Label Inference Attacks from Log-loss Scores

Abhinav Aggarwal, Shiva P. Kasiviswanathan,
Zekun Xu, Oluwaseyi Feyisetan, Nathanael Teissier

Amazon

ICML 2021

Is it possible to recover all the ground truth labels from the Log-loss scores, even when these scores are noised?

[1] Whitehill, J. *"Climbing the kaggle leaderboard by exploiting the log-loss oracle."* AAAI 2018.

Is it possible to recover all the ground truth labels from the Log-loss scores, even when these scores are noised?

First introduced by Whitehill in the context of Kaggle competitions, where an algorithm that could recover some test labels (without any noise) was presented using a heuristic based on MCMC simulation[1].

[1] Whitehill, J. *"Climbing the kaggle leaderboard by exploiting the log-loss oracle."* AAAI 2018.

(Binary) Cross-entropy Loss

Given a vector $\mathbf{u} = (u_1, \ldots, u_N) \in [0, 1]^N$ and a labeling $\sigma \in \{0, 1\}^N$, the log-loss on $\mathbf{u}$ with respect to σ, denoted by $\textsc{LLoss}(\mathbf{u}; \sigma)$, is defined as follows:

$$\textsc{LLoss}(\mathbf{u}; \sigma) := \frac{-1}{N} \ln \left(\prod_{i=1}^{N} u_i^{\sigma_i} (1 - u_i)^{1-\sigma_i} \right).$$

Label Inference

Let $\sigma \in \{0,1\}^N$ be an (unknown) labeling. The label inference problem is that of recovering σ given $\textsc{LLoss}\left(\mathbf{u}_1;\sigma\right),\ldots,\textsc{LLoss}\left(\mathbf{u}_M;\sigma\right)$. Here, M is the number of queries and $\mathbf{u}_i \in [0,1]^N$ are the prediction vectors.

Overview of Our Results

Amount of Noise in Responses	Precision	# Log-loss Queries	#Arithmetic Operations
No noise	Arbitrary	1	$\tilde{O}(N)$
No noise	ϕ-bits	$\Theta\left(1 + N\phi 2^{-\phi/4}\right)$	$O(N)$
τ-accurate	Arbitrary	1	$O(2^N)$
τ-accurate	ϕ-bits	$O\left(\frac{N}{\log N} + \frac{N}{\log(\phi/N\tau)}\right)$	$O\left(\frac{poly(N,\phi/\tau)}{\log(\phi/N\tau)}\right)$

Table 1: Overview of our results for binary label inference. Here, $N \geq 1$ is the number of labels to be inferred. We present attacks under both arbitrary and bounded precision arithmetic models. The τ-accurate means that the error on the responses is atmost $|\tau|$. The last column represents the number of arithmetic operations needed at the adversary. All our adversaries are polytime except for the third row.

Label Inference from Raw Scores

Single Query Label Inference under Arbitrary Precision with Polynomial-time Adversary

Our task here is to compute all labels in a single query, when allowed arbitrary precision and polynomial time local computation.[2]

[2] Aggarwal *et al.* "*On Primes, Log-Loss Scores and (No) Privacy.*" EMNLP 2020.

Single Query Label Inference under Arbitrary Precision with Polynomial-time Adversary

Our task here is to compute all labels in a single query, when allowed arbitrary precision and polynomial time local computation.[2]

Set $u_i = \frac{p_i}{1+p_i}$, where p_i is the i^{th} prime.

[2] Aggarwal *et al.* "*On Primes, Log-Loss Scores and (No) Privacy.*" EMNLP 2020.

Single Query Label Inference under Arbitrary Precision with Polynomial-time Adversary

Our task here is to compute all labels in a single query, when allowed arbitrary precision and polynomial time local computation.[2]

Set $u_i = \frac{p_i}{1+p_i}$, where p_i is the i^{th} prime.

$$\text{LLoss}(\mathbf{u};\sigma) = \frac{-1}{N} \ln\left(\frac{\prod_{i=1}^{N} p_i^{\sigma_i}}{(1+p_1)\cdots(1+p_N)}\right)$$

$$\implies \prod_{i=1}^{N} p_i^{\sigma_i} = (1+p_1)\cdots(1+p_N)e^{-N\cdot\text{LLoss}(\mathbf{u};\sigma)}$$

The product on the left can be uniquely recovered.

[2] Aggarwal *et al.* "*On Primes, Log-Loss Scores and (No) Privacy.*" EMNLP 2020.

Label Inference under Bounded Precision with Polynomial-time Adversary

With bounded precision, using large primes and assuming an accurate rational form for $(1 + p_1) \cdots (1 + p_N) e^{-N \cdot \text{LLoss}(\mathbf{u}; \sigma)}$ is not possible.

Label Inference under Bounded Precision with Polynomial-time Adversary

With bounded precision, using large primes and assuming an accurate rational form for $(1+p_1)\cdots(1+p_N)e^{-N\cdot \text{LLoss}(\mathbf{u};\sigma)}$ is not possible.

We can issue multiple queries to infer only a few labels at a time — use $\mathbf{u} = \left[\frac{p_1}{1+p_1}, \cdots, \frac{p_m}{1+p_m}, \frac{1}{2}, \cdots, \frac{1}{2}\right]$.

Label Inference under Bounded Precision with Polynomial-time Adversary

With bounded precision, using large primes and assuming an accurate rational form for $(1+p_1)\cdots(1+p_N)e^{-N\cdot\text{LLoss}(\mathbf{u};\sigma)}$ is not possible.

We can issue multiple queries to infer only a few labels at a time — use $\mathbf{u} = \left[\frac{p_1}{1+p_1}, \cdots, \frac{p_m}{1+p_m}, \frac{1}{2}, \cdots, \frac{1}{2}\right]$.

Theorem

Let $\phi \geq 9$. There exists a polynomial-time adversary for the label inference problem in the $\text{FPA}(\phi)$ *model using* $\Theta\left(1 + N\phi 2^{-\phi/4}\right)$ *queries.*

$\text{FPA}(\phi)$: Finite Precision Arithmetic with ϕ bits of precision.

Label Inference under Bounded Precision with Polynomial-time Adversary

With bounded precision, using large primes and assuming an accurate rational form for $(1+p_1)\cdots(1+p_N)e^{-N\cdot \text{LLOSS}(\mathbf{u};\sigma)}$ is not possible.

We can issue multiple queries to infer only a few labels at a time — use $\mathbf{u} = \left[\frac{p_1}{1+p_1}, \cdots, \frac{p_m}{1+p_m}, \frac{1}{2}, \cdots, \frac{1}{2}\right]$.

Theorem

Let $\phi \geq 9$. There exists a polynomial-time adversary for the label inference problem in the FPA(ϕ) *model using* $\Theta\left(1 + N\phi 2^{-\phi/4}\right)$ *queries.*

FPA(ϕ): Finite Precision Arithmetic with ϕ bits of precision.
Observe the tight bound in the Theorem statement – lower bound derived using the Prime number theorem ($p_m = \Theta(m \log m)$).

Label Inference from Noised Scores

Robust Label Inference

Let $\tau > 0$ and $\sigma \in \{0, 1\}^N$ be the (unknown) labeling. The τ-robust label inference problem is that of recovering σ given $\ell_1, \ldots, \ell_M$, where for all $i \in [M]$, it holds that $|\text{LLOSS}(\mathbf{u}_i; \sigma) - \ell_i| \leq \tau$. Here, M is the number of queries and $\mathbf{u}_i \in [0, 1]^N$ are the prediction vectors.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Algorithm 1 Label Inference with Bounded Error in the APA Model (Exponential Adversary)

1: **Input:** N, upper bound on error $\tau > 0$
2: **Output:** Labeling $\hat{\sigma} \in \{0,1\}^N$
3: Let $\mathbf{u} = [u_1, \ldots, u_N]$, where $u_i \leftarrow 3^{2^i N\tau}/(1 + 3^{2^i N\tau})$.
4: Obtain the loss score ℓ using $\mathbf{u}$ as the prediction vector.
5: **Return** $\hat{\sigma} \leftarrow \arg\min_{\sigma \in \{0,1\}^N} |\textsc{LLoss}(\mathbf{u}, \sigma) - \ell|$.

APA: Arbitrary Precision Arithmetic.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Let us see why this works.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Let us see why this works.

The main idea is to make the outputs of the Log-loss function distinct for each labeling.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Let us see why this works.

The main idea is to make the outputs of the Log-loss function distinct for each labeling.

Define $\Delta(\mathbf{u}) := \min_{\sigma_1,\sigma_2\in\{0,1\}^N} |\textsc{LLoss}(\mathbf{u};\sigma_1) - \textsc{LLoss}(\mathbf{u};\sigma_2)|$.

Our task is to find $\mathbf{u}$ such that $\Delta(\mathbf{u}) > 2\tau$ (necessary and sufficient).

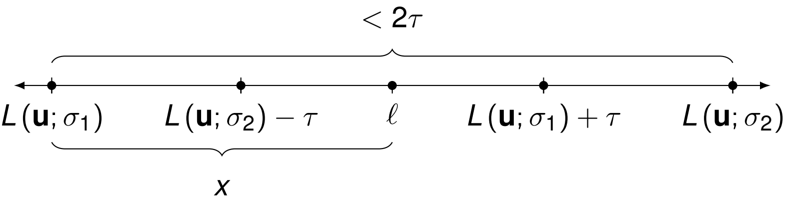
τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Let us see why this works.

The main idea is to make the outputs of the Log-loss function distinct for each labeling.

Define $\Delta(\mathbf{u}) := \min_{\sigma_1,\sigma_2 \in \{0,1\}^N} |\text{LLoss}(\mathbf{u};\sigma_1) - \text{LLoss}(\mathbf{u};\sigma_2)|$.

Our task is to find $\mathbf{u}$ such that $\Delta(\mathbf{u}) > 2\tau$ (necessary and sufficient).



$L \equiv \text{LLoss}$

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

How to find a vector $\mathbf{u}$ such that $\Delta(\mathbf{u}) > 2\tau$?

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

How to find a vector $\mathbf{u}$ such that $\Delta(\mathbf{u}) > 2\tau$?

For any set S, let $\mu(S) := \min_{S_1, S_2 \subseteq S} \left| \sum_{s_1 \in S_1} s_1 - \sum_{s_2 \in S_2} s_2 \right|$ denote the magnitude of the minimum difference between any two subset sums in S. For example, the set $S = \{1, 2, 4, \ldots, 2^m\}$ has $\mu(S) = 1$ for all m.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

How to find a vector $\mathbf{u}$ such that $\Delta(\mathbf{u}) > 2\tau$?

For any set S, let $\mu(S) := \min_{S_1, S_2 \subseteq S} \left| \sum_{s_1 \in S_1} s_1 - \sum_{s_2 \in S_2} s_2 \right|$ denote the magnitude of the minimum difference between any two subset sums in S. For example, the set $S = \{1, 2, 4, \ldots, 2^m\}$ has $\mu(S) = 1$ for all m.

Lemma

If all entries in $\mathbf{v} = [v_1, \ldots, v_N]$ *are distinct and positive, then*

$$\Delta\left(\frac{\mathbf{v}}{1+\mathbf{v}}\right) = \frac{1}{N}\mu(\ln \mathbf{v}),$$

where $\ln \mathbf{v} := [\ln v_1, \ldots, \ln v_N]$.

We use this lemma by setting $\mathbf{u} = \frac{\mathbf{v}}{1+\mathbf{v}}$ (element-wise).

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Thus, setting $\Delta(\mathbf{u}) > 2\tau$ is equivalent to setting $\frac{1}{N}\mu(\ln \mathbf{v}) > 2\tau$, or $\mu(\ln \mathbf{v}) > 2N\tau$. How do we ensure this?

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Thus, setting $\Delta(\mathbf{u}) > 2\tau$ is equivalent to setting $\frac{1}{N}\mu(\ln \mathbf{v}) > 2\tau$, or $\mu(\ln \mathbf{v}) > 2N\tau$. How do we ensure this?

Recall $\mu\left(\{1, 2, 4, \ldots, 2^N\}\right) = 1$.

Set $v_i = 3^{2^i N\tau}$. Then, $\ln v_i = 2^i N\tau \ln 3 = 2^{i-1}\left(2N\tau \ln 3\right)$.

We can now compute $\mu\left(\{\ln v_1, \ldots, \ln v_N\}\right) = 2N\tau \ln 3 > 2N\tau$.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Thus, setting $\Delta(\mathbf{u}) > 2\tau$ is equivalent to setting $\frac{1}{N}\mu(\ln \mathbf{v}) > 2\tau$, or $\mu(\ln \mathbf{v}) > 2N\tau$. How do we ensure this?

Recall $\mu\left(\{1, 2, 4, \ldots, 2^N\}\right) = 1$.

Set $v_i = 3^{2^i N\tau}$. Then, $\ln v_i = 2^i N\tau \ln 3 = 2^{i-1}\left(2N\tau \ln 3\right)$.

We can now compute $\mu\left(\{\ln v_1, \ldots, \ln v_N\}\right) = 2N\tau \ln 3 > 2N\tau$.

From this, we obtain the desired prediction vector for τ-robust label inference as:

$$u_i = \frac{v_i}{1 + v_i} = \frac{3^{2^i N\tau}}{1 + 3^{2^i N\tau}}.$$

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Is it possible to avoid exponentially large entries in $v_i = 3^{2^i N \tau}$?

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Is it possible to avoid exponentially large entries in $v_i = 3^{2^i N\tau}$?

Generalized Result from Euler's Theorem:

Theorem

For any set $S \subset \mathbb{Q}^+$ with $\mu(S) > \lambda$ for some $\lambda \in [0, \infty)$, it must hold that $||S||_\infty = \Omega(\lambda 2^{|S|})$.

τ-Robust Label Inference under Arbitrary Precision with Exponential-time Adversary

Is it possible to avoid exponentially large entries in $v_i = 3^{2^i N\tau}$?

Generalized Result from Euler's Theorem:

Theorem

For any set $S \subset \mathbb{Q}^+$ with $\mu(S) > \lambda$ for some $\lambda \in [0, \infty)$, it must hold that $||S||_\infty = \Omega(\lambda 2^{|S|})$.

Bound for robust vector construction:

Theorem

For sufficiently large N and all $\tau > 0$, any vector $\mathbf{u} = \frac{\mathbf{v}}{1+\mathbf{v}}$ must have $||\mathbf{v}||_\infty = \Omega\left(e^{2^N N\tau}\right)$ to allow τ-robust label inference using $\mathbf{u}$.

τ-Robust Label Inference under Bounded Precision with Polynomial-time Adversary

τ-Robust Label Inference under Bounded Precision with Polynomial-time Adversary

Recall label algorithm in the APA model:

Algorithm 2 Label Inference in APA Model with Exponential Adversary

1: **Input:** N, upper bound on error $\tau > 0$
2: **Output:** Labeling $\hat{\sigma} \in \{0, 1\}^N$
3: Let $\mathbf{u} = [u_1, \ldots, u_N]$, where $u_i \leftarrow 3^{2^i N\tau} / (1 + 3^{2^i N\tau})$.
4: Obtain the loss score ℓ using $\mathbf{u}$ as the prediction vector.
5: **Return** $\hat{\sigma} \leftarrow \arg\min_{\sigma \in \{0,1\}^N} |\mathcal{L}_{\mathbf{u}}(\sigma) - \ell|$.

τ-Robust Label Inference under Bounded Precision with Polynomial-time Adversary

Recall label algorithm in the APA model:

Algorithm 3 Label Inference in APA Model with Exponential Adversary

1: **Input:** N, upper bound on error $\tau > 0$
2: **Output:** Labeling $\hat{\sigma} \in \{0,1\}^N$
3: Let $\mathbf{u} = [u_1, \ldots, u_N]$, where $u_i \leftarrow 3^{2^i N\tau}/(1+3^{2^i N\tau})$.
4: Obtain the loss score ℓ using $\mathbf{u}$ as the prediction vector.
5: **Return** $\hat{\sigma} \leftarrow \arg\min_{\sigma \in \{0,1\}^N} |\mathcal{L}_{\mathbf{u}}(\sigma) - \ell|$.

Limitations for FPA model and polynomial time adversary:

1. Intermediate computations are exponentially large for $u_i \leftarrow 3^{2^i N\tau}/(1+3^{2^i N\tau})$.
2. Iterating over all labelings in $\arg\min_{\sigma \in \{0,1\}^N} |\mathcal{L}_{\mathbf{u}}(\sigma) - \ell|$ infeasible.

τ-Robust Label Inference under Bounded Precision with Polynomial-time Adversary

A similar trick to the unnoised case works here – infer a few labels at a time using

$$\mathbf{u} = \left[\frac{3e^{2N\tau}}{1+3e^{2N\tau}}, \frac{3e^{4N\tau}}{1+3e^{4N\tau}}, \ldots, \frac{3e^{2^m N\tau}}{1+3e^{2^m N\tau}}, \frac{1}{2}, \ldots, \frac{1}{2}\right].$$

τ-Robust Label Inference under Bounded Precision with Polynomial-time Adversary

A similar trick to the unnoised case works here – infer a few labels at a time using

$$\mathbf{u} = \left[\frac{3e^{2N\tau}}{1+3e^{2N\tau}}, \frac{3e^{4N\tau}}{1+3e^{4N\tau}}, \ldots, \frac{3e^{2^m N\tau}}{1+3e^{2^m N\tau}}, \frac{1}{2}, \ldots, \frac{1}{2}\right].$$

Theorem

For any error bounded by $\tau > 0$ and $\phi \geq 8 + \lceil N\tau \ln 2 \rceil$, there exists a polynomial-time adversary for the τ-label inference problem in the FPA(ϕ) *model using* $O\left(\frac{N}{\log N} + \frac{N}{\log(\phi/N\tau)}\right)$ *queries.*

Inference is done $m = \min\left\{\lceil \log_2 N \rceil, \left\lfloor \log_2\left(\frac{\phi-8}{N\tau \ln 2}\right)\right\rfloor\right\}$ labels at a time.

Empirical Observations

Experiments on Real Datasets

The list of datasets we use is as follows, fetched from the UCI machine learning dataset repository[3]:

1. **D1** (IMDB movie review for sentiment analysis – 0 (negative review) or 1 (positive review);
2. **D2** (Banknote Authentication) – 0 (fine) or 1 (forged);
3. **D3** (Wisconsin Cancer) – 0 (benign) and 1 (malignant);
4. **D4** (Haberman's Survival) – 0 (survived) and 1 (died).

[3] https://archive.ics.uci.edu/ml/machine-learning-databases

Experiments on Real Datasets

Table 2: Experimental results for unnoised label inference with polynomial time adversary. Here, N is the number of test samples in the dataset and $\mathbf{Acc}_q$ is the fraction of labels correctly inferred with q queries.

Dataset	N	$\mathbf{Acc}_1$	$\mathbf{Acc}_{N/5}$	$\mathbf{Time}_{N/5}$
D1	25,000	0.4891	1.0	53.41 ms
D2	1,372	0.4446	1.0	0.2 ms
D3	569	0.3448	1.0	0.06 ms
D4	306	0.2647	1.0	0.03 ms

Experiments on Real Datasets

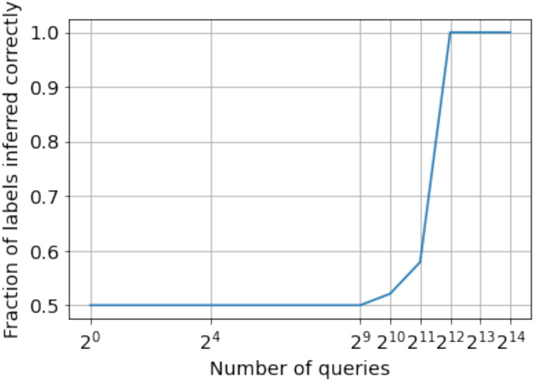


Figure 1: Accuracy of (unnoised) label inference on dataset **D1** as a function of the number of queries used by the adversary. For $N/5 = 5000$ queries, all labels have been correctly inferred.

Experiments for noised label inference on Simulated Binary Labelings

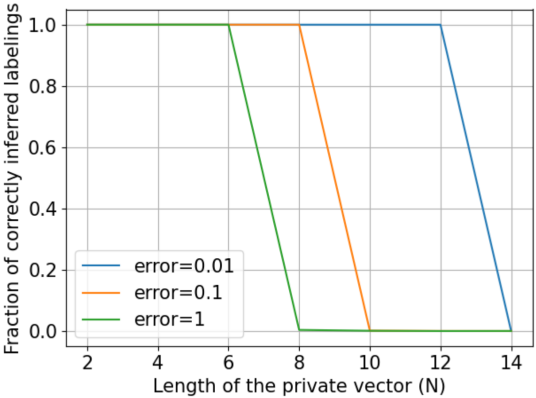


Figure 2: Accuracy of single-query (noised) label inference on simulated binary labelings with bounded error (scale = 0.01, 0.1, and 1).

Concluding Remarks

We demonstrated that log-loss scores can leak information about the ground truth labels, even when noised arbitrarily. This information can be exploited using specially constructed prediction vectors, without any access to the underlying dataset or model training.

Concluding Remarks

We demonstrated that log-loss scores can leak information about the ground truth labels, even when noised arbitrarily. This information can be exploited using specially constructed prediction vectors, without any access to the underlying dataset or model training.

How can we defend against these attacks?

1. Compute loss scores on random subsets.
2. Randomized Response – will impart protection through plausible deniability.

Concluding Remarks

We demonstrated that log-loss scores can leak information about the ground truth labels, even when noised arbitrarily. This information can be exploited using specially constructed prediction vectors, without any access to the underlying dataset or model training.

How can we defend against these attacks?

1. Compute loss scores on random subsets.
2. Randomized Response – will impart protection through plausible deniability.

Future work: Characterize the class of loss functions for which robust label inference is feasible.

Thank you for attending the talk!