Supplementary Materials for Black-Box Data Poisoning Attacks on Crowdsourcing

In this supplementary document, we provide the following details to support the main text.

Section A: Mathematical proofs.

Section B: More details about the proposed algorithm.

A. The detailed proof

A.1. Proof for Theorem 1

Theorem 1. *let $\mathbf{P}^{(j)}$ denote the confusion matrix and π_k^* denote the prior of class l_k, f is equivalent to the Dawid-Skene model when $\mathbf{W}^{(ij)} = \ln \mathbf{P}^{(j)}$ and $w_k^* = \ln \pi_k^*$.*

Proof. Since $\left(\bar{w}_{-1}^{(i)}, \bar{w}_{+1}^{(i)}\right) = \sum_j \mathbf{v}^{(ij)} (\mathbf{W}^{(ij)})^{\mathrm{T}} + \left(w_{-1}^*, w_{+1}^*\right)$, when $\mathbf{W}^{(ij)} = \ln \mathbf{P}^{(j)}$ and $w_k^* = \ln \pi_k^*$, we obtain:

$$\bar{w}_{-1}^{(i)} = \ln \Pi_j \left(p_{-1-1}^{(j)}\right)^{\frac{1}{2} t_{ij}\left(1-y_{ij}\right)} \cdot \left(p_{-1+1}^{(j)}\right)^{\frac{1}{2} t_{ij}\left(1+y_{ij}\right)} \cdot \pi_{-1}^*$$

$$\bar{w}_{+1}^{(i)} = \ln \Pi_j \left(p_{+1-1}^{(j)}\right)^{\frac{1}{2} t_{ij}\left(1-y_{ij}\right)} \cdot \left(p_{+1+1}^{(j)}\right)^{\frac{1}{2} t_{ij}\left(1+y_{ij}\right)} \cdot \pi_{+1}^*.$$

Then, we obtain that

$$\begin{aligned} f\left(\mathbf{Y}_{i*}\right) &= \operatorname*{argmax}_{k \in \{-1,+1\}} \bar{w}_k^{(i)} \\ &= \operatorname*{argmax}_{k \in \{-1,+1\}} \frac{\exp \bar{w}_k^{(i)}}{\exp \bar{w}_{-1}^{(i)} + \exp \bar{w}_{+1}^{(i)}} \\ &= \operatorname*{argmax}_{k \in \{-1,+1\}} \rho_k, \end{aligned}$$

where ρ_k is the posterior of class l_k. Therefore, f denotes DS label aggregation model. □

A.2. Proof for Corollary 1

Corollary 1. *f is equivalent to ZenCrowd, when $w_k^* = 0$ and*

$$\mathbf{W}^{(ij)} = \begin{pmatrix} \ln(p_j^*) & \ln(1-p_j^*) \\ \ln(1-p_j^*) & \ln(p_j^*) \end{pmatrix}, \tag{1}$$

where p_j^* denotes the reliability parameters of workers.

Proof. Since $\left(\bar{w}_{-1}^{(i)}, \bar{w}_{+1}^{(i)}\right) = \sum_j \mathbf{v}^{(ij)} (\mathbf{W}^{(ij)})^{\mathrm{T}} + \left(w_{-1}^*, w_{+1}^*\right)$, we obtain

$$\bar{w}_{-1}^{(i)} = \ln \Pi_j \left(p_j^*\right)^{\frac{1}{2} t_{ij}\left(1-y_{ij}\right)} \cdot \left(1-p_j^*\right)^{\frac{1}{2} t_{ij}\left(1+y_{ij}\right)}$$

$$\bar{w}_{+1}^{(i)} = \ln \Pi_j \left(1-p_j^*\right)^{\frac{1}{2} t_{ij}\left(1-y_{ij}\right)} \cdot \left(p_j^*\right)^{\frac{1}{2} t_{ij}\left(1+y_{ij}\right)},$$

when $w_k^* = 0$ and

$$\mathbf{W}^{(ij)} = \begin{pmatrix} \ln(p_j^*) & \ln(1-p_j^*) \\ \ln(1-p_j^*) & \ln(p_j^*) \end{pmatrix}. \tag{2}$$

Finally, we obtain that

$$f\left(\mathbf{Y}_{i*}\right) = \operatorname{argmax}_{l_k} \rho_k,$$

Therefore, f denotes the aggregation rule of the ZenCrowd model. □

A.3. Proof for Theorem 2

Theorem 2. *When $\mathbf{W}^{(ij)} = \mathbf{I}$ and $w_k^* = 0$, f is equivalent to majority voting, where $\mathbf{I}$ is the identity matrix.*

Proof. When $\mathbf{W}^{(ij)} = \mathbf{I}$ and $w_k^* = 0$, we obtain

$$\bar{w}_{-1}^{(i)} = \frac{1}{2} t_{ij}(1-y_{ij}),$$

$$\bar{w}_{+1}^{(i)} = \frac{1}{2} t_{ij}(1+y_{ij}).$$

Finally, we obtain that

$$f\left(\mathbf{Y}_{i*}\right) = \operatorname{argmax}_{l_k} \sum_j \mathbb{1}(y_{ij} = l_k) \tag{3}$$

where $\mathbb{1}(x = k)$ is the indicator function, which is 1, when $x = y$, or otherwise. Therefore, f denotes the aggregation rule of majority voting. □

A.4. Proof for Theorem 3

Theorem 3. *When $\mathbf{W}^{(ij)} = d_j \mathbf{I}$ and $w_k^* = 0$, f is equivalent to weighted majority voting, where d_j denotes the weight of u_j who provides a label to instance $\mathbf{x}_i$.*

Proof. When $\mathbf{W}^{(ij)} = d_j \mathbf{I}$ and $w_k^* = 0$, we obtain

$$f\left(\mathbf{Y}_{i*}\right) = \text{argmax}_{l_k} \sum\nolimits_j d_j \mathbb{1}(y_{ij} = l_k) \tag{4}$$

Therefore, f denotes the aggregation rule of weighted majority voting. $\square$

A.5. Gradients with mathematical proofs

A.5.1. Gradient $\nabla_{\tilde{y}_{ij'}} \Psi$.

Theorem 4. *Given the Lagrangian Ψ and ψ, and $\mathbf{v} = (-1, +1)$, we have the gradient*

$$\nabla_{\tilde{y}_{ij'}} \Psi = \frac{1}{|\mathcal{X}|} \cdot (\sigma(\bar{w}_{+1}^{(i)}) - \sigma(\hat{w}_{+1}^{(i)})) \cdot \left(\frac{\partial \hat{w}_{+1}^{(i)}}{\partial \tilde{y}_{ij'}} - \frac{\partial \hat{w}_{-1}^{(i)}}{\partial \tilde{y}_{ij'}} \right) + \frac{\lambda \tilde{t}_{ij'}}{|\tilde{\mathcal{U}}| \sum_i \tilde{t}_{ij'}} \left(\frac{1 - \sigma(\bar{w}_{+1}^{(i)})}{1 - \tilde{y}_{ij'}} - \frac{\sigma(\bar{w}_{+1}^{(i)})}{\tilde{y}_{ij'}} \right),$$

where $\sigma(\cdot)$ is softmax function which makes weights $(\hat{w}_{-1}^{(i)}, \hat{w}_{+1}^{(i)})$ or $(\bar{w}_{-1}^{(i)}, \bar{w}_{+1}^{(i)})$ be distributions.

Proof. The final aggregated label of an instance depends on the weights $(\hat{w}_{-1}^{(i)}, \hat{w}_{+1}^{(i)})$ or $(\bar{w}_{-1}^{(i)}, \bar{w}_{+1}^{(i)})$. Thus, we reformulate the loss function as follows.

$$L = -\frac{1}{|\mathcal{X}|} \sum\nolimits_i v(\sigma(\hat{\mathbf{w}}^{(i)}), \sigma(\bar{\mathbf{w}}^{(i)})) + \frac{\lambda}{|\tilde{\mathcal{U}}|} \sum\nolimits_{j'} \frac{\sum_i \tilde{t}_{ij'} v\left(\tilde{\mathbf{D}}_{ij'}, \sigma\left(\bar{\mathbf{w}}^{(i)}\right)\right)}{\sum_i \tilde{t}_{ij'}}, \tag{5}$$

where $\hat{\mathbf{w}}^{(i)} = \left(\hat{w}_{-1}^{(i)}, \hat{w}_{+1}^{(i)}\right)$, $\bar{\mathbf{w}}^{(i)} = \left(\bar{w}_{-1}^{(i)}, \bar{w}_{+1}^{(i)}\right)$ and $\sigma(\cdot)$ is softmax function which makes them be distributions. Correspondingly, v is the cross-entropy and $\tilde{\mathbf{D}}_{ij'} = (1 - \tilde{y}_{ij'}, \tilde{y}_{ij'})$.

Then, L directly hinges on $\sigma(\hat{\mathbf{w}}^{(i)})$ and $\sigma(\bar{\mathbf{w}}^{(i)})$ instead of $f\left(\mathbf{Y}'_{i*}\right)$ and $f\left(\mathbf{Y}_{i*}\right)$. We relax the outer subproblem of the bilevel program as follows.

$$\begin{aligned} \min_{\tilde{\mathbf{Y}}, \tilde{\mathbf{T}}} \Psi = L + \psi(\sum\nolimits_i \sum\nolimits_{j'} \frac{1}{2}(1 + \text{sign}(t_{ij'} - 1/2)) - B) \\ s.t.\ \ \tilde{\mathbf{T}} \in [0,1]^{|\mathcal{X}| \times |\tilde{\mathcal{U}}|} \\ \tilde{\mathbf{Y}} \in [0,1]^{|\mathcal{X}| \times |\tilde{\mathcal{U}}|}, \end{aligned} \tag{6}$$

where $\hat{\mathbf{w}}^{(i)} = \left(\bar{w}_{-1}^{(i)}, \bar{w}_{+1}^{(i)}\right) + \sum_{j'} \tilde{\mathbf{v}}^{(ij')} (\tilde{\mathbf{W}}^{(ij')})^{\mathrm{T}} + \left(w_{-1}^*, w_{+1}^*\right)$. With the chain rule, we compute the gradient $\nabla_{\tilde{y}_{i,j'}} \Psi$ as follows.

$$\nabla_{\tilde{y}_{i,j'}} \Psi = \nabla_{\tilde{y}_{i,j'}} d_1 + \nabla_{\tilde{y}_{i,j'}} d_2, \tag{7}$$

where $d_1 = -\frac{1}{|\mathcal{X}|} \sum_i v(\sigma(\hat{\mathbf{w}}^{(i)}), \sigma(\bar{\mathbf{w}}^{(i)}))$ and $d_2 = \frac{\lambda}{|\tilde{\mathcal{U}}|} \sum_{j'} \frac{\sum_i \tilde{t}_{ij'} v\left(\tilde{\mathbf{D}}_{ij'}, \sigma\left(\bar{\mathbf{w}}^{(i)}\right)\right)}{\sum_i \tilde{t}_{ij'}}$. Then, we compute $\frac{\partial d_1}{\partial \tilde{y}_{ij'}}$ and $\frac{\partial d_2}{\partial \tilde{y}_{ij'}}$ as follows.

$$\begin{aligned} \frac{\partial d_1}{\partial \tilde{y}_{ij'}} &= \frac{1}{|\mathcal{X}|} \cdot (\sigma(\bar{w}_{+1}^{(i)}) - \sigma(\hat{w}_{+1}^{(i)})) \cdot \left(\frac{\partial \hat{w}_{+1}^{(i)}}{\partial \tilde{y}_{ij'}} - \frac{\partial \hat{w}_{-1}^{(i)}}{\partial \tilde{y}_{ij'}} \right), \\ \frac{\partial d_2}{\partial \tilde{y}_{ij'}} &= \frac{\lambda \tilde{t}_{ij'}}{|\tilde{\mathcal{U}}| \sum_i \tilde{t}_{ij'}} \left(\frac{1 - \sigma(\bar{w}_{+1}^{(i)})}{1 - \tilde{y}_{ij'}} - \frac{\sigma(\bar{w}_{+1}^{(i)})}{\tilde{y}_{ij'}} \right), \end{aligned} \tag{8}$$

where $\left(\frac{\partial \hat{w}_{-1}^{(i)}}{\partial \tilde{y}_{ij'}}, \frac{\partial \hat{w}_{+1}^{(i)}}{\partial \tilde{y}_{ij'}}\right) = \frac{1}{2} \tilde{t}_{ij'} \mathbf{v} (\tilde{\mathbf{W}}^{(ij')})^{\mathrm{T}}$ and $\mathbf{v} = (-1, +1)$. $\square$

Finally, we derive the gradient of Ψ w.r.t. $\tilde{y}'_{ij'}$ as follows.

$$\nabla_{\tilde{y}'_{ij'}} \Psi = \nabla_{\tilde{y}_{ij'}} \Psi \cdot \nabla_{\tilde{y}'_{ij'}} \tilde{y}_{ij'},$$

where $\nabla_{\tilde{y}'_{ij'}} \tilde{y}_{ij'} = \text{sigmoid}(\tilde{y}'_{ij'}) \cdot (1 - \text{sigmoid}(\tilde{y}'_{ij'}))$.

A.5.2. Gradient $\nabla_{\tilde{t}_{ij'}} \Psi$

Theorem 5. *Given the Lagrangian Ψ and a constant θ, we have the gradient*

$$\nabla_{\tilde{t}_{ij'}} \Psi = \frac{1}{|\mathcal{X}|} \cdot (\sigma(\bar{w}_{+1}^{(i)}) - \sigma(\hat{w}_{+1}^{(i)})) \cdot \left(\frac{\partial \hat{w}_{+1}^{(i)}}{\partial \tilde{t}_{ij'}} - \frac{\partial \hat{w}_{-1}^{(i)}}{\partial \tilde{t}_{ij'}} \right) + \frac{\sum_{i'} \lambda \tilde{t}_{i'j'} (v(\tilde{\mathbf{D}}_{ij'}, \sigma(\bar{w}_{+1}^{(i)})) - v(\tilde{\mathbf{D}}_{i'j'}, \sigma(\bar{w}_{+1}^{(i)})))}{|\mathcal{X}| (\sum_i \tilde{t}_{ij'})^2} + \frac{2\theta\psi \cdot e^{2\theta \tilde{t}_{ij'} - 1}}{(e^{2\theta \tilde{t}_{ij'} - 1} + 1)^2},$$

where $v(p, q)$ is the cross-entropy of two distributions and $\tilde{\mathbf{D}}_{ij'} = (1 - \tilde{y}_{ij'}, \tilde{y}_{ij'})$.

Proof. With the chain rule, we compute the gradient $\nabla_{\tilde{t}_{i,j'}} \Psi$ as follows.

$$\nabla_{\tilde{t}_{i,j'}} \Psi = \nabla_{\tilde{t}_{i,j'}} d_1 + \nabla_{\tilde{t}_{i,j'}} d_2 + \nabla_{\tilde{t}_{i,j'}} d_3, \tag{9}$$

where, $d_3 = \psi(\sum_i \sum_{j'} \frac{1}{2}(1 + \text{sign}(\tilde{t}_{ij'} - 1/2)) - B)$. Similarly, we obtain $\frac{\partial d_1}{\partial \tilde{t}_{ij'}}$ and $\frac{\partial d_2}{\partial \tilde{t}_{ij'}}$.

$$\begin{aligned} \frac{\partial d_1}{\partial \tilde{t}_{ij'}} &= \frac{1}{|\mathcal{X}|} \cdot (\sigma(\bar{w}_{+1}^{(i)}) - \sigma(\hat{w}_{+1}^{(i)})) \cdot \left(\frac{\partial \hat{w}_{+1}^{(i)}}{\partial \tilde{t}_{ij'}} - \frac{\partial \hat{w}_{-1}^{(i)}}{\partial \tilde{t}_{ij'}} \right) \\ \frac{\partial d_2}{\partial \tilde{t}_{ij'}} &= \frac{\lambda \sum_{i'} \tilde{t}_{i'j'} (v(\tilde{\mathbf{D}}_{ij'}, \sigma(\bar{w}_{+1}^{(i)})) - v(\tilde{\mathbf{D}}_{i'j'}, \sigma(\bar{w}_{+1}^{(i)})))}{|\mathcal{X}| (\sum_i \tilde{t}_{ij'})^2}, \end{aligned} \tag{10}$$

where

$$\frac{\partial \hat{w}_{+1}^{(i)}}{\partial \tilde{t}_{ij'}} = \frac{1}{2}(\tilde{w}_{+1,-1}^{(ij')} \cdot (1 - \tilde{y}_{ij'}) + \tilde{w}_{+1,+1}^{(ij')}(1 + \tilde{y}_{ij'})) \tag{11}$$

$$\frac{\partial \hat{w}_{-1}^{(i)}}{\partial \tilde{t}_{ij'}} = \frac{1}{2}(\tilde{w}_{-1,-1}^{(ij')} \cdot (1 - \tilde{y}_{ij'}) + \tilde{w}_{-1,+1}^{(ij')}(1 + \tilde{y}_{ij'})) \tag{12}$$

It is hard to compute the third gradient, as the sign function $h_1(x) = \mathrm{sign}(x)$ is not continuous. To address this problem, we approximate $h_1(x) = \mathrm{sign}(x)$ by $h_2(x) = \tanh(\theta x)$, when $x \in (-1, 1)$. Then, the third gradient can be computed as follows.

$$\frac{\partial d_3}{\partial \tilde{t}_{ij'}} = \frac{2\theta\psi \cdot e^{2\theta \tilde{t}_{ij'}-1}}{(e^{2\theta \tilde{t}_{ij'}-1} + 1)^2}. \tag{13}$$

□

Similarly, the gradient $\nabla_{\tilde{t}'_{ij'}} \Psi$ is calculated as: $\nabla_{\tilde{t}'_{ij'}} \Psi = \nabla_{\tilde{t}_{ij'}} \Psi \cdot \nabla_{\tilde{t}'_{ij'}} \tilde{t}_{ij'}$, where $\nabla_{\tilde{t}'_{ij'}} \tilde{t}_{ij'} = \mathrm{sigmoid}(\tilde{t}'_{ij'}) \cdot (1 - \mathrm{sigmoid}(\tilde{t}'_{ij'}))$.

B. More details about the proposed algorithm.

The first line of pseudo-code in SubPac initializes the labeling strategy, and lines 2 and 3 are both convergence conditions for the algorithm. Line 4 is responsible for updating the parameters of the label aggregation method. Lines 5 -8 are responsible for updating the task selection strategy of malicious workers and Lines 9 -12 are responsible for updating the labeling strategies of malicious workers. Line 13 is responsible for updating the Lagrangian multipliers. Line 14 returns the malicious attack strategy.

B.1. Applied to multiple option settings

It is straightforward to extend the proposed method to the multi-option setting of the labeling task. To do so, we first extend $\mathbf{v}^{(ij)}$ and $\mathbf{W}^{(ij)}$ to the multi-option setting.

$$\mathbf{v}^{(ij)} = t_{ij} \cdot (y_{ij}^1, \cdots, y_{ij}^k, \cdots, y_{ij}^K), \tag{14}$$

$$\mathbf{W}^{(ij)} = (w_{kh}^{(ij)})_{K \times K}, \tag{15}$$

where K denotes the number of options and y_{ij}^k denotes the indicator whether worker u_j provides label l_k to instance $\mathbf{x}_i$. For instance, if $K = 5$ and the label from u_j to $\mathbf{x}_i$ is l_3 option, $\mathbf{v}^{(ij)} = (0, 0, 1, 0, 0)$ which is the one-hot encoding of worker u_j to instance $\mathbf{x}_i$. $\tilde{\mathbf{v}}^{(ij')}$ and $\tilde{\mathbf{W}}^{(ij')}$ can be extended similarly to the multi-option setting. Then, we obtain the general representation of label aggregation models before and after attacks. The attack strategy in such a scenario can be derived by solving the following optimization problem:

$$\begin{aligned} \min_{\tilde{\mathbf{Y}}, \tilde{\mathbf{T}}} L, \quad s.t.\ & f(\mathbf{Y}'_{i*}) = \max\nolimits_{l_k} \hat{w}_k^{(i)} \\ & \sum\nolimits_i \sum\nolimits_{j'} \tilde{t}_{ij'} = B, \tilde{\mathbf{T}} \in \{0,1\}^{|\mathcal{X}| \times |\tilde{\mathcal{U}}|}, \\ & \tilde{\mathbf{Y}} \in \{l_1, l_2, \cdots, l_K\}^{|\mathcal{X}| \times |\tilde{\mathcal{U}}|}. \end{aligned} \tag{16}$$

In the reparameterization trick, we use the softmax function.

$$\tilde{\mathbf{Y}} = \mathrm{softmax}(\tilde{\mathbf{Y}}'), \tag{17}$$

$$\tilde{\mathbf{T}} = \mathrm{softmax}(\tilde{\mathbf{T}}'), \tag{18}$$

which enables $\tilde{\mathbf{Y}}$ to be updated in a gradient-based optimization algorithm. In order to derive the optimal strategy, we use $\tilde{y}_{ij'} = \mathrm{argmax}_{l_k} \tilde{y}_{ij'}^k$ instead in Algorithm 1.

B.2. The definition of attack success rate in crowdsourcing

We define the attack success rate of substitution-based attacks on the target model for the measurement of attack transferability. We denote by $\mathcal{F} = \{(f_{c'}, \lambda_{c'})\}_{c'=1}^{|\mathcal{F}|}$ the set of two-tuples, where $f_{c'}$ denotes the c'-th victim model and $\lambda_{c'}$ is the attack designed for it. We denote by $f_{c'}^{\lambda_c}, c \in \{1, 2, \cdots, |\mathcal{F}|\}$ the label aggregation model $f_{c'}$ under attack λ_c. The attack success rate of poisoning attack λ_c designed for substitute f_c on the target model $f_{c'}$ is defined as follows:

$$\mathrm{Asr}_{c'c} = \frac{1}{N'} \sum\nolimits_i \mathbf{1}\left(f_{c'}\left(\mathbf{Y}_{i*}\right) = z_i \wedge f_{c'}^{\lambda_c}\left(\mathbf{Y}'_{i*}\right) \neq z_i\right), \tag{19}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $N' = \sum_i \mathbf{1}(f_{c'}(\mathbf{Y}_{i*}) = z_i)$ denotes the number of correct labels aggregated from normal ones with $f_{c'}$. And $\sum_i \mathbf{1}\left(f_{c'}\left(\mathbf{Y}_{i*}\right) = z_i \wedge f_{c'}^{\lambda_c}\left(\mathbf{Y}'_{i*}\right) \neq z_i\right)$ denotes that among the N' instances, the number of incorrect labels aggregated from all the labels by using $f_{c'}^{\lambda_c}$.

B.3. Discussion of the optimality

Our optimization objective function is non-convex for which it is (computationally) infeasible to find the global optimum; however, in practice a gradient-based optimization algorithm allows us to find local minima which are "good enough" as shown in experiments. (Note this is similar to training a general deep learning model.) We provide the theorems 4 and 5 in A.5. of the supplementary materials to derive the gradient. We set parameter λ as

$$\frac{\tilde{N} * 2^{\tilde{N}} * \mathrm{sigmoid}\left(\sum_k \sigma(\mathbf{W}_k^{i,k}) * \ln^{\sigma(\mathbf{W}^{i,k})}\right)}{M'} \tag{20}$$

where k denotes the index of class of a instance, i denotes the index of the instance, $\tilde{N}$ denote the proportion of instances labeled by malicious workers, M' denotes the number of malicious workers.