Non-omniscient backdoor injection with a single poison sample: Proving the one-poison hypothesis for linear regression and linear classification

Thorsten Peinemann[1], Paula Arnold[1], Sebastian Berndt[2], Thomas Eisenbarth,[1] and Esfandiar Mohammadi[1]

[1]University of Lübeck, Lübeck, Germany
[2]Technische Hochschule Lübeck, Lübeck, Germany
[1]{*t.peinemann, p.arnold, thomas.eisenbarth, esfandiar.mohammadi*}*@uni-luebeck.de*
[2]*sebastian.berndt@th-luebeck.de*

arXiv:2508.05600v1 [cs.LG] 7 Aug 2025

Abstract

Backdoor injection attacks are a threat to machine learning models that are trained on large data collected from untrusted sources; these attacks enable attackers to inject malicious behavior into the model that can be triggered by specially crafted inputs. Prior work has established bounds on the success of backdoor attacks and their impact on the benign learning task, however, an open question is what amount of poison data is needed for a successful backdoor attack. Typical attacks either use few samples, but need much information about the data points or need to poison many data points.

In this paper, we formulate the one-poison hypothesis: an adversary with one poison sample and limited background knowledge can inject a backdoor with zero backdooring-error and without significantly impacting the benign learning task performance. Moreover, we prove the one-poison hypothesis for linear regression and linear classification. For adversaries that utilize a direction that is unused by the benign data distribution for the poison sample, we show that the resulting model is functionally equivalent to a model where the poison was excluded from training. We build on prior work on statistical backdoor learning to show that in all other cases, the impact on the benign learning task is still limited. We also validate our theoretical results experimentally with realistic benchmark data sets.

1 Introduction

Machine learning on publicly available data is highly prone to data poisoning attacks that aim to inject backdoors into a model. Malicious parties could, e.g., randomly sprinkle poisoned data points on the Internet, hoping for a successful backdoor injection. Yet, data poisoning attacks are theoretically not sufficiently understood to assess whether such attacks will succeed, not even for linear models.

Prior work that provides theoretical bounds on backdoor injection attacks requires a significant fraction of the data to be poisoned in order to establish a bound on the accuracy of a backdoor attack. Poisoning such a large number of data points raises multiple problems for an attacker. One particular issue is the fact that such a large number of untypical data points can be easily detected and point to malicious behaviour, revealing the attack. In contrast, the modification of only few data points allows for plausible deniability, as these modification might, e.g., be due to measurement errors. Hence, Wang et al. [16] raise the question of "when will a backdoor attack succeed with a vanishing [fraction] ρ" of poisoned data points? Chaudhari et al. [2], Zhong et al. [19], Tan et al. [14] empirically illustrate that a one-poison attack is possible for specific tasks with rich data points. Their results, however, leave the question open for which other tasks such a one-poison attack can succeed, what the theoretical bounds for the success of the backdoor attack are, and how well the original, benign task is learned. For an omniscient attacker knowing *all* training data points, Hoang [6] provide initial bounds for a one-poison attack.

Hence, previous backdoor attacks were either able to perform a one-poison attack but needed exact knowledge about *all* data points or needed to poison a significant fraction of the data points, but needed less knowledge about the other data points. One might thus wonder whether this tradeoff between the number of poisoned data points and the needed knowledge is inherent. We study this open question from the literature that we pose as the *one-poison hypothesis*:

One-Poison Hypothesis. *Any machine learning model can be backdoored by a non-omniscient attacker with zero backdooring-error with a single poison sample with no significant harm to benign learning with probability almost 1.*

1.1 Main contributions

To understand this problem space, we confirm this hypothesis for linear regression and linear classification.

- We prove that with little knowledge of the training data, a single poisoned sample suffices to backdoor a

Symbol	Description
x	data point
n	size of training data
d	Dimensionality of data space, i.e. $\mathbb{R}^d$
η	poison pattern strength
$1-\varphi$	targeted prediction of poisoned sample (regression)
x_p	single poison sample
C	regularization parameter
$\mu_{\text{cl}}/\mu_{\text{bd}}/\mu_{\text{poi}}$	benign/backdoored/poisoned data distribution
$\hat{f}_{\text{cl}}/\hat{f}_{\text{poi}}$	classifier trained on benign/poisoned data
$\mathcal{L}^{\text{L2H}}/l_{\text{L2H}}$	regularized/per sample Hinge loss
$\mathcal{L}^{\text{SE}}/l_{\text{SE}}$	regularized/per sample squared error loss
$r_n^{\text{cl}}\left(\hat{f}_{\text{poi}}\right)$	statistical risk on benign data, expectation taken over training data
$r_n^{\text{poi}}\left(\hat{f}_{\text{poi}}\right)$	statistical risk on poisoned data, expectation taken over training data

Table 1: Notation Table

linear classification or linear regression model with no attack error with probability almost 1. In contrast to previous work, our bounds are proven and confirmed on real-world data.

- We prove that, if all samples from the benign data distribution have zero magnitude projected onto some direction, then both clean and poisoned model are functionally equivalent for any clean data sample, if the attacker chooses that direction for its single poison sample.
- We build on the prior work by Wang et al. [16] for classification and extend their work for regression to show that in all other cases, the impact of the poison sample on the benign learning task is still limited.
- We validate our theoretical results by evaluating them on realistic benchmark.

2 Overview

We provide an overview of our technique to confirm the one-poison hypothesis for linear classification and linear regression. Our proofs guarantee a successful backdoor attack and limited impact, in some cases even no impact, on the benign learning task. In our proof for guaranteeing a successful backdoor (cf. Theorem 2), we construct an attacker whose goal it is to leave a sufficiently large imprint on the classifier / regressor in any one direction. During inference stage, the adversary then triggers the backdoor behavior using a poison patch that activates the imprint and thus changes the prediction. Our proof for linear classification follows a similar idea as for linear regression. Hence, we only provide an intuition for linear regression.

Our main contribution is a one-poison attack for non-omniscient attackers.[1] We identify a simple sufficient property for injecting a backdoor through a single poisonous data point. For that we need the notion of the magnitude of a data set in a direction that is characterized by a vector $u \in \mathbb{R}^D$. The magnitude of a data set D in a direction u is the sum over each data point's length projected onto u, i.e., $\sum_{x\in D} \|\text{proj}(x,u)\|_2 = \sum_{x\in D} \|x^T u/\|u\|_2^2 \cdot u\|_2$. The sufficient condition that we identified requires that mean and variance of this magnitude for a single direction u are known to the adversary. We show that using this knowledge, an adversary can inject a backdoor with zero backdoor-error and if the magnitude is 0 (in mean and variance), the poison sample does not affect the benign learning task at all (cf. Theorem 3). We also show that if this magnitude is not 0, the impact on the benign learning task is limited (cf. Corollary 4).

In the simplest case of our proof for linear regression, assume that there is a direction $u \in \mathbb{R}^D$ such that the samples from the benign data distribution have zero projected magnitude onto u. The attacker constructs their single poison along the direction of u. They desire a specific property of the regressor, that the poison sample is predicted by the regressor with large enough value $1-\varphi$ ($\varphi \in (0,1)$). To obtain this property, the adversary enforces the loss's gradient to be non-zero, whenever this property is not satisfied. For training a linear regressor with regularized squared error loss, the gradient of the loss function for a classifier $T \in \mathbb{R}^D$ is the difference of the classifier T and a sum of training data points, weighted with a regularization factor C. Only the classifier disturbs the impact of the poison sample in the gradient as benign training data points have no magnitude projected onto u by assumption. We show that if the magnitude of the poison sample is now larger than $\sqrt{\varphi(1-\varphi)}/\varphi\sqrt{C}$ for regularization factor C, then it will outdo the regularization in the gradient.

During inference, the attacker patches the pattern with increased magnitude onto any test data point to change prediction of the data point. The increased magnitude of the poison patch increases the impact of the imprinted poison sample in the prediction of the regressor. This can arbitrarily change the prediction of the test data point when the poison patch is applied.

For the special case of zero magnitude of training data in direction u, we also show functional equivalence between clean and poisoned regressor. The high-level idea is that no benign data interferes with poisoned data in the learning phase and vice versa, because they are orthogonal. We show, that the task of optimizing the model can the be split into two independent tasks and their results are added together: One task is optimizing a benign model part and the other optimizing a poison model part. The poison model part is orthogonal to the benign model part, so predicting a benign test sample is not influenced by the benign model part, resulting in functional equivalence.

For bounding the impact of the poison sample on the benign learning task in general, we build on results on statistical poisoned learning for linear classification by Wang et al. [16] and extend their work to linear regression. These

[1] For an omniscient attacker, Hoang [6] has shown that a one-poison attack basically follows the same idea as in the impossibility of byzantine agreement [1].

proofs bound the discrepancy between poisoned model and optimal clean model by bounding first the discrepancy between poisoned model and optimal poisoned model and then the discrepancy between optimal poisoned model and optimal clean model. The first discrepancy will be expected to be small when a good training algorithm is deployed and the second discrepancy will also be expected to be small since there is only a single sample that differs between the poisoned task and the benign task.

2.1 Related work

Data poisoning backdoor attacks. Data poisoning backdoor attacks add maliciously crafted samples to the training data of a machine learning model to implant hidden triggers into the model, that when activated, leads to malicious behavior while the model continues to perform normally in the absence of such an active trigger. To the best of our knowledge the first work to explore data poisoning attacks is that of Gu et al. [5], which similar to our work crafts a poison sample with a target label that is specified by the adversary.

One poison sample backdoor. There are only few papers on backdoor injection specifically with one poison sample, to the best of our knowledge. Theoretical works on that topic assume that the adversary is omniscient, i.e., knows all training data, whereas we do not make such an assumption. Blanchard et al. [1] show, that a sum of elements can be arbitrarily changed by controlling a single element when all other elements are known. Hoang [6] apply the principle of Blanchard et al.'s work to linear and logistic regression, since the gradient of the loss functions for linear regression and logistic regression are both essentially a weighted sum of training data points. Hoang can make the gradient a $\mathbf{0}$-gradient for an attacker chosen regressor, so that this regressor is optimal for the optimization task. For achieving the $\mathbf{0}$-gradient Hoang use gradient inversion, i.e., generating a data point for a specific desired gradient, which is straightforward for linear and logistic regression. Tan et al. [14], Zhong et al. [19], Chaudhari et al. [2] empirically demonstrate that a single poison document / passage suffices to backdoor retriever models. Their evaluation shows attack success rates ranging from $28\% - 100\%$.

Theoretical understanding of backdoor attacks. Manoj and Blum [10] find that the excess memorization capacity of a model, i.e., the fact that a model can learn more tasks than the benign learning tasks, is necessary for a backdoor. They show that overparameterized linear models, i.e., linear models of dimension d where the input data from $\mathbb{R}^d$ resides in a smaller subspace of dimension $s < d$, can be poisoned with learning error on benign data of $\varepsilon_{\text{clean}}$ and backdoor data of ε_{adv}. The number of poison samples required is $\Omega(\varepsilon_{\text{adv}}^{-1}((d+1) + \log 1/\delta))$ where $d+1$ is the VC-dimension of linear classifiers and δ is the failure probability. This number can be a very large constant for small learning error.

Xian et al. [17] propose the 'adaptability hypothesis' to explain the success of a backdoor attack: A good backdoor attack should not change the predicted value too much before and after the backdoor attack. Their findings suggest that a good attack should have direction of low variance. For kernel smoothing algorithms and a specific benign data distributions that follows a multivariate normal distribution they prove that for $n \to \infty$, a large poison strength can counter a small poison ratio and yield no harm on the benign learning task and zero backdoor error.

Wang et al. [16] provide generalization bounds for benign task learning and backdoor learning for data poisoning. Their bound on the statistical risk of the poisoned classifier on data with backdoor trigger scales with $1/\rho \cdot r_{\text{poi}}$, ignoring further additive error terms, where r_{poi} is the statistical risk on input from training distribution. For $1/\rho = n$, i.e., a single poison sample, this bound can quickly become impractical due to the factor n on the statistical risk on training data. Wang et al. further show for benign data distributions that follow a multivariate normal, the direction of smallest variance is most successful for a backdoor attack. For the case of distributions with a direction where the distribution is a point, they show that any backdoor attack will be successful but only asymptotically. A backdoor attack is considered successful by Wang et al. if for dataset size n, one has $\limsup_{n\to\infty} r_{\text{bd}}/r_{\text{cl}} \leq C$, where r_{bd} is the statistical risk of a poisoned classifier on data with backdoor trigger and r_{cl} is the statistical risk of a clean model on benign data. Consequently, a very large constant C might fulfill this definition, while the risk on data with backdoor trigger might not be meaningfully bounded.

Li and Liu [8] provide bounds on benign learning task accuracy and backdoor attack success rate of backdoor attacks on overparameterized CNNs assuming data is constructed from random patches that follow a multivariate normal distribution for a poison ratio that is sub-linear in the training dataset size n.

Yu et al. [18] propose a generalization bound for benign learning and backdoor learning for neural networks and also more general hypothesis spaces. Their bound on the statistical risk of the poisoned model on data with backdoor patch roughly scales with L/ρ, where L is the cross-entropy loss on poisoned training data. For $\rho = 1/n$, i.e., a single poison sample, this bound can also quickly become impractical due to the factor n on the cross-entropy loss.

3 Preliminaries

Linear regression. We train a linear regressor minimizing the regularized squared error loss $\mathcal{L}_{\text{sq}}$ with regularization parameter C:

$$\min_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{sq}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \min_{\hat{f}_{\text{cl}}} \frac{1}{2}\|\hat{f}_{\text{cl}}\|_2^2 + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} (y_i - \hat{f}_{\text{cl}}^T x_i)^2.$$

The gradient of regularized squared error loss is

$$\nabla_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{sq}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \hat{f}_{\text{cl}} + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} (\hat{f}_{\text{cl}}^T x_i - y_i) x_i.$$

We can predict test samples x in inference stage: $\hat{f}_{\text{cl}}^T x$. The following lemma will be useful throughout the paper.

Lemma 1. *Let a and b be two orthogonal vectors. Then $\|a+b\|_2^2 = \|a\|_2^2 + \|b\|_2^2$.*

Proof. We have $\|a+b\|_2^2 = (a+b)^T(a+b) = a^Ta+2a^Tb+b^Tb = \|a\|_2^2 + 2a^Tb + \|b\|_2^2$. As a and b are orthogonal, $a^Tb = 0$, which concludes the proof. □

Linear classification. We train a linear classifier minimizing the regularized Hinge loss $\mathcal{L}_{\text{Hinge}}$ with regularization parameter C (training stage):

$$\begin{aligned}&\min_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{Hinge}}(D_{\text{cl}}, \hat{f}_{\text{cl}})\\ =&\min_{\hat{f}_{\text{cl}}} \frac{1}{2}\|\hat{f}_{\text{cl}}\|_2^2 + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} \max\left(0, 1 - y_i \hat{f}_{\text{cl}}^T x_i\right).\end{aligned} \tag{1}$$

The gradient of regularized Hinge loss is

$$\nabla_{\hat{f}_{\text{cl}}} \mathcal{L}_{\text{Hinge}}(D_{\text{cl}}, \hat{f}_{\text{cl}}) = \hat{f}_{\text{cl}} - C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{cl}}^T x_i y_i < 1} x_i y_i.$$

We can predict test samples x in inference stage: $\hat{f}_{\text{cl}}^T x$. In this work, we also consider the Lagrangian dual of the linear classification problem of Equation (1). For obtaining the Lagrangian dual, we first formulate the primal optimization problem which is equivalent to Equation (1):

$$\begin{aligned}&\min_{w,\xi} \frac{1}{2}\|w\|_2^2 + C \cdot \sum_{i=1}^n \xi_i \\ &\text{s.t.} -\xi_i \leq 0 \text{ and } (1 - y_i w^T x_i) - \xi_i \leq 0 \quad \text{for } i = 1, ..., n.\end{aligned} \tag{2}$$

We can then construct the Lagrangian of Equation (2):

$$\begin{aligned}\mathcal{L}(w, \xi, \alpha, r) = &\frac{1}{2}\|w\|_2^2 + C \cdot \sum_{i=1}^n \xi_i \\ &- \sum_{i=1}^n \alpha_i \left(y_i w^T x_i - 1 + \xi_i\right) - \sum_{i=1}^n r_i \xi_i,\end{aligned} \tag{3}$$

where the α_i and r_i are Lagrange multipliers constrained to being ≥ 0. We formulate the primal of Equation (2) equivalently as

$$\min_{w,\xi} \max_{\alpha,r} \mathcal{L}(w, \xi, \alpha, r). \tag{4}$$

Then, the dual of the problem is defined as

$$\max_{\alpha,r} \min_{w,\xi} \mathcal{L}(w, \xi, \alpha, r) \ . \tag{5}$$

For solving the dual, we first set the derivative of $\mathcal{L}$ with respect to w to zero. From the derivative, we obtain that for any α, r, the minimizing w is $w = \sum_{i=1}^n \alpha_i x_i y_i$. Inserting into the Lagrangian of Equation (3) yields the following form of the dual optimization problem of Equation (5):

$$\begin{aligned}\max_{(\alpha)_{i=1}^n} \quad &\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad &0 \leq \alpha_i \leq C \quad \text{for } i = 1, ..., n.\end{aligned} \tag{6}$$

Threat model. The adversary provides a single maliciously crafted sample to the training data set to plant a backdoor into the trained model. The backdoor should have no backdoor error, i.e., 100% of samples with injected backdoor patch get classified as positive (classification) or, respectively, get labeled with +1 (regression). The single poison sample should also not degrade learning the benign task. The adversary does not need to know the exact training data or even the exact distribution of the benign data. They need, however, an estimate on mean and variance of the benign data when projected onto some direction $u \in \mathbb{R}^D$ that the adversary uses to launch their attack. To this end, we define a random variable $s_m^{(u)} = \sum_{i=1}^m \|x_i^T u/\|u\|_2^2 \cdot u\|_2$ that for m data points $x_i \sim \mu_{\text{cl}}$ $(i = 1, 2, ..., m)$ from benign data distribution μ_{cl}, sums up the lengths when those data points are projected onto a direction $u \in \mathbb{R}^d$, and let $\mu_{\text{signal}}, \sigma_{\text{signal}}^2$ be mean and variance of $s_m^{(R)}$. We assume the size n of the benign training data to be fixed and known to the adversary.

4 One-poison hypothesis for linear regression

In this section, we show validity of the one-poison hypothesis for linear regression. Our proof for linear classification follows a similar idea. We briefly discuss linear classification in Section 4.3, for the complete proofs, we refer to the appendix.

4.1 Single poison sample backdoor for linear regression

Our first result shows that a single poison sample suffices to backdoor a linear regressor with zero backdoor error with probability almost 1:

Theorem 2. *Let $0 < \delta, \varphi < 1$. Let the size n of the benign data D_{cl} be fixed. Let $R \in \mathbb{R}^{d\times d} : \left(\det(R) = 1 \wedge R^TR = I_d\right)$, then define the poison sample $x_p = R \cdot (0, 0, ..., 0, \eta)$. Let $s_m^{(u)}$ be defined as in the threat model (cf. Section 3) with mean and variance μ_{signal}, σ_{signal}^2. Assume that there is some bound $[-K, K]$ $(K \in \mathbb{R}^+)$ on regression labels and regressor output. With poison pattern strength $\eta > \frac{1}{2\varphi}(2K(\mu_{signal} + \sigma_{signal}^2/\sqrt{\delta}) + \sqrt{4(\mu_{signal} + \sigma_{signal}^2/\sqrt{\delta})^2 + 4K^2\varphi(1-\varphi)/C})$ and patch function $\text{patch}(x) = x + R \cdot (0, 0, ..., 0, |(-\hat{f}_{poi}^T x + 1)/(\hat{f}_{poi}^T x_p) \cdot \eta|)$, the adversary achieves with probability $1 - \delta$ that $\forall x \in \Omega_{|y=-1} \colon \hat{f}_{poi}^T \text{patch}(x) = 1$.*

Proof. We define the gradient of squared error loss w.r.t. poisoned model $\hat{f}_{\text{poi}}$ as $g := \nabla L(\hat{f}_{\text{poi}})$, which is equal to

$$C \cdot \sum_{x_i \in D_{\text{cl}}} (\hat{f}_{\text{poi}}^T x_i - y_i) x_i + C \cdot (\hat{f}_{\text{poi}}^T x_p - 1) x_p + \hat{f}_{\text{poi}}.$$

To exploit the backdoor during inference stage, the adversary desires the property $\hat{f}_{\text{poi}}^T x \geq 1$. The adversary constructs a single poison sample $x_p \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, so that for

any regressor that does not satisfy this property, its gradient g will have $g^T x_p < 0$ almost certainly which implies $g \neq \mathbf{0}$, i.e., that regressor is not optimal. We investigate the probability of the adversary being successful:

$$\begin{aligned}
&\Pr[g^T x_p < 0] = \Pr[\big(C \cdot \sum_{x_i \in D_{\text{cl}}} (\hat{f}_{\text{poi}}^T x_i - y_i) x_i \\
&+ C \cdot \underbrace{(\hat{f}_{\text{poi}}^T x_p - 1)}_{<0 \text{ since } \hat{f}_{\text{poi}}^T x_p < 1} x_p + \hat{f}_{\text{poi}}\big)^T x_p < 0] \\
&= \Pr[C \cdot \sum_{x_i \in D_{\text{cl}}} x_p^T (\hat{f}_{\text{poi}}^T x_i - y_i) x_i \\
&+ C \cdot x_p^T (\hat{f}_{\text{poi}}^T x_p - 1) x_p + \hat{f}_{\text{poi}}^T x_p < 0] \\
&= \Pr[C \cdot \sum_{x_i \in D_{\text{cl}}} (R \cdot (0, ..., 0, \eta))^T (\hat{f}_{\text{poi}}^T x_i - y_i)(R \cdot x_i^{(R)}) \\
&+ C \cdot x_p^T x_p (\hat{f}_{\text{poi}}^T x_p - 1) + \hat{f}_{\text{poi}}^T x_p < 0]
\end{aligned}$$

Here, $x_i^{(R)} := R^{-1} x_i$ and $x_{i,d}^{(R)}$ are the d'th component.

$$\begin{aligned}
&= \Pr[C \cdot \eta \cdot \underbrace{\sum_{x_i \in D_{\text{cl}}} x_{i,d}^{(R)} (\hat{f}_{\text{poi}}^T x_i - y_i)}_{\mathbb{L}^{\text{SE}} :=} \\
&+ C \cdot \underbrace{x_p^T x_p}_{=\eta^2} \underbrace{(\hat{f}_{\text{poi}}^T x_p - 1)}_{<-\varphi} + \underbrace{\hat{f}_{\text{poi}}^T x_p}_{<1-\varphi} < 0] \\
&\geq \Pr[C \cdot \mathbb{L}^{\text{SE}} \cdot \eta - C \cdot \eta^2 \cdot \varphi + 1 - \varphi < 0]. \qquad (7)
\end{aligned}$$

The adversary aims to bound $\mathbb{L}^{\text{SE}}$ to obtain a magnitude of positive η that guarantees $g^T x_p < 0$ with high probability. To this end, for n data points in training data, the adversary establishes a bound on the sum of the magnitudes of benign training data points projected onto x_p, and uses μ_{signal} and σ_{signal}^2 and Chebyshev's inequality [3]: From

$$\Pr[|s_{|D_{\text{cl}}|}^{(u)} - \mu_{\text{signal}}| \geq k] \leq \sigma_{\text{signal}}^2 / k^2 \stackrel{!}{=} \delta$$

it follows that $k = \sigma_{\text{signal}}^2 / \sqrt{\delta}$. The adversary using the assumption of bounded regressor output and bounded regression label on $[-K, K]$ bounds $|(\hat{f}_{\text{poi}}^T x_i - y_i)| \leq 2 \cdot K$ inside $\mathbb{L}_{\text{SE}}$, then sets $\hat{s} := 2 \cdot K \cdot (\mu_{\text{signal}} + k)$ and obtains $\Pr[\mathbb{L}^{\text{SE}} < \hat{s}] \geq 1 - \delta$. Solving the quadratic inequality of Equation (7) yields the solution $\eta^* > \frac{\hat{s} + \sqrt{\hat{s}^2 + \frac{4\varphi(1-\varphi)}{C}}}{2\varphi}$ and

$$\Pr[C \cdot \mathbb{L}^{\text{SE}} \cdot \eta - C \cdot \eta^2 \cdot \varphi + 1 - \varphi < 0] \geq 1 - \delta.$$

Consequently, when the poisoned regressor attains optimum, i.e., its gradient g satisfies $g = \mathbf{0}$, it has to hold that $\hat{f}_{\text{poi}}^T x_p \geq 1 - \varphi$ or else with high probability $g \neq \mathbf{0}$. Now the backdoor can be triggered via the patch function for every sample $x \in X$ and $\hat{f}_{\text{poi}}^T \text{patch}(x)$ thus equals

$$\begin{aligned}
&= \hat{f}_{\text{poi}}^T \big(x + R \cdot (0, 0, ..., 0, |(-\hat{f}_{\text{poi}}^T x + 1)/(\hat{f}_{\text{poi}}^T x_p) \cdot \eta|)\big) \\
&= \hat{f}_{\text{poi}}^T x + |(-\hat{f}_{\text{poi}}^T x + 1)/(\hat{f}_{\text{poi}}^T x_p)| \cdot \hat{f}_{\text{poi}}^T \\
&\underbrace{\big(R \cdot (0, 0, ..., 0, |\eta|)\big)}_{=x_p} \\
&= \hat{f}_{\text{poi}}^T x + \underbrace{|(-\hat{f}_{\text{poi}}^T x + 1)/(\hat{f}_{\text{poi}}^T x_p)| \cdot |\hat{f}_{\text{poi}}^T x_p|}_{=-\hat{f}_{\text{poi}}^T x + 1} = 1. \qquad \square
\end{aligned}$$

The prior theorem assumes a bound $[-K, K]$ on regressor output and regression label, but note that this is merely a technical assumption because otherwise the impact of benign data could be arbitrarily large. In reality, regressor output and regression label will certainly always have some upper and lower bound. While the adversary can pick any rotation R for the poison sample x_p, when R is selected such that all samples from benign data distribution projected onto the poison are zero-centered and have zero magnitude, then $\mu_{\text{signal}} = 0$ and $\sigma_{\text{signal}}^2 = 0$. In this special case, the adversary does not need to bound the impact of benign training data and their success probability is exactly 1. The bound on η simplifies to $\eta > \sqrt{\varphi(1-\varphi)}/(\varphi\sqrt{C})$ since all terms with $\mu_{\text{signal}}, \sigma_{\text{signal}}^2$ can be dropped.

4.2 Impact on benign learning task

We now analyze the impact of the single poison sample on the benign learning task. First, we show that if all samples from benign data distribution projected onto the poison sample are zero-centered and have zero magnitude, then the backdoor attack described in Theorem 2 does not impact the benign learning task at all.

Theorem 3. *Assume that $\mu_{signal} = 0$ and $\sigma_{signal}^2 = 0$, i.e., all samples from benign data distribution projected onto the poison are zero-centered and have zero magnitude. Then for all $x \in X$, an optimal regressor $\hat{f}_{poi}$ is functionally equivalent to an optimal $\hat{f}_{cl}$ that is obtained when the single poison x_p is omitted in training.*

Proof. Let x_p be defined as in Theorem 2. From the definition, it follows that $\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{sq}}(D_{\text{poi}}, \hat{f}_{\text{poi}})$ is equal to

$$\begin{aligned}
&= \min_{\hat{f}_{\text{poi}}} \frac{1}{2} \|\hat{f}_{\text{poi}}\|_2^2 + C \cdot \sum_{(x_i, y_i) \in D_{\text{cl}}} (y_i - \hat{f}_{\text{cl}}^T x_i)^2 \\
&\quad + C \cdot (1 - \hat{f}_{\text{cl}}^T x_p)^2 \ .
\end{aligned}$$

Now, we split the classifier $\hat{f}_{\text{poi}}$ into two parts: the part of $\hat{f}_{\text{poi}}$ projected onto x_p, i.e., $\text{proj}_{x_p}(\hat{f}_{\text{poi}}) := \hat{f}_{\text{poi}}^T x_p / \|x_p\|^2 \cdot x_p$, and the remainder of $\hat{f}_{\text{poi}}$, i.e., $\text{rem}_{x_p}(\hat{f}_{\text{poi}}) := \hat{f}_{\text{poi}} - \text{proj}_{x_p}(\hat{f}_{\text{poi}})$. We can thus write $\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{sq}}(D_{\text{poi}}, \hat{f}_{\text{poi}})$ as

$$\begin{aligned}
&= \min_{\hat{f}_{\text{poi}}} \frac{1}{2} \|\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \\
&\quad + C \cdot \sum_{(x_i, y_i) \in D_{\text{cl}}} (y_i - (\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}}))^T x_i)^2 \\
&\quad + C \cdot (1 - (\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}}))^T x_p)^2
\end{aligned}$$

By assumption, $\mu_{\text{signal}} = \sigma^2_{\text{signal}} = 0$ and thus

$$
\begin{aligned}
&= \min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \qquad (*)\\
&\quad + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} (y_i - (\underbrace{\text{proj}_{x_p}(\hat{f}_{\text{poi}})^T x_i}_{=0} +\\
&\quad \text{rem}_{x_p}(\hat{f}_{\text{poi}})^T x_i))^2\\
&\quad + C \cdot (1 - (\text{proj}_{x_p}(\hat{f}_{\text{poi}})^T x_p + \underbrace{\text{rem}_{x_p}(\hat{f}_{\text{poi}})^T x_p}_{=0}))^2\\
&= \min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\text{proj}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 + \frac{1}{2}\|\text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2\\
&\quad + C \sum_{(x_i,y_i)\in D_{\text{cl}}} (y_i - \text{rem}_{x_p}(\hat{f}_{\text{poi}})^T x_i)^2\\
&\quad + C \cdot (1 - \text{proj}_{x_p}(\hat{f}_{\text{poi}})^T x_p)^2\\
&= \min_{\text{rem}_{x_p}(\hat{f}_{\text{poi}})} \mathcal{L}_{\text{sq}}(\text{rem}_{x_p}(\hat{f}_{\text{poi}}), D_{\text{cl}})\\
&\quad + \min_{\text{proj}_{x_p}(\hat{f}_{\text{poi}})} \mathcal{L}_{\text{sq}}(\text{proj}_{x_p}(\hat{f}_{\text{poi}}), \{x_p\})
\end{aligned}
$$

Here, we used that $\text{proj}_{x_p}(\hat{f}_{\text{poi}})$ and $\text{rem}_{x_p}(\hat{f}_{\text{poi}})$ are orthogonal. Lemma 1 thus implies $\|\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 = \|\text{proj}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 + \|\text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2$. Using this equality in (*) gives us the term completely separating the projection from the remainder.

To deduce the functional equivalence, we need to show that minimizing $\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*$, which is defined as $\arg\min_{\text{rem}_{x_p}(\hat{f}_{\text{poi}})} \mathcal{L}_{\text{sq}}(\text{rem}_{x_p}(\hat{f}_{\text{poi}}), D_{\text{cl}})$, is the same as minimizing $f^* = \arg\min_f \mathcal{L}_{\text{sq}}(f, D_{\text{cl}})$. Intuitively, this is true, as any vector of the form $\alpha \cdot x_p$ ($\alpha \in \mathbb{R} \setminus \{0\}$) is orthogonal to $\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*$. Adding $\alpha \cdot x_p$ to $\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*$ does not change the prediction of any benign training data point $(x, y) \in \Omega$ and can thus not reduce the data point's loss. More formally, let l_{SE} be a data point's squared error loss, then

$$
\begin{aligned}
&l_{\text{SE}}(\text{rem}_{x_p}(\hat{f}_{\text{poi}})^* + \alpha x_p, x)\\
=&(y - (\text{rem}_{x_p}(\hat{f}_{\text{poi}})^* + \alpha x_p)^T x)^2\\
=&(y - ((\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*)^T x + \alpha x_p^T x))^2\\
=&l_{\text{SE}}(\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*, x).
\end{aligned}
$$

Due to the orthogonality, we can again use Lemma 1 to show that the addition of $\alpha \cdot x_p$ only increases the norm of the classifier, as $\|\text{rem}_{x_p}(\hat{f}_{\text{poi}})^* + \alpha x_p\|_2^2 = \|\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*\|_2^2 + \|\alpha x_p\|_2^2 > \|\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*\|_2^2$.

We thus obtain the functional equivalence for $x \in X$ by using $(\min_{\text{proj}_{x_p}(\hat{f}_{\text{poi}})} \mathcal{L}_{\text{sq}}(\text{proj}_{x_p}(\hat{f}_{\text{poi}}), \{x_p\}))^T x = 0$, as

$$
(\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{sq}}(D_{\text{poi}}, \hat{f}_{\text{poi}}))^T x = (\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{sq}}(D_{\text{cl}}, \hat{f}_{\text{poi}}))^T x \ .
$$

□

Now we move on to the most general case where the benign data distribution can be any distribution. For this, we extend prior work [16] to regression:

Corollary 4. *Corollary of [16, Theorem 1] Let μ_{cl} be the benign data distribution. We define the backdoor and poisoned distributions as*

$$
\mu_{bd}(x) = \mathbb{1}\{x = x_p\},\ \mu_{poi}(x) = (1 - 1/n)\mu_{cl} + 1/n\mu_{bd}.
$$

Let n be fixed. Let $\hat{f}_{poi}$ be the regressor trained on n samples from the poisoned distribution μ_{poi}. Let $l(.,.) : [0,1] \times [0,1] \mapsto \mathbb{R}^+$ be a general loss function that is (C, α)-Hölder continuous for $0 < \alpha \leq 1$ that measures the discrepancy between two regressors. The statistical risk on benign input is bounded as

$$
r_n^{cl}(\hat{f}_{poi}) \leq \frac{1}{1 - 1/n} r_n^{poi}(\hat{f}_{poi}) + C(\mu_{cl}(x_p) \cdot 1/n \cdot 2)^{\alpha} \ .
$$

Proof. First, since l is (C, α)-Hölder continuous:

$$
\begin{aligned}
&r_n^{\text{cl}}(\hat{f}_{\text{poi}}) = \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{cl}}(x))\right]\right]\\
&\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{poi}}(x)) + C\,|f^*_{\text{poi}}(x) - f^*_{\text{cl}}(x)|^{\alpha}\right]\right]\\
&\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{poi}}(x))\right]\right]\\
&+ C \mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[|f^*_{\text{poi}}(x) - f^*_{\text{cl}}(x)|^{\alpha}\right].
\end{aligned}
$$

Using Jensen's inequality, this is at most

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{poi}}(x))\right]\right]\\
&+ C\left(\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[|f^*_{\text{poi}}(x) - f^*_{\text{cl}}(x)|\right]\right)^{\alpha} \ .
\end{aligned}
$$

We bound each term on the right-hand side independently. For the first term, we can bound

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{poi}}(x))\right]\right]\\
&\leq (1 - 1/n)^{-1} \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x\sim\mu_{\text{poi}}}\left[l(\hat{f}_{\text{poi}}(x), f^*_{\text{poi}}(x))\right]\right]\\
&\leq (1 - 1/n)^{-1} r_n^{\text{poi}}(\hat{f}_{\text{poi}}).
\end{aligned}
$$

As for the second term, by definition of the Bayes optimal regressor for squared error loss [20], we have $f^*_{\text{cl}}(x) = \mathop{\mathbb{E}}_{\mu_{\text{cl}}}[Y|X = x]$ and, similarly,

$$
f^*_{\text{poi}}(x) = \mathop{\mathbb{E}}_{\mu_{\text{poi}}}[Y|X = x] =
$$

$$
\begin{cases}
\mathop{\mathbb{E}}_{\mu_{\text{cl}}}[Y|X = x] & \text{if } x \neq x_p,\\
(1 - 1/n)\mathop{\mathbb{E}}_{\mu_{\text{cl}}}[Y|X = x_p] + 1/n \mathop{\mathbb{E}}_{\mu_{\text{bd}}}[Y|X = x_p] & \text{if } x = x_p.
\end{cases}
$$

Combining prior equations, we get

$$
f^*_{\text{poi}} - f^*_{\text{cl}} = \begin{cases}
0 & \text{if } x \neq x_p,\\
1/n \cdot \mathop{\mathbb{E}}_{\mu_{\text{bd}}}[Y|X = x_p] & \\
-1/n \mathop{\mathbb{E}}_{\mu_{\text{cl}}}[Y|X = x_p] & \text{if } x = x_p.
\end{cases}
$$

Accordingly,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[|f^*_{\text{poi}}(x) - f^*_{\text{cl}}(x)|\right] \\
&= \Pr_{\mu_{\text{cl}}}[X = x_p] \cdot |1/n \mathop{\mathbb{E}}_{\mu_{\text{bd}}}[Y|X = x_p] \\
&- \mathop{\mathbb{E}}_{\mu_{\text{cl}}}[Y|X = x_p]| + (1 - \Pr_{\mu_{\text{cl}}}(X = x_p) \cdot 0 \\
&= \mu_{\text{cl}}(x_p)1/n| \mathop{\mathbb{E}}_{\mu_{\text{bd}}}[Y|X = x_p] - \mathop{\mathbb{E}}_{\mu_{\text{cl}}}[Y|X = x_p]| \\
&\leq \mu_{\text{cl}}(x_p) \cdot 1/n \cdot 2.
\end{aligned}$$

Now, we plug the two bounds together:

$$r_n^{\text{cl}}\big(\hat{f}_{\text{poi}}\big) \leq \frac{1}{1 - 1/n} r_n^{\text{poi}}\big(\hat{f}_{\text{poi}}\big) + C(\mu_{\text{cl}}(x_p) \cdot 1/n \cdot 2)^{\alpha}.$$

□

The attentive reader might be wondering why we consider i.i.d. sampling of the data set that is different to our prior assumption that the poison sample is guaranteed to be added to training data. Slightly increasing the poison ratio in the proof from $1/n$ to k/n for a small constant k gives the probability of sampling at least one poison sample $1 - (1 - k/n)^n$. Under large enough data size n, this turns to $1 - (1 - k/n)^n \underset{n\to\infty}{\to} 1 - e^{-k}$ which quickly goes to 1 when increasing k. The bound on the statistical risk on benign data for k poison samples is $r_n^{\text{cl}}\big(\hat{f}_{\text{poi}}\big) \leq \frac{1}{1-k/n} r_n^{\text{poi}}\big(\hat{f}_{\text{poi}}\big) + C(\mu_{\text{cl}}(x_p) \cdot k/n)^{\alpha}$, so the bound is only slightly higher.

4.3 One-poison hypothesis for linear classification

We extend all of our results stated here to the case of linear classification using regularized Hinge loss, both in primal optimization and dual optimization. The attack works mostly similar, but there are a few differences. The loss function is different than regularized squared error loss: In the gradient w.r.t. classifier f each data point x with label y is weighted by either 1 if $f^T xy < 1$ and 0 otherwise. Consequently, the poison sample x_p's impact on gradient is not scaled down when $f^T x_p$ approaches 1, so the adversary does not need to account for this. Also, during inference stage, the adversary must construct a poison patch for test data point x that changes prediction from $f^T x < 0$ to $f^T \text{patch}(x) \geq 0$ so there is no need to aim for a specific predicted value. Linear classification is often performed using dual optimization, for instance using the popular `LIBLINEAR` library [4]. We show using the strong duality property of linear classification, i.e., that maximum objective value of primal and dual are equal, that our results for primal optimization also follow for dual optimization. For full proofs, we refer to the appendix.

5 Experiments

5.1 Experimental setup

All experiments were run on an Intel Xeon Platinum 8168 2.7 GHz CPU with 32GB of RAM.

Model. We validate our developed theory for linear regression and linear classification using the scikit-learn Python library [13], trained for 1,000 iterations with regularization $C = 1$. For our functional equivalence ablation for linear classification we use a vanilla implementation of liblinear [4] instead of the one of scikit-learn to remove any source of randomness.

Data sets. We evaluate regression data sets Parkinsons [15] and Abalone [12], and classification data sets Spambase [7] and Phishing [11], each partitioned half and half in training and test data. The data sets demand deducing motor impairment of patients from biomedical voice measurements, predicting the age of abalone from physical measurements, detecting spam mail from word counts, and detecting phishing websites from website meta data. We choose these data sets as they represent realistic data sets of real-world measurements.

Baselines. We train a clean model only on benign data. We also evaluate a mean regressor producing the mean regression label from training data for linear regression, and a majority-vote classifier outputting the majority label from training data for linear classification. For linear regression, we report the mean squared error, for linear classification, we report accuracy. Both metrics correspond to their learning task, i.e., reducing squared error (linear regression) and improving accuracy (linear classification). We report mean and standard deviation over five runs of random data splits.

Attack. We set the poison label to 1 and the magnitude of poison patch $\eta = 1$ for Parkinsons and Abalone, $\eta = 10$ for Spambase and Phishing. For poison direction, we compute a principal component analysis on each data set and extract an eigenvector in direction with smallest variance. This variance is $6.5e^{-12}$ for Parkinsons, 0.0001 for Abalone, 0.0004 for Spambase and 0.03 for Phishing. A single poison sample is added to the training data set.

5.2 Results

Single poison sample backdoor. Table 2 and Table 3 show that a single poison sample suffices to inject a backdoor in linear regressors and classifiers. All test samples with poison patches are predicted with the poison label instead of the correct label. The single poison sample does not significantly increase prediction error, showing bounded impact on the benign learning task.

Ablation: Functional equivalence of clean model and poisoned model with 0-dimension in data. To validate our theoretic results, we add a dimension to the benign data and set each benign sample's value in that dimension to 0, then calculate for 100 random test samples the L1-distance of predictions between clean model and poisoned model. We clip the predictions to two decimal places. In all settings, the L1 distance is 0, showing functional equivalence of clean model and poisoned model.

6 Conclusion

In this paper, we proved the one-poison hypothesis for both linear regression and linear classification. Our attack shows

Parkinsons		
Regressor	Benign Task MSE	Backdoor MSE
Mean regr.	0.202 ± 0.002	
Clean regr.	0.165 ± 0.002	3.852 ± 1.954
Poisoned regr.	0.166 ± 0.002	0.000 ± 0.000
Abalone		
Regressor	Benign Task MSE	Backdoor MSE
Mean regr.	0.054 ± 0.001	
Clean regr.	0.033 ± 0.001	7.391 ± 0.043
Poisoned regr.	0.034 ± 0.001	0.000 ± 0.000

Table 2: Clean and poisoned regressor test MSE of benign and backdoor task. Single poison added to training data. Mean regressor outputs mean of labels from training data.

Spambase		
Classifier	Benign Task (%)	Backdoor Task (%)
Majority vote	60.00 ± 0.04	
Clean	82.89 ± 0.05	18.33 ± 36.65
Poisoned	81.81 ± 0.04	100.00 ± 0.00
Phishing		
Classifier	Benign Task (%)	Backdoor Task (%)
Majority vote	56.44 ± 0.01	
Clean	92.47 ± 0.01	0.02 ± 0.05
Poisoned	92.37 ± 0.01	100.00 ± 0.00

Table 3: Clean and poisoned classifier test accuracy of benign and backdoor task. Single poison added to training data. Majority vote predicts majority label from training data.

that such models can be successfully attacked by poisoning a single data point with limited knowledge about the other data points. Our bounds are formally proven, apply to real-world instance sizes, and are verified experimentally also.

While typical countermeasures such as differential privacy look very promising [9], they also come with a large performance or accuracy penalty. Promising future directions are thus the development of efficient countermeasures and the transfer of our results to more complex models.

References

[1] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.

[2] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*, 2024.

[3] Pafnutii Lvovich Chebyshev. Des valeurs moyennes. *Journal de Mathématique Pures et Appliquées*, 12(2): 177–184, 1867.

[4] 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.

[5] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

[6] Lê-Nguyên Hoang. The poison of dimensionality. *arXiv preprint arXiv:2409.17328*, 2024.

[7] Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase. UCI Machine Learning Repository, 1999. URL https://archive.ics.uci.edu/dataset/94/spambase.

[8] Boqi Li and Weiwei Liu. A theoretical analysis of backdoor poisoning attacks in convolutional neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[9] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: attacks and defenses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 4732–4738. AAAI Press, 2019. ISBN 9780999241141.

[10] Naren Manoj and Avrim Blum. Excess capacity and backdoor poisoning. *Advances in Neural Information Processing Systems*, 34:20373–20384, 2021.

[11] Rami Mustafa A. Mohammad, Fadi A. Thabtah, and Lee McCluskey. An assessment of features related to phishing websites using an automated technique. *2012 International Conference for Internet Technology and Secured Transactions*, pages 492–497, 2012. URL https://api.semanticscholar.org/CorpusID:5716727.

[12] Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1995. URL https://archive.ics.uci.edu/dataset/1/abalone.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, and Huan Liu. Glue pizza and eat rocks - exploiting vulnerabilities in retrieval-augmented generative models. In *EMNLP*, pages 1610–1626. Association for Computational Linguistics, 2024.

[15] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010. doi: 10.1109/TBME.2009.2036000.

[16] Ganghua Wang, Xun Xian, Jayanth Srinivasa, Xuan Bi Ashish Kundu, Mingyi Hong, , and Jie Ding. Demystifying Poisoning Backdoor Attacks from a Statistical Perspective. In *International Conference on Learning Representations (ICLR)*, 2024.

[17] Xun Xian, Ganghua Wang, Jayanth Srinivasa, Ashish Kundu, Xuan Bi, Mingyi Hong, and Jie Ding. Understanding backdoor attacks through the adaptability hypothesis. In *International Conference on Machine Learning*, pages 37952–37976. PMLR, 2023.

[18] Lijia Yu, Shuang Liu, Yibo Miao, Xiao-Shan Gao, and Lijun Zhang. Generalization bound and new algorithm for clean-label backdoor attack. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57559–57596. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24i.html.

[19] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *EMNLP*, pages 13764–13775. Association for Computational Linguistics, 2023.

[20] Eric R Ziegel. The elements of statistical learning, 2003.

A One-poison hypothesis for linear classification

In this section, we answer the one-poison hypothesis for linear classification: We first show that a single poison sample suffices to backdoor a linear classifier with zero backdoor error with probability almost 1. We extend this result to the case where the learner uses dual optimization to fit their classifier. We then show that the poison sample does not impact the benign learning task at all when all data from the benign data distribution projected onto the poison sample is zero-centered and has zero magnitude. For all other cases, we show that the impact is still limited.

A.1 Single poison sample backdoor for linear classification

Our first result shows that a single poison sample suffices to backdoor a linear classifier with zero backdoor error with probability almost 1:

Theorem 5. *Let $0 < \delta < 1$. Let the size n of the benign data D_{cl} be fixed. Let $R \in \mathbb{R}^{d\times d}$: $\big(\det(R) = 1 \wedge R^T R = I_d\big)$, then define the poison sample $x_p = R \cdot (0, 0, ..., 0, \eta)$. Let $s_m^{(u)}$ be defined as in the threat model (cf. Section 3) with mean and variance $\mu_{signal}, \sigma^2_{signal}$. With poison pattern strength $\eta > \frac{1}{2}(\mu_{signal} + \sigma^2_{signal}/\sqrt{\delta} + \sqrt{(\mu_{signal} + \sigma^2_{signal}/\sqrt{\delta})^2 + 4/C})$ and patch function $\mathrm{patch}(x) = x + R \cdot (0, 0, ..., 0, K \cdot \eta)$, the adversary achieves with probability $1 - \delta$ that*

$$\forall x \in \Omega_{|y=-1} \colon \hat{f}_{poi}^T x > -K \Rightarrow \hat{f}_{poi}^T \mathrm{patch}(x) \geq 0 \ .$$

Proof. We define the gradient w.r.t. poisoned model $\hat{f}_{\text{poi}}$ as

$$g := \nabla_{\hat{f}_{\text{poi}}} L(\hat{f}_{\text{poi}}) = C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{poi}}^T x_i y_i < 1} -x_i y_i - C \cdot x_p \cdot \mathbb{1}\{\hat{f}_{\text{poi}}^T x_p < 1\} + \hat{f}_{\text{poi}}.$$

To exploit the backdoor during inference stage, the adversary desires the property $\hat{f}_{\text{poi}}^T x \geq 1$. The adversary constructs a single poison sample $x_p \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, so that for any classifier that does not satisfy this property, its gradient g will have $g^T x_p < 0$ almost certainly which implies $g \neq \mathbf{0}$, i.e., that classifier is not optimal. We investigate the probability of the adversary being successful:

$$\begin{aligned}
&\Pr[g^T x_p < 0] = \Pr[\big(C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{poi}}^T x_i y_i < 1} -x_i y_i \\
&- C \cdot x_p \cdot \underbrace{\mathbb{1}\{\hat{f}_{\text{poi}}^T x_p < 1\}}_{=1 \text{ by assumption}} + \hat{f}_{\text{poi}}\big)^T x_p < 0] \\
&= \Pr[C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{poi}}^T x_i y_i < 1} -x_p^T x_i y_i \\
&- C \cdot x_p^T x_p + \hat{f}_{\text{poi}}^T x_p < 0] \\
&= \Pr[C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{poi}}^T x_i y_i < 1} -(R \cdot (0, ..., 0, \eta))^T
\end{aligned}$$

Let $x_i^{(R)} := R^{-1} x_i$ and $x_{i,d}^{(R)}$ its d'th component, then

$$\begin{aligned}
&(R \cdot x_i^{(R)}) y_i \\
&- C \cdot x_p^T x_p + \hat{f}_{\text{poi}}^T x_p < 0] \\
&= \Pr[C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}: \hat{f}_{\text{poi}}^T x_i y_i < 1} -(0, ..., 0, \eta)^T x_i^{(R)} y_i \\
&- C \cdot x_p^T x_p + \hat{f}_{\text{poi}}^T x_p < 0] \\
&= \Pr[C \cdot \eta \cdot \underbrace{\sum_{x_i\in D_{\text{cl}}: \hat{f}_{\text{poi}}^T x_i y_i < 1} x_{i,d}^{(R)} y_i}_{\mathbb{L}^{\text{L2H}} :=} - C \cdot \underbrace{x_p^T x_p}_{=\eta^2} \\
&+ \underbrace{\hat{f}_{\text{poi}}^T x_p}_{<1} < 0] \geq \Pr[C \cdot \mathbb{L}^{\text{L2H}} \cdot \eta - C \cdot \eta^2 + 1 < 0]. \quad (8)
\end{aligned}$$

The adversary aims to bound $\mathbb{L}^{\text{L2H}}$ in order to obtain a magnitude of positive η that guarantees $g^T x_p < 0$ with high probability. To this end, for n data points in training data, the adversary establishes a bound on the sum of the magnitudes of benign training data points projected onto x_p, and uses μ_{signal} and σ^2_{signal} and Chebyshev's inequality [3]:

$$\begin{aligned}
&\Pr[|s_{|D_{\text{cl}}|}^{(u)} - \mu_{\text{signal}}| \geq k] \leq \sigma^2_{\text{signal}}/k^2 \stackrel{!}{=} \delta \\
&\Rightarrow k = \sigma^2_{\text{signal}}/\sqrt{\delta}.
\end{aligned}$$

The adversary assumes all data points to be active in the gradient and sets $\hat{s} := \mu_{\text{signal}} + k$ and obtains

$$\Pr[\mathbb{L}^{\text{L2H}} < \hat{s}] \geq 1 - \delta.$$

Then, solving the quadratic inequality of Equation (8) yields the solution $\eta^* > \frac{\hat{s} + \sqrt{\hat{s}^2 + \frac{4}{C}}}{2}$ and with that

$$\Pr[C \cdot \mathbb{L}^{\text{L2H}} \cdot \eta^* - C \cdot (\eta^*)^2 + 1 < 0] \geq 1 - \delta.$$

Consequently, when the poisoned classifier attains optimum, i.e., its gradient g satisfies $g = \mathbf{0}$, it has to hold that $\hat{f}_{\text{poi}}^T x_p \geq 1$ or else with high probability $g \neq \mathbf{0}$. Now the backdoor can be triggered via the patch function for every sample $x \in X \colon \hat{f}_{\text{poi}}^T x > -K$:

$$\begin{aligned}
\hat{f}_{\text{poi}}^T \mathrm{patch}(x) &= \hat{f}_{\text{poi}}^T \big(x + R \cdot (0, 0, ..., 0, K \cdot \eta)\big) \\
&= \hat{f}_{\text{poi}}^T x + K \cdot \hat{f}_{\text{poi}}^T \big(R \cdot (0, 0, ..., 0, \eta)\big) \\
&= \hat{f}_{\text{poi}}^T x + K \cdot \underbrace{\hat{f}_{\text{poi}}^T x_p}_{\geq 1} \geq \underbrace{\hat{f}_{\text{poi}}^T}_{\geq -K} x + K \geq 0.
\end{aligned}$$

□

In the prior theorem, the adversary can pick any rotation R for the poison sample x_p. When R is selected such that all samples from benign data distribution projected onto the poison are zero-centered and have zero magnitude, then $\mu_{\text{signal}} = 0$ and $\sigma^2_{\text{signal}} = 0$. In this special case, the adversary does not need to bound the impact of benign training data. Consequently, the bound on η simplifies to $\eta > 1/\sqrt{C}$ and the backdoor success probability is always exactly 1.

A.2 Single poison sample backdoor for dual optimization of linear classification

For linear classification instead of solving the task directly, one often solves the dual of the problem using dual optimization, e.g., via dual optimizers like `liblinear`. We show that solving the dual exhibits the same susceptibility to backdoor attacks as in the original task.

Corollary 6. *Consider the same adversary as in Theorem 5. If the learner utilizes dual optimization for obtaining a linear classifier, then with probability $1-\delta$ the learned model is backdoored.*

Proof. The dual learner obtains a solution (α, w) with $w = \sum_{i=1}^{n} \alpha_i x_i y_i$ that is dual optimal, i.e., is optimal for Equation (6). Because strong duality holds for linear SVM, the objective value d^* that (α, w) attains in the dual is also an optimal objective value for the primal of Equation (4), so (α, w) is also a minimizer of the primal of Equation (4). Now for the sake of contradiction, assume that w does not satisfy the adversary's desired backdoor property, i.e., assume that $w^T x_p < 1$. By Theorem 5, when choosing η appropriately, with probability arbitrarily close to 1, the gradient g_t projected onto poison direction is not zero and so $g_t \neq 0$. This implies that w can be further changed to achieve smaller objective value in Equation (1). This further implies that changing w can also reduce objective value in Equation (2), and also in Equation (4). This contradicts that (α, w) is a minimizer of the primal of Equation (4). □

A.3 Impact on benign learning task

We now analyze the impact of the single poison sample on the benign learning task. First, we show that if all samples from benign data distribution projected onto the poison sample are zero-centered and have zero magnitude, then the backdoor attack described in Theorem 5 does not impact the benign learning task.

Theorem 7. *Assume that $\mu_{signal} = 0$ and $\sigma^2_{signal} = 0$, i.e., all samples from benign data distribution projected onto the poison are zero-centered and have zero magnitude. Then for all $x \in X$, an optimal linear classifier w is functionally equivalent to an optimal w' that is obtained when the single poison x_p is omitted in training.*

Proof. Let x_p be defined as in Theorem 5. By definition, we have

$$\begin{aligned}
&\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{Hinge}}(D_{\text{poi}}, \hat{f}_{\text{poi}}) \\
=&\min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\hat{f}_{\text{poi}}\|_2^2 + C \cdot \sum_{(x_i,y_i)\in D_{\text{poi}}} \max\left(0, 1 - y_i \hat{f}_{\text{cl}}^T x_i\right) \\
=&\min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\hat{f}_{\text{poi}}\|_2^2 + C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} \max\left(0, 1 - y_i \hat{f}_{\text{poi}}^T x_i\right) \\
&+ C \cdot \max\left(0, 1 - \hat{f}_{\text{poi}}^T x_p\right) \ .
\end{aligned}$$

Now, we split the classifier $\hat{f}_{\text{poi}}$ into two parts: the part of $\hat{f}_{\text{poi}}$ projected onto x_p, i.e., $\text{proj}_{x_p}(\hat{f}_{\text{poi}}) := \hat{f}_{\text{poi}}^T x_p / \|x_p\|^2 \cdot x_p$, and the remainder of $\hat{f}_{\text{poi}}$, i.e., $\text{rem}_{x_p}(\hat{f}_{\text{poi}}) := \hat{f}_{\text{poi}} - \text{proj}_{x_p}(\hat{f}_{\text{poi}})$. We can thus write $\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{Hinge}}(D_{\text{poi}}, \hat{f}_{\text{poi}})$ as

$$\begin{aligned}
=&\min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \\
&+ C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} \\
&\max\left(0, 1 - y_i(\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}}))^T x_i\right) \\
&+ C \cdot \max\left(0, 1 - (\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}}))^T x_p\right).
\end{aligned}$$

By assumption, $\mu_{\text{signal}} = \sigma^2_{\text{signal}} = 0$, and we thus conclude

$$\begin{aligned}
=&\min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \qquad (*) \\
&+ C \cdot \sum_{(x_i,y_i)\in D_{\text{cl}}} \\
&\max\left(0, 1 - y_i(\underbrace{\text{proj}_{x_p}(\hat{f}_{\text{poi}})^T x_i}_{=0} + \text{rem}_{x_p}(\hat{f}_{\text{poi}})^T x_i)\right) \\
&+ C \cdot \max\left(0, 1 - (\text{proj}_{x_p}(\hat{f}_{\text{poi}})^T x_p\right. \\
&\left. + \underbrace{\text{rem}_{x_p}(\hat{f}_{\text{poi}})^T x_p}_{=0})\right).
\end{aligned}$$

As $\text{proj}_{x_p}(\hat{f}_{\text{poi}})$ and $\text{rem}_{x_p}(\hat{f}_{\text{poi}})$ are orthogonal, Lemma 1 implies

$$\begin{aligned}
&\|\text{proj}_{x_p}(\hat{f}_{\text{poi}}) + \text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \\
&= \|\text{proj}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 + \|\text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \ .
\end{aligned}$$

Using this equality in (*) gives us a term completely separating the projection from the remainder as

$$\begin{aligned}
&\min_{\hat{f}_{\text{poi}}} \frac{1}{2}\|\text{proj}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 + \frac{1}{2}\|\text{rem}_{x_p}(\hat{f}_{\text{poi}})\|_2^2 \\
&+ C \sum_{(x_i,y_i)\in D_{\text{cl}}} \max\left(0, 1 - y_i \text{rem}_{x_p}(\hat{f}_{\text{poi}})^T x_i\right) \\
&+ C \cdot \max\left(0, 1 - \text{proj}_{x_p}(\hat{f}_{\text{poi}})^T x_p\right) \\
=&\min_{f_1 \,:\, \text{proj}_{x_p}(f_1)=\mathbf{0}} \mathcal{L}_{\text{Hinge}}(f_1, D_{\text{cl}}) \\
&+ \min_{f_2 \,:\, \text{rem}_{x_p}(f_2)=\mathbf{0}} \mathcal{L}_{\text{Hinge}}(f_2, \{x_p\}) \ .
\end{aligned}$$

To deduce the functional equivalence of the classifiers, we need to show that minimizing

$$\arg \min_{f_1 \,:\, \text{proj}_{x_p}(f_1)=\mathbf{0}} \mathcal{L}_{\text{Hinge}}(f_1, D_{\text{cl}}) \qquad (9)$$

is the same as minimizing

$$\arg\min_f \mathcal{L}_{\text{Hinge}}(f, D_{\text{cl}}) \ .$$

Intuitively, this is true, because any vector of the form $\alpha \cdot x_p$ ($\alpha \in \mathbb{R} \setminus \{0\}$) is orthogonal to $\text{rem}_{x_p}(\hat{f}_{\text{poi}})^*$. Adding $\alpha \cdot x_p$ to a solution f_1^* of Equation (9) does not change the prediction of any benign training data point $(x, y) \in$

Ω and, consequently, cannot reduce any data point's loss. More formally, let l_{L2H} be a data point's Hinge loss, then

$$\begin{aligned}
&l_{\text{L2H}}(f_1^* + \alpha x_p, x)\\
=&\max\left(0, 1 - y(f_1^* + \alpha x_p)^T x\right)\\
=&\max\left(0, 1 - y((f_1^*)^T x + \alpha \underbrace{x_p^T x}_{=0})\right)\\
=&l_{\text{L2H}}(f_1^*, x).
\end{aligned}$$

As $\text{proj}_{x_p}(f_1) = \mathbf{0}$, we can again use Lemma 1 to show that the addition of $\alpha \cdot x_p$ can only increase the norm of the f_1^*, as

$$\begin{aligned}
&\|f_1^* + \alpha x_p\|_2^2\\
=&\|f_1^*\|_2^2 + \|\alpha x_p\|_2^2 > \|f_1^*\|_2^2 \ .
\end{aligned}$$

We thus obtain the functional equivalence for $x \in X$ due to the equality

$$\begin{aligned}
&(\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{Hinge}}(D_{\text{poi}}, \hat{f}_{\text{poi}}))^T x\\
=&(\min_{f} \mathcal{L}_{\text{Hinge}}(D_{\text{cl}}, f))^T x\\
&+ \underbrace{(\min_{f_2 : \text{rem}_{x_p}(f_2) = \mathbf{0}} \mathcal{L}_{\text{Hinge}}(f_2, \{x_p\}))^T x}_{=0}\\
=&(\min_{\hat{f}_{\text{poi}}} \mathcal{L}_{\text{Hinge}}(D_{\text{cl}}, \hat{f}_{\text{poi}}))^T x \ . \qquad \square
\end{aligned}$$

Again, this result also holds when using dual optimization instead of solving the linear classification task directly, as we show in the following result:

Theorem 8. *Consider the same setting as in Theorem 7. If the learner utilizes dual optimization for training linear classifier w, then this is functionally equivalent to optimizing a classifier w' with poison sample x_p omitted in training.*

Proof. The dual learner obtains a solution (α, w) with $w = \sum_{i=1}^n \alpha_i x_i y_i$ that is dual optimal, i.e., is optimal for Equation (6). Because strong duality holds for linear SVM, the objective value d^* that (α, w) attains in the dual is also optimal objective value for the primal of Equation (4), so (α, w) is also a minimizer of the primal of Equation (4). Consequently, w is also a minimizer of Equation (2) and of Equation (1). By Theorem 7, any optimum of Equation (1) satisfies the clean learning functional equivalence, so this holds for this specific w as well. $\square$

Now, we move on to the most general case where the benign data distribution can be any data distribution. For this, we build on prior work [16] to show that the impact on the benign learning task is still limited:

Corollary 9. *Corollary of [16, Theorem 1] Let μ_{cl} be the benign data distribution. We define the backdoor and poisoned distributions as*

$$\begin{aligned}
\mu_{bd}(x) &= \mathbb{1}\{x = x_p\},\\
\mu_{poi}(x) &= (1 - 1/n)\mu_{cl} + 1/n \mu_{bd} \ .
\end{aligned}$$

Let n be fixed. Let $\hat{f}_{poi}$ be the classifier trained on n samples from the poisoned distribution μ_{poi}. Let $l(.,.) : [0,1] \times [0,1] \mapsto \mathbb{R}^+$ be a general loss function that is (C, α)-Hölder continuous for $0 < \alpha \leq 1$ that measures the discrepancy between two classifiers. The statistical risk on benign input is bounded as

$$r_n^{cl}(\hat{f}_{poi}) \leq \frac{1}{1 - 1/n} r_n^{poi}(\hat{f}_{poi}) + C(\mu_{cl}(x_p) \cdot 1/n)^\alpha \ .$$

Proof. Let μ_{cl} be any clean distribution. We define the poisoned distributions as

$$\begin{aligned}
\mu_{\text{bd}}(x) &= \mathbb{1}\{x = x_p\},\\
\mu_{\text{poi}}(x) &= (1 - 1/n)\mu_{\text{cl}} + 1/n \mu_{\text{bd}} \ .
\end{aligned}$$

We derive an upper bound on $r_n^{\text{cl}}(\hat{f}_{\text{poi}})$. First, since l is (C, α)-Hölder continuous:

$$\begin{aligned}
&r_n^{\text{cl}}(\hat{f}_{\text{poi}}) = \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f_{\text{cl}}^*(x))\right]\right]\\
&\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f_{\text{poi}}^*(x)) + C\,|f_{\text{poi}}^*(x) - f_{\text{cl}}^*(x)|^\alpha\right]\right]\\
&\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f_{\text{poi}}^*(x))\right]\right]\\
&+ C \mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[|f_{\text{poi}}^*(x) - f_{\text{cl}}^*(x)|^\alpha\right]\\
&\leq \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f_{\text{poi}}^*(x))\right]\right]\\
&+ C \left(\mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[|f_{\text{poi}}^*(x) - f_{\text{cl}}^*(x)|\right]\right)^\alpha
\end{aligned}$$

using Jensen's inequality in the last step. We bound each term on the right-hand side independently. First, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x \sim \mu_{\text{cl}}}\left[l(\hat{f}_{\text{poi}}(x), f_{\text{poi}}^*(x))\right]\right]\\
&\leq (1 - 1/n)^{-1} \mathop{\mathbb{E}}_{D_{\text{poi}}}\left[\mathop{\mathbb{E}}_{x \sim \mu_{\text{poi}}}\left[l(\hat{f}_{\text{poi}}(x), f_{\text{poi}}^*(x))\right]\right]\\
&\leq (1 - 1/n)^{-1} r_n^{\text{poi}}(\hat{f}_{\text{poi}}).
\end{aligned}$$

As for the second term by definition of the Bayes optimal classifier considering 0-1-loss [20], we have

$$f_{\text{cl}}^*(x) = \Pr_{\mu_{\text{cl}}}[Y = 1 | X = x].$$

Similarly,

$$\begin{aligned}
f_{\text{poi}}^*(x) &= \Pr_{\mu_{\text{poi}}}[Y = 1 | X = x]\\
&= \begin{cases}
\Pr_{\mu_{\text{cl}}}[Y = 1 | X = x] & \text{if } x \neq x_p,\\
(1 - 1/n) \Pr_{\mu_{\text{cl}}}[Y = 1 | X = x] & \\
+ 1/n \Pr_{\mu_{\text{bd}}}[Y = 1 | X = x] & \text{if } x = x_p.
\end{cases}
\end{aligned}$$

Combining prior equations, we get

$$f_{\text{poi}}^* - f_{\text{cl}}^* = \begin{cases}
0 & \text{if } x \neq x_p,\\
1/n \cdot \Pr_{\mu_{\text{bd}}}[Y = 1 | X = x] & \\
-1/n \cdot \Pr_{\mu_{\text{cl}}}[Y = 1 | X = x] & \text{if } x = x_p.
\end{cases}$$

Accordingly,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mu_{\text{cl}}}\left[|f^*_{\text{poi}}(x) - f^*_{\text{cl}}(x)|\right] \\
&= \Pr_{\mu_{\text{cl}}}[X = x_p]\cdot \\
&\left|{}^1\!/_n \cdot \underbrace{\Pr_{\mu_{\text{bd}}}[Y = 1|X = x_p]}_{=1} - {}^1\!/_n \cdot \Pr_{\mu_{\text{cl}}}[Y = 1|X = x_p]\right| \\
&+ (1 - \Pr_{\mu_{\text{cl}}}[X = x_p]) \cdot 0 \\
&= \mu_{\text{cl}}(x_p) \cdot {}^1\!/_n \cdot \underbrace{\left|1 - \Pr_{\mu_{\text{cl}}}[Y = 1|X = x_p]\right|}_{\leq 1} \\
&\leq \mu_{\text{cl}}(x_p) \cdot {}^1\!/_n.
\end{aligned}
$$

using that $\Pr_{\mu_{\text{bd}}}[Y = 1|X = x_p] = 1$ in the equality since the poison label is 1. Now, we plug the two bounds together:

$$
r_n^{\text{cl}}\big(\hat{f}_{\text{poi}}\big) \leq \frac{1}{1 - {}^1\!/_n} r_n^{\text{poi}}\big(\hat{f}_{\text{poi}}\big) + C(\mu_{\text{cl}}(x_p) \cdot {}^1\!/_n)^\alpha. \qquad \square
$$

The attentive reader might be wondering why we consider i.i.d. sampling of the data set that is different to our prior assumption that the poison sample is guaranteed to be added to training data. Slightly increasing the poison ratio in the proof from ${}^1\!/_n$ to ${}^k\!/_n$ for a small constant k gives the probability of sampling at least one poison sample $1 - (1 - {}^k\!/_n)^n$. Under large enough data size n, this turns to $1 - (1 - {}^k\!/_n)^n \underset{n\to\infty}{\to} 1 - e^{-k}$ which quickly goes to 1 when increasing k. The bound on the statistical risk on benign data for k poison samples is $r_n^{\text{cl}}\big(\hat{f}_{\text{poi}}\big) \leq \frac{1}{1-{}^k\!/_n} r_n^{\text{poi}}\big(\hat{f}_{\text{poi}}\big) + C(\mu_{\text{cl}}(x_p) \cdot {}^k\!/_n)^\alpha$, so the bound is only slightly higher.