges, we resort to formulating the optimization as a closed-form update of the original pre-trained model θ^*, while only knowing the subset $\mathcal{D}_p$. Specifically, we adopt influence functions (Hampel, 1974) to find an closed-form model update $\Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p)$ that we add to the original model θ^* (trained over $\mathcal{D}^{tr} = \mathcal{D}_p \cup \mathcal{D}_c$) for the generated poisoning samples. In this way, by capturing the changes to the pre-trained model θ^* in a closed-form update, we can provide significant speed-ups over existing retraining based methods (Huang et al., 2020). Our closed-form updates are not only limited to the feature-level manipulations, but also the labels. To map the changes of the training data in retrospection to close-form updates of model parameters, we can formulate

$$\theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p} = \arg\min_{\theta} \mathcal{L}_\xi(\theta; \mathcal{D}^{tr}) = \mathcal{L}(\theta; \mathcal{D}^{tr}) + \xi \sum_{\tilde{Z}_p \in \tilde{\mathcal{D}}_p} l(\tilde{Z}_p, \theta) - \xi \sum_{Z_p \in \mathcal{D}_p} l(Z_p, \theta). \quad (9)$$

The above generalization allows for the substitution of Z_p with $\tilde{Z}_p$ by slightly increasing the weight of $\tilde{Z}_p$ by a small value ξ and correspondingly decreasing Z_p. Below, we introduce our rigorously refined attacks based on the first-order and second-order closed-form gradient updates.

First-order case. To derive the first-order based update, when ξ is small and l is differential with respect to θ, we can use a first-order Taylor series at θ^* to approximate $\mathcal{L}_\xi(\theta; \mathcal{D}^{tr})$ in Eq. (9) by

$$\begin{aligned}\mathcal{L}_\xi(\theta^*_{\xi,Z_p\to\tilde{Z}_p}; \mathcal{D}^{tr}) \approx{}& \mathcal{L}(\theta^*; \mathcal{D}^{tr}) + \xi(l(\tilde{Z}_p, \theta^*) \\ &- l(Z_p, \theta^*)) + \Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p) \cdot (\nabla_\theta \mathcal{L}(\theta^*; \mathcal{D}^{tr}) \\ &+ \xi(\nabla_\theta l(\tilde{Z}_p; \theta^*) - \nabla_\theta l(Z_p; \theta^*))),\end{aligned} \quad (10)$$

where θ^* is obtained over $\mathcal{D}^{tr}$. Given that the poisoned model $\theta^*_{\xi,Z_p\to\tilde{Z}_p}$ is a minimum of $\mathcal{L}_\xi(\cdot; \mathcal{D}^{tr})$, we can assume that $\mathcal{L}_\xi(\theta^*_{\xi,Z_p\to\tilde{Z}_p}; \mathcal{D}^{tr}) < \mathcal{L}_\xi(\theta^*; \mathcal{D}^{tr})$. Integrating this into the Taylor series approximation and using the condition that $\nabla_\theta \mathcal{L}(\theta^*; \mathcal{D}^{tr}) = 0$, based on Eq. (9), we now can have $\xi\Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p) \cdot (\nabla_\theta l(\tilde{Z}_p, \theta^*) - \nabla_\theta l(Z_p, \theta^*)) < 0$. Given $\xi > 0$, our attention shifts to analyzing the dot product within the equation. For two given vectors μ_1, μ_2, the dot product can be expressed as $\mu_1 \cdot \mu_2 = ||\mu_1||||\mu_2|| \cos(\mu_1, \mu_2)$, where $\cos(\mu_1, \mu_2)$ is the cosine between μ_1 and μ_2. The minimum cosine, -1, occurs when $\mu_1 = -\mu_2$. Therefore, we can arrive at

$$\Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p) = \sum_{Z_p \in \mathcal{D}_p} \nabla_\theta l(Z_p, \theta^*) - \sum_{\tilde{Z}_p \in \tilde{\mathcal{D}}_p} \nabla_\theta l(\tilde{Z}_p, \theta^*),$$

which indicates the optimal direction for adjustment from θ^* is $\sum_{\tilde{Z}_p \in \tilde{\mathcal{D}}_p} \nabla_\theta l(\tilde{Z}_p, \theta^*) - \sum_{Z_p \in \mathcal{D}_p} \nabla_\theta l(Z_p, \theta^*)$. The actual step size is unknown and requires calibration with a small constant τ to determine the appropriate update magnitude. Based on this, we can have

$$\begin{aligned}\theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p} \approx{}& \theta^* - \\ &\tau\Big(\sum_{\tilde{Z}_p \in \tilde{\mathcal{D}}_p} \nabla_\theta l(\tilde{Z}_p, \theta^*) - \sum_{Z_p \in \mathcal{D}_p} \nabla_\theta l(Z_p, \theta^*)\Big).\end{aligned} \quad (11)$$

Intuitively, this update shifts the model parameters from $\sum_{\tilde{Z}_p \in \tilde{\mathcal{D}}_p} \nabla_\theta l(\tilde{Z}_p, \theta^*)$ to $\sum_{Z_p \in \mathcal{D}_p} \nabla_\theta l(Z_p, \theta^*)$, with τ dictating the update's step size.

Next, we can use a first-order Taylor series around θ^* to approximate $\ell_3(X_v; \theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p}, Q_{1-\varepsilon})$ in Eq. (8) as follows

$$\begin{aligned}&\min_{\tilde{\mathcal{D}}_p} \ell_3(X_v; \theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p}, Q_{1-\varepsilon}) \\ ={}&\min_{\tilde{\mathcal{D}}_p} \ell_3(X_v; \theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p}, Q_{1-\varepsilon}) - \ell_3(X_v; \theta^*, Q_{1-\varepsilon}) \\ \approx{}&\min_{\tilde{\mathcal{D}}_p} \nabla_\theta \ell_3(X_v; \theta^*, Q_{1-\varepsilon}) \cdot [\theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p} - \theta^*] \\ ={}&\min_{\tilde{\mathcal{D}}_p} -\tau \nabla_\theta \ell_3(X_v; \theta^*, Q_{1-\varepsilon}) \cdot \Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p).\end{aligned} \quad (12)$$

Therefore, to induce a modification $\theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p} - \theta^*$ that can most increase the adversarial loss ℓ_3 on victim examples $\{(X_v, Y_v)\}_{v=1}^V$, we can minimize the above equation when ξ is small, i.e., maximizing $\nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon}) \cdot \Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p)$. Now the objective is to solve

$$\begin{aligned}&\arg\max_{\tilde{\mathcal{D}}_p} \Phi(\tilde{\mathcal{D}}_p, \theta) = \nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon}) \cdot \Psi(\mathcal{D}_p, \\ &\tilde{\mathcal{D}}_p)/(\|\nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon})\|\|\Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p)\|),\end{aligned} \quad (13)$$

which achieves the maximized attack goal by aligning the directions of $\nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon})$ and $\Psi(\mathcal{D}_p, \tilde{\mathcal{D}}_p)$. To compute the adversarial loss $\nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon})$, we adopt the technique in Foret et al. (2020) to first approximate ℓ_3 by leveraging a first-order method

$$\begin{aligned}&\hat{\zeta} = \rho \cdot \mathrm{sign}(\nabla_\theta \ell_2(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon}))|\nabla_\theta \ell_2(\{X_v\}_{v=1}^V; \\ &\theta^*, Q_{1-\varepsilon})|^{q-1}/(||\nabla_\theta \ell_2(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon})||_p^q)^{\frac{1}{p}},\end{aligned} \quad (14)$$

where $1/p + 1/q = 1$. We set $p = 2$, following Foret et al. (2020), unless otherwise stated. Then, we can have the approximation to calculate $\nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon})$ via replacing θ^* with $\theta^* + \hat{\zeta}$

$$\begin{aligned}\nabla_\theta \ell_3(\{X_v\}_{v=1}^V; \theta^*, Q_{1-\varepsilon}) \approx{}& \nabla_\theta \ell_2(\{X_v\}_{v=1}^V; \\ &\theta, Q_{1-\varepsilon})|_{\theta=\theta^*+\hat{\zeta}}.\end{aligned} \quad (15)$$

In this way, by fixing $\nabla_\theta \ell_3$, we can solve Eq. (13) to find effective poisoning samples $\tilde{\mathcal{D}}_p$ via gradient descent. In this way, the poisoned model is specifically tailored to exhibit malicious behavior towards the victim's data samples.

Second-order case. When the loss $\mathcal{L}(\theta;\mathcal{D}^{tr})$ is twice differentiable and strictly convex, there exists an inverse Hessian matrix $H_{\theta^*}^{-1}$, which allows for the approximation of changes to the model (Ling, 1984). In particular, the optimality conditions for Eq. (9) can be directly determined by $0 = \mathcal{L}(\theta^*;\mathcal{D}^{tr}) + \xi(l(\tilde{Z},\theta^*_{\xi,Z\to\tilde{Z}}) - l(Z,\theta^*_{\xi,Z\to\tilde{Z}}))$. If ξ is sufficiently small, we can use a first-order Taylor series at θ^* to approximate the conditions as

$$\begin{aligned}0 \approx \mathcal{L}(\theta^*;\mathcal{D}^{tr}) + \xi(l(\tilde{Z},\theta^*) - l(Z,\theta^*)) + (\theta^*_{\xi,Z\to\tilde{Z}} - \theta^*)\cdot \\ (\nabla^2\mathcal{L}(\theta^*;\mathcal{D}^{tr}) + \xi(\nabla^2 l(\tilde{Z};\theta^*) - \nabla^2 l(Z;\theta^*))).\end{aligned} \tag{16}$$

Given the optimality condition $\nabla\mathcal{L}(\theta^*;\mathcal{D}^{tr}) = 0$ for θ^*, using the Hessian of the loss function, we can rearrange this solution and get $\theta^*_{\xi,\mathcal{D}_p\to\tilde{\mathcal{D}}_p} - \theta^* = -\xi H_{\theta^*}^{-1}(\sum_{\tilde{Z}_p\in\tilde{\mathcal{D}}_p}\nabla_\theta l(\tilde{Z}_p,\theta^*) - \sum_{Z_p\in\mathcal{D}_p}\nabla_\theta l(Z_p,\theta^*))$, where we additionally omit higher-order terms. Then, we can set $\xi = 1$ to replace sample Z completely by $\tilde{Z}$, which leads to the below second-order update

$$\begin{aligned}\theta^*_{\mathcal{D}_p\to\tilde{\mathcal{D}}_p} \approx \theta^* - H_{\theta^*}^{-1}(\sum_{\tilde{Z}_p\in\tilde{\mathcal{D}}_p}\nabla_\theta l(\tilde{Z}_p,\theta^*) \\ - \sum_{Z_p\in\mathcal{D}_p}\nabla_\theta l(Z_p,\theta^*)).\end{aligned} \tag{17}$$

Combining this with Eq. (13), we can easily derive the second-order based attack framework.

Theorem 4.1. *Assume that $\mathcal{L}(\theta)$ is local convex and differentiable. Let $\mathcal{D}_p = \{(X_i,Y_i)\}_{i=1}^P$, $\mathcal{L}(\theta^*)$ be the initial optimal solution, and $\mathcal{L}(\theta_u^*)$ be the updated optimal solution. Given a bound $\epsilon > 0$ with the perturbation $\|\delta_i\|_2 \leq \epsilon$, assume that $\|\theta^* - \theta_u^*\|_2$ has a upper bound B_θ, the gradient ∇l is L_z-Lipschitz with respect to X at θ^* and L_1 -Lipschitz with respect to θ. We get $\theta^*_{\mathcal{D}_p\to\tilde{\mathcal{D}}_p}$ from θ^* by our closed-form updates. Then the following upper bounds hold: For the first-order update of our approach, if $\tau \leq \frac{1}{L_1}$ we have $\mathcal{L}_\xi(\theta^*_{\mathcal{D}_p\to\tilde{\mathcal{D}}_p}) - \mathcal{L}(\theta_u^*) \leq \epsilon L_z|\mathcal{D}_p|B_\theta$. For the second-order update of our approach, we have $\mathcal{L}_\xi(\theta^*_{\mathcal{D}_p\to\tilde{\mathcal{D}}_p}) - \mathcal{L}(\theta_u^*) \leq \epsilon B_\theta L_z|\mathcal{D}_p| + \left(1+\frac{1}{2}L_1^2\right)L_1(\epsilon L_z|\mathcal{D}_p|)^2$.*

Theorem 4.1 gives a finite-sample bound to quantify the difference between our two approximation methods and the optimal solution. It demonstrates that as we decrease the perturbation bound ϵ and the number of poisoned data $|\mathcal{D}_p|$, our methods approximate $\mathcal{L}(\theta_u^*)$ more closely. The procedure for optimizing the above losses is postponed to the full version of this paper. Notably, we can easily generalize this algorithm to perform other different attack types, e.g., adding irrelevant labels or removing specific labels regardless of the correctness of labels. This further demonstrates the significant threats of poisoning attacks against CP. Theorem 4.2 shows that under specific conditions regarding step sizes, the victim model will converge to a stationary point of the adversarial loss when the main training loss is optimized using stochastic gradient descent.

Theorem 4.2. *Let $\ell_3(\theta)$ be bounded below and have a Lipschitz continuous gradient with constant $L > 0$. Assume that the victim model is trained by stochastic gradient descent (SGD) with step sizes α_t , i.e. by sampling a random index $\tilde{i}_t$ uniformly from $\{1,\ldots,n\}$ and then updating $\theta^{t+1} = \theta^t - \alpha_t\nabla\mathcal{L}_{\tilde{i}_t}(\theta^t)$. If the gradient descent steps $\alpha_t > 0$ satisfy $\alpha_t L < \omega\Phi(\mathcal{D}_p,\theta^t)\frac{\|\nabla\ell_3(\theta^t)\|}{\|\nabla\mathcal{L}(\theta^t)\|}$ and $\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\tilde{i}_t}(\theta^t)\right\|^2\right] \leq \|\nabla\mathcal{L}(\theta^t)\|^2$ for some fixed $\omega < 2$, then $\mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] < \mathbb{E}\left[\ell_3\left(\theta^t\right)\right]$. If in addition $\exists\mu > 0$, t_0 and $\forall t \geq t_0, \Phi(\mathcal{D}_p,\theta^t) > \mu$, we then can have $\lim_{t\to\infty}\|\nabla\ell_3(\theta^t)\| \to 0$.*

Proof. For $\ell_3\left(\theta^{t+1}\right)$, we can have the following

$$\begin{aligned}\ell_3\left(\theta^{t+1}\right) =& \ell_3\left(\theta^t - \alpha_t\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)\right) \\ \leq& \ell_3\left(\theta^t\right) - \alpha_t\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)^\top\nabla\ell_3\left(\theta^t\right) \\ &+ \frac{1}{2}\alpha_t^2 L\left\|\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)\right\|^2.\end{aligned} \tag{18}$$

If we take the expected value of both sides of this expression (where the expectation is taken over the randomness in the sample selection $\tilde{i}_t$), we get

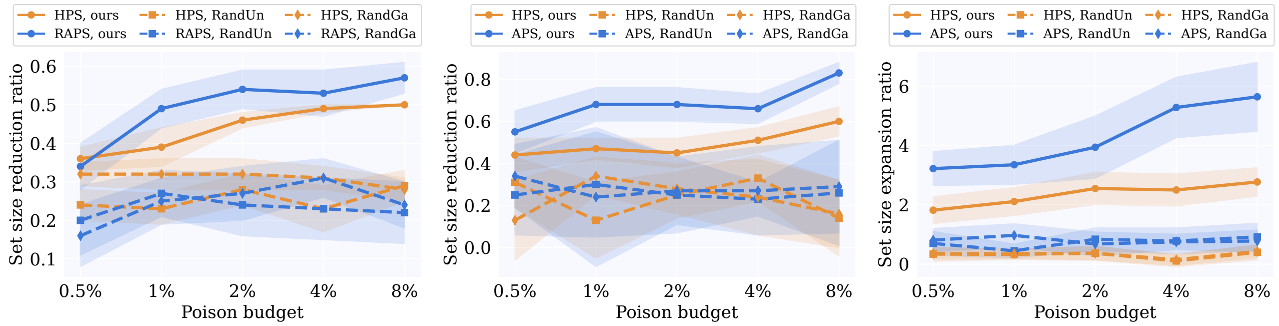
$$\begin{aligned}\mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] \leq \mathbb{E}\left[\ell_3\left(\theta^t\right)\right] - \alpha_t\mathbb{E}\left[\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)^\top\nabla\ell_3\left(\theta^t\right)\right] \\ + \frac{\alpha_t^2 L\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)\right\|^2\right]}{2}.\end{aligned} \tag{19}$$

Now, the expected value of $\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)$ given θ^t is $\mathbb{E}\left[\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)\mid\theta^t\right] = \sum_{i=1}^n\nabla\mathcal{L}_i\left(\theta^t\right)\cdot P\left(\tilde{i}_t = i\mid\theta^t\right) = \sum_{i=1}^n\nabla\mathcal{L}_i\left(\theta^t\right)\cdot\frac{1}{n} = \nabla\mathcal{L}\left(\theta^t\right)$. Based on this, we can have

$$\begin{aligned}\mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] \leq \mathbb{E}\left[\ell_3\left(\theta^t\right)\right] - \alpha_t\nabla\mathcal{L}\left(\theta^t\right)^\top\nabla\ell_3\left(\theta^t\right) \\ + \frac{\alpha_t^2 L\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)\right\|^2\right]}{2}.\end{aligned} \tag{20}$$

According to the assumption $\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\tilde{i}_t}\left(\theta^t\right)\right\|^2\right] \leq \|\nabla\mathcal{L}\left(\theta^t\right)\|^2$, we get

$$\begin{aligned}\mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] \leq \mathbb{E}\left[\ell_3\left(\theta^t\right)\right] - (\alpha_t\frac{\|\nabla\ell_3\left(\theta^t\right)\|}{\|\nabla\mathcal{L}\left(\theta^t\right)\|}\cos\left(\gamma^t\right) \\ - \frac{1}{2}\alpha_t^2 L)\left\|\nabla\mathcal{L}\left(\theta^t\right)\right\|^2.\end{aligned} \tag{21}$$



(a) CIFAR-10 (b) CIFAR-100 (c) Tiny-ImageNet

Figure 1. Performance of overconfidence CP attacks on CIFAR-10 and CIFAR-100, and underconfidence CP attacks on Tiny-ImageNet.

As such, the adversarial loss decreases for nonzero step sizes if $\frac{\|\nabla \ell_3(\theta^t)\|}{\|\nabla \mathcal{L}(\theta^t)\|}\cos(\gamma^t) > \frac{1}{2}\alpha_t L$ for some $1/2 < c < \infty$. This follows from our assumption on the parameter ω. Therefore, we can get $\mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] < \mathbb{E}\left[\ell_3\left(\theta^{t}\right)\right]$. Reinserting this estimate into Eq. (21) reveals that

$$\mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] \leq \mathbb{E}\left[\ell_3\left(\theta^{t}\right)\right] - \frac{\cos^2\gamma^t}{2L}\left\|\nabla \ell_3\left(\theta^t\right)\right\|^2.$$

Due to monotonicity we may sum over all descent inequalities, yielding $\sum_{t=0}^{t=T-1}\mathbb{E}\left[\ell_3\left(\theta^{t}\right)\right] - \mathbb{E}\left[\ell_3\left(\theta^{t+1}\right)\right] \geq \sum_{t=0}^{t=T-1}\frac{\cos^2\gamma^t}{2L}\left\|\nabla \ell_3\left(\theta^t\right)\right\|^2$, then

$$\begin{aligned}\ell_3\left(\theta^0\right) - \ell_3^* &\geq \ell_3\left(\theta^0\right) - \mathbb{E}\left[\ell_3\left(\theta^T\right)\right] \\ &\geq \sum_{t=0}^{t=T-1}\frac{\cos^2\gamma^t}{2L}\left\|\nabla \ell_3\left(\theta^t\right)\right\|^2 \qquad (22)\end{aligned}$$

where ℓ_3^* is the global optimum of ℓ_3. When $T \to \infty$ we can find

$$\sum_{t=0}^{\infty}\frac{\cos^2\gamma^t}{2L}\left\|\nabla \ell_3\left(\theta^t\right)\right\|^2 < \infty. \qquad (23)$$

According to the assumption that $\cos\gamma^t$ is bounded below by some fixed $\mu > 0$ except finitely many iterates for all (i.e., the angle between adversarial and training gradient is less than 90°), we have the convergence to a stationary point as $\sum_{t=0}^{\infty}\frac{\mu^2}{2L}\left\|\nabla \ell_3\left(\theta^t\right)\right\|^2 < \infty$. Therefore, we can get

$$\lim_{t\to\infty}\left\|\nabla \ell_3\left(\theta^t\right)\right\| \to 0. \qquad (24)$$

□

Discussions on poisoning attacks against full conformal prediction. In full conformal prediction, it assumes that both training and test data are exchangeable. Therefore, directly crafting perturbation-based poisoning samples would violate the data exchangeability assumption. This would increase the risks of being detected by just checking coverage results. One straightforward way is to inject exchangeable samples without perturbations. However, such a method is limited in attack effectiveness and the availability of a large number of exchangeable points. To study the effects of poisoning attacks on full conformal prediction while maintaining validity, we can employ transfer learning-based attack settings (Shen et al., 2021; Shafahi et al., 2018), where the adversary has knowledge of a pre-trained model and the victim model is fine-tuned on this pre-trained model. Due to space limitation, more details about poisoning attacks against full conformal prediction can be found in the full version of this paper.

5. Experiments

In this section, we perform extensive experiments to validate our proposed poisoning attacks against conformal prediction. Due to space limitation, more experimental details and results (e.g., more datasets, and attacks scenarios for unequalized and unfair coverage subgroups) are given in the full version of this paper.

Datasets and models. In experiments, we adopt the following image classification datasets: Tiny-ImageNet (Deng et al., 2009) and CIFAR-10/100 (Krizhevsky et al.). We consider various DNN models, including MobileNet-V2 (Sandler et al., 2018), ResNet-18 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2014), and a 5-layer ConvNet.

Baselines. As there is no existing work on data poisoning attacks against conformal prediction, in our experiments, we adopt the *RandUn* and *RandGa* baselines to assess the effectiveness of the proposed poisoning strategies. Specifically, we use random uniform noise and Gaussian noise as poisoning perturbations for the *RandUn* and *RandGa* baselines, respectively.

Evaluation metrics. To evaluate the attack effectiveness, we measure the *set size reduction ratio* as $(set\_size_{benign} - set\_size_{victim})/set\_size_{benign}$ and *set size expansion ratio* as $(set\_size_{victim} - set\_size_{benign})/set\_size_{benign}$ of the target samples on the victim model. In addition, we analyze *prediction*

Table 1. Set size reduction ratio of overconfidence CP attacks under data poisoning defenses.

Defense method	HPS	APS	RAPS	RSCP
Ours + No defense	0.46 ± 0.02	0.47 ± 0.04	0.54 ± 0.04	0.49 ± 0.05
Ours + MaxUp (Gong et al., 2021)	0.34 ± 0.05	0.32 ± 0.07	0.49 ± 0.06	0.26 ± 0.07
Ours + Adversarial Poisoning (Geiping et al., 2021a)	0.39 ± 0.04	0.28 ± 0.09	0.41 ± 0.06	0.24 ± 0.07
Ours + EPIC (Yang et al., 2022)	0.38 ± 0.06	0.33 ± 0.09	0.42 ± 0.07	0.38 ± 0.10

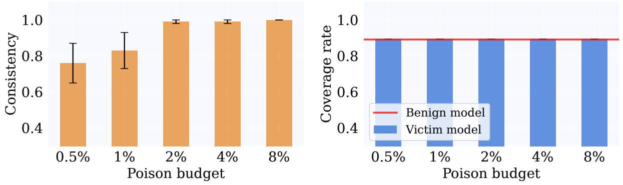
consistency (whether the prediction labels are consistent) and *empirical coverage rate* between benign and victim models to show the stealthiness of our attacks.

The adopted conformal methods. In experiments, we adopt the following popular conformal methods: RSCP (Gendler et al., 2021), an adversarial robust CP method against adversarial attacks; APS (Romano et al., 2020), designed to improve conditional coverage; RAPS (Angelopoulos et al., 2020), a regularized variant of APS for generating smaller sets; and HPS (Lei et al., 2013; Vovk et al., 2005), which relies on softmax output.

Implementation details. In experiments, we allocate 10% data for calibration and maintain a default coverage rate $(1 - \varepsilon)$ of 0.9. We limit the perturbation bound ϵ to $16/255$. The poisoning attacks are implemented through training the models from scratch (Huang et al., 2020; Huai et al., 2020b), utilizing the SGD optimizer with a learning rate of 0.01 and a batch size of 128. We evaluate the attack results in each experiment by randomly sampling a target class. We generate poisons and evaluate them on 8 newly initialized victim models. We repeat each experiment 10 times and report the mean and standard errors.

5.1. Attack Performance against Conformal Prediction

In Figure 1, we present the performance of overconfidence CP attacks on CIFAR-10 and CIFAR-100, as well as the underconfidence CP attacks on Tiny-ImageNet. We observe that our proposed attacks significantly outperform RandUn and RandGa baselines in terms of set size reduction ratio and set size expansion ratio across various poison budgets. For example, consider Figure 1a, where overconfidence CP attacks are conducted on CIFAR-10 with HPS and RAPS. The benign set size of HPS is 2.0 (implying a maximum reduction ratio of 0.5 in order to obtain a set size of 1), and the benign set size of RAPS is about 2.84 (with a maximum reduction ratio of 0.64). Our proposed attacks achieve a reduction ratio of 0.48 with HPS and 0.54 with RAPS using 2% poison budget, while the baselines achieve reduction ratios below 0.32. Therefore, our proposed attacks can effectively manipulate the uncertainty of CP and successfully trick the model into being overconfident or underconfident for target samples.



(a) Prediction consistency

(b) Empirical coverage rate

Figure 2. Stealthiness of overconfidence CP attacks on CIFAR-10.

In addition, in Figure 2, we demonstrate the stealthiness of overconfidence CP attacks on CIFAR-10 with HPS. Our proposed attacks achieve a high prediction consistency and similar empirical convergence rates compared to the benign model. This underscores the stealthiness of our attacks when targeting uncertainty in CP.

5.2. Attack Performance under Data Poisoning Defenses

In this section, we explore the performance of our proposed attacks under existing data poisoning defenses. In Table 1, we report the set size reduction ratio of overconfidence CP attacks under MaxUp (Gong et al., 2021), Adversarial Poisoning (Geiping et al., 2021a), and EPIC (Yang et al., 2022), using 2% poison budget on CIFAR-10. Specifically, MaxUp generates augmented data with random perturbations, aiming to minimize the worst-case loss of the augmented data. Adversarial Poisoning is a variant of adversarial training that builds a robust model using adversarially poisoned data. EPIC identifies and eliminates effective poison data in gradient space during training to prevent poisoning attacks. Notably, our proposed attacks remain effective even under these existing poisoning defenses since we specifically target the nonconformity scores in our attack framework. For example, it still achieves a reduction ratio of 0.34 under MaxUp with HPS, compared to 0.46 without defense. Therefore, our proposed attacks demonstrate a satisfying set size reduction ratio across existing defense mechanisms, indicating the utility and effectiveness of our approach.

5.3. Ablation Study

First, we compare the performance and running time of overconfidence CP attacks with different optimizations against HPS on CIFAR-10. The results in Table 2 reveal that our proposed attacks, both in first-order and second-order op-

Table 2. Set size reduction ratio and running time (min) of overconfidence CP attacks with varying optimizations.

Poison budget	Ours – first-order		Ours – second-order		MetaPoison (Huang et al., 2020)	
	Reduction ratio	Running time	Reduction ratio	Running time	Reduction ratio	Running time
0.5%	0.36 ± 0.03	12.75 ± 0.09	0.38 ± 0.05	95.76 ± 0.37	0.31 ± 0.04	233.48 ± 0.38
1%	0.39 ± 0.05	14.96 ± 0.25	0.40 ± 0.05	179.77 ± 0.96	0.36 ± 0.05	300.04 ± 1.93
2%	0.46 ± 0.02	19.26 ± 0.14	0.48 ± 0.02	369.23 ± 4.81	0.38 ± 0.07	494.72 ± 2.35

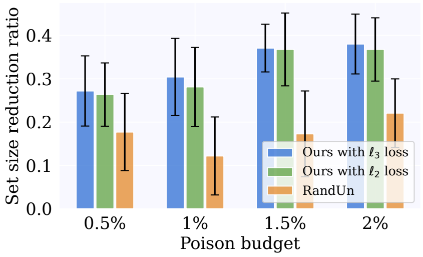


Figure 3. Impact of selected loss.

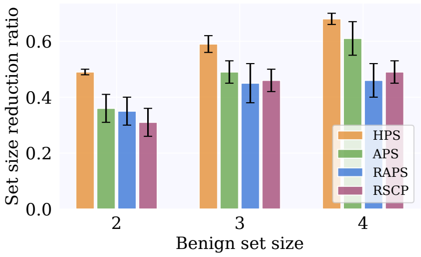


Figure 4. Impact of benign set size.

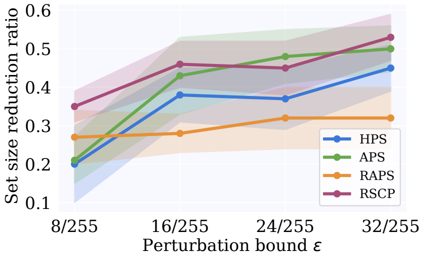


Figure 5. Impact of perturbation bound.

timizations, achieve a significantly higher set size reduction ratio and require much less running time compared to MetaPoison (Huang et al., 2020) optimization.

Next, we examine the performance of overconfidence CP attacks using the ℓ_2 loss and ℓ_3 loss on various poison budgets in the practical black-box scenario. Note that unlike ℓ_2, ℓ_3 considers the worst-case poisoned model. As shown in Figure 3, our attacks demonstrate significantly higher set size reduction ratios compared to the RandUn baseline. When comparing the two loss functions, we observe that under budgets of 0.5% and 1%, attacks employing the ℓ_3 loss achieve higher reduction ratios of 0.01 and 0.02, respectively, than the ℓ_2 loss. This indicates we can make an improvement by taking into account the worst case of the model when conducting the poisoning attacks.

Furthermore, we conduct overconfidence CP attacks on varying benign set sizes, using 2% poison budget on CIFAR-10. As shown in Figure 4, our proposed attacks consistently reduce the uncertainty for target samples across different benign set sizes. Typically, a larger prediction set implies more uncertainty and poses a greater challenge for attacks due to the need to manipulate more labels. Nonetheless, our attacks persist in showcasing their capability to reduce the set size. Our optimization approach specifically targets each nonconformity score associated with labels in the prediction set, ensuring that the attacked prediction set exclusively contains the predicted label, thereby reducing uncertainty.

Lastly, in Figure 5, we illustrate the performance of overconfidence CP attacks across different perturbation bounds employing ℓ_3 loss, using 2% poison budget on CIFAR-10 in the practical black-box scenario. The results show that our proposed attacks generally achieve higher set size reduction ratios with larger perturbation bounds. Even with a small perturbation bound (e.g., 16/255), our proposed attacks exhibit remarkable performance. The reason is that, as the perturbation bound increases, the adversary has more space to adjust the features of victim samples, allowing them to explore a broader range and find perturbations that deceive the model more effectively.

6. Conclusion and Future Work

For the first time to our best knowledge, in this paper, we study the vulnerabilities of CP to data poisoning attacks, and devise a bi-level attack framework for crafting effective poisoning points in black-box scenarios. Specifically, in our proposed strategy, we first propose to calculate the worst poisoning model before using it to update poisoning points, to maintain a strong poisoning effect across various models for maximizing the impact of our attacks. Additionally, we also design approximate relaxations for handling the discrete uncertainty set sizes and the non-convex, non-differentiable quantile. Further, we introduce rigorous optimization methods that refine our strategies for efficiently creating effective poisoning points using closed-form updates, thus bypassing the need for full model retraining or complete dataset access. Our extensive experiments in both full and split CP settings demonstrate our attacks' effectiveness in manipulating uncertainty, surpassing traditional poisoning methods. Moreover, we discover that existing defenses are inadequate against our advanced attack strategies.

In the future, we will extend our proposed attacks to a broader range of machine learning models, CP methods, and larger datasets. Notably, the attack strategies proposed in this paper could potentially be used by malicious users to attack real CP systems. To mitigate the potential negative consequences and impacts, we will design robust CP algorithms that can effectively defend against such poisoning attacks in our future work.

Impact Statement

In this paper, we introduce a novel class of data poisoning attacks tailored to compromise conformal prediction systems by manipulating the uncertainty estimate. This approach reveals vulnerabilities of conformal prediction, thereby shedding light on potential security breaches in the predicted conformal results for uncertainty estimation. Our results highlight the urgent need for further research to protect against such significant threats and improve the security and reliability of such uncertainty estimation methods.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alarab, I. and Prakoonwit, S. Adversarial attack for uncertainty estimation: identifying critical regions in neural networks. *Neural Processing Letters*, 54(3):1805–1821, 2022.

Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668. PMLR, 2022.

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., and Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Chen, J. and Gu, Q. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1739–1747, 2020.

Cretu, G. F., Stavrou, A., Locasto, M. E., Stolfo, S. J., and Keromytis, A. D. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 81–95. IEEE, 2008.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pp. 3329–3339. PMLR, 2021.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Conformal prediction sets with limited false positives. In *International Conference on Machine Learning*, pp. 6514–6532. PMLR, 2022.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Geiping, J., Fowl, L., Somepalli, G., Goldblum, M., Moeller, M., and Goldstein, T. What doesn't kill you makes you robust (er): How to adversarially train against data poisoning. *arXiv preprint arXiv:2102.13624*, 2021a.

Geiping, J., Fowl, L. H., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021b.

Gendler, A., Weng, T.-W., Daniel, L., and Romano, Y. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Ghosh, S., Shi, Y., Belkhouja, T., Yan, Y., Doppa, J., and Jones, B. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pp. 681–690. PMLR, 2023.

Gluch, G. and Urbanke, R. Query complexity of adversarial attacks. In *International Conference on Machine Learning*, pp. 3723–3733. PMLR, 2021.

Gong, C., Ren, T., Ye, M., and Liu, Q. Maxup: Lightweight adversarial training with data augmentation improves neural network training. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 2474–2483, 2021.

Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hobbhahn, M., Kristiadi, A., and Hennig, P. Fast predictive uncertainty for classification with bayesian deep networks. In *Uncertainty in Artificial Intelligence*, pp. 822–832. PMLR, 2022.

Huai, M., Sun, J., Cai, R., Yao, L., and Zhang, A. Malicious attacks against deep reinforcement learning interpretations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 472–482, 2020a.

Huai, M., Wang, D., Miao, C., Xu, J., and Zhang, A. Pairwise learning with differential privacy guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 694–701, 2020b.

Huai, M., Zheng, T., Miao, C., Yao, L., and Zhang, A. On the robustness of metric learning: an adversarial perspective. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(5):1–25, 2022.

Huang, W. R., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

Humbert, P., Le Bars, B., Bellet, A., and Arlot, S. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pp. 14153–14177. PMLR, 2023.

Jagielski, M., Severi, G., Pousette Harger, N., and Oprea, A. Subpopulation data poisoning attacks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3104–3122, 2021.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Ledda, E., Angioni, D., Piras, G., Fumera, G., Biggio, B., and Roli, F. Adversarial attacks against uncertainty quantification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4599–4608, 2023.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Levine, A. and Feizi, S. Deep partition aggregation: Provable defenses against general poisoning attacks. In *International Conference on Learning Representations*, 2020.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction with temporal quantile adjustments. *Advances in Neural Information Processing Systems*, 35:31017–31030, 2022.

Ling, R. F. Residuals and influence in regression, 1984.

Liu, Z., Wang, T., Huai, M., and Miao, C. Backdoor attacks via machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14115–14123, 2024.

Ma, Y., Zhu, X., and Hsu, J. Data poisoning against differentially-private learners: attacks and defenses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 4732–4738, 2019.

Ndiaye, E. Stable conformal prediction sets. In *International Conference on Machine Learning*, pp. 16462–16479. PMLR, 2022.

Neekhara, P., Dolhansky, B., Bitton, J., and Ferrer, C. C. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 923–932, 2021.

Peri, N., Gupta, N., Huang, W. R., Fowl, L., Zhu, C., Feizi, S., Goldstein, T., and Dickerson, J. P. Deep k-nn defense against clean-label data poisoning attacks. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 55–70. Springer, 2020.

Qian, W., Zhao, C., Le, W., Ma, M., and Huai, M. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1932–1942, 2023.

Qian, W., Zhao, C., Li, Y., Ma, F., Zhang, C., and Huai, M. Towards modeling uncertainties of self-explaining neural networks via conformal prediction. *arXiv preprint arXiv:2401.01549*, 2024.

Ren, Q., Deng, H., Chen, Y., Lou, S., and Zhang, Q. Bayesian neural networks avoid encoding complex and perturbation-sensitive concepts. In *International Conference on Machine Learning*, pp. 28889–28913. PMLR, 2023.

Romano, Y., Sesia, M., and Candès, E. J. Classification with valid and adaptive coverage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pp. 9389–9398. PMLR, 2021.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Shen, L., Ji, S., Zhang, X., Li, J., Chen, J., Shi, J., Fang, C., Yin, J., and Wang, T. Backdoor pre-trained models can transfer to all. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3141–3158, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017.

Stutz, D., Cemgil, A. T., Doucet, A., et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Suya, F., Mahloujifar, S., Suri, A., Evans, D., and Tian, Y. Model-targeted poisoning attacks with provable convergence. In *International Conference on Machine Learning*, pp. 10000–10010. PMLR, 2021.

Tao, L., Feng, L., Yi, J., Huang, S.-J., and Chen, S. Better safe than sorry: Preventing delusive adversaries with adversarial training. *Advances in Neural Information Processing Systems*, 34:16209–16225, 2021.

Teng, J., Wen, C., Zhang, D., Bengio, Y., Gao, Y., and Yuan, Y. Predictive inference with feature conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Trinh, T. Q., Heinonen, M., Acerbi, L., and Kaski, S. Tackling covariate shift with node-based bayesian neural networks. In *International Conference on Machine Learning*, pp. 21751–21775. PMLR, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, K.-C., Vicol, P., Lucas, J., Gu, L., Grosse, R., and Zemel, R. Adversarial distillation of bayesian neural network posteriors. In *International conference on machine learning*, pp. 5190–5199. PMLR, 2018.

Wang, X., He, X., Wang, J., and He, K. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16158–16167, 2021.

Wang, X., Li, Y., Xu, Z., and Luo, Y. Nested information representation of multi-dimensional decision: An improved promethee method based on npltss. *Information Sciences*, 607:1224–1244, 2022.

Weber, M., Xu, X., Karlaš, B., Zhang, C., and Li, B. Rab: Provable robustness against backdoor attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1311–1328. IEEE, 2023.

Wen, K., Ma, T., and Li, Z. How does sharpness-aware minimization minimize sharpness? *arXiv preprint arXiv:2211.05729*, 2022.

Wicker, M., Laurenti, L., Patane, A., and Kwiatkowska, M. Probabilistic safety for bayesian neural networks. In *Conference on uncertainty in artificial intelligence*, pp. 1198–1207. PMLR, 2020.

Yang, Y., Liu, T. Y., and Mirzasoleiman, B. Not all poisons are created equal: Robust training against data poisoning. In *International Conference on Machine Learning*, pp. 25154–25165. PMLR, 2022.

Yang, Z., He, X., Li, Z., Backes, M., Humbert, M., Berrang, P., and Zhang, Y. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pp. 39299–39313. PMLR, 2023.

Yuan, M., Wicker, M., and Laurenti, L. Gradient-free adversarial attacks for bayesian neural networks. *arXiv preprint arXiv:2012.12640*, 2020.

Zhao, C., Qian, W., Li, Y., Li, W., and Huai, M. Rethinking adversarial robustness in the context of the right to be forgotten. 2023.

Zhao, C., Qian, W., Ying, R., and Huai, M. Static and sequential malicious attacks in the context of selective forgetting. *Advances in Neural Information Processing Systems*, 36, 2024.