A Interventional Expectation satisfying Neyman-Orthogonality

In graphical causal models, causal inference is fulfilled with do-operator to intervene on treatments t, thanks to Pearl *et al.* [6]. For the inference, interventional probability $p(y \mid \mathrm{do}(\mathcal{T} = t))$ is introduced to eliminate a linking edge from covariates x to treatments t and capture the probability of outcomes y for considering all possible covariates x given fixed treatments t, which can be written as follows:

$$p(y \mid \mathrm{do}(\mathcal{T} = t)) = \sum_{x \in \mathcal{X}} p(y \mid x, t) p(x). \tag{1}$$

To utilize it for all data samples, we need to calculate the interventional expectation for Eq. (1) as follows:

$$\begin{aligned} \mathbb{E}[y \mid \mathrm{do}(\mathcal{T} = t)] &= \sum_{y \in \mathcal{Y}} y \times p(y \mid \mathrm{do}(\mathcal{T} = t)) \\ &= \sum_{y \in \mathcal{Y}} y \times \sum_{x \in \mathcal{X}} p(y \mid x, t) p(x) \quad (\because \text{Eq. (1)}) \\ &= \sum_{x \in \mathcal{X}} \left[\sum_{y \in \mathcal{Y}} y \times p(y \mid x, t) \right] p(x) \\ &= \sum_{x \in \mathcal{X}} \mathbb{E}[y \mid x, t] p(x) \\ &= \mathbb{E}_{x \in \mathcal{X}} \left[\mathbb{E}[y \mid x, t] \right]. \end{aligned} \tag{2}$$

To perform double machine learning (DML), Eq. (2) necessarily satisfies Neyman-Orthogonality to robustly predict causal parameter θ despite erroneous nuisance parameters f and g. Therefore, we expand two more different versions of the interventional expectation to satisfy Neyman-Orthogonality. The following formulation represents one version of it and this one is related to inverse probability weighting (IPW) that helps to rebalance the importance of outcomes y:

$$\begin{aligned} \mathbb{E}[y \mid \mathrm{do}(\mathcal{T} = t)] &= \sum_{x \in \mathcal{X}} \mathbb{E}[y \mid x, t] \frac{p(t \mid x)}{p(t \mid x)} p(x) \\ &= \sum_{x \in \mathcal{X}} \left[\sum_{y \in \mathcal{Y}} y \times p(y \mid x, t) \right] \frac{p(x, t)}{p(t \mid x)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y \frac{p(x, y, t)}{p(t \mid x)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{t' \in \mathcal{T}} \mathbb{1}(t' = t) y \frac{p(x, y, t')}{p(t' \mid x)} \\ &= \mathbb{E}_{x \in \mathcal{X}, y \in \mathcal{Y}, t' \in \mathcal{T}} \left[y \frac{\mathbb{1}(t' = t)}{p(t' \mid x)} \right], \end{aligned} \tag{3}$$

where $\mathbb{1}$ denotes a indicator function, which is activated to one if and only if when t' equals to t, otherwise turned off to zero. Another version for the interventional expectation is a modified version of IPW to rebalance prediction outputs $\mathbb{E}[y \mid x, t]$ get from a nuisance parameter f, instead of directly rebalancing outcomes y. This version of formulation can be written as follows:

$$\begin{aligned}\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)] &= \sum_{x\in\mathcal{X}} \mathbb{E}[y \mid x, t]\frac{p(t \mid x)}{p(t \mid x)}p(x) \\ &= \sum_{x\in\mathcal{X}} \mathbb{E}[y \mid x, t]\frac{p(x, t)}{p(t \mid x)} \\ &= \sum_{x\in\mathcal{X}}\sum_{t'\in\mathcal{T}} \mathbb{E}[y \mid x, t']\frac{\mathbb{1}(t'=t)}{p(t' \mid x)}p(x, t') \\ &= \mathbb{E}_{x\in\mathcal{X}, t'\in\mathcal{T}}\left[\mathbb{E}[y \mid x, t']\frac{\mathbb{1}(t'=t)}{p(t' \mid x)}\right].\end{aligned} \tag{4}$$

Next, we combine the three equations Eq.(2-4) to newly represent the interventional expectation, which can be written as follows:

$$\begin{aligned}&\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)] \\ &= \underbrace{\mathbb{E}_{x\in\mathcal{X}}\left[\mathbb{E}[y \mid x, t]\right]}_{\mathbb{E}[y|\mathrm{do}(\mathcal{T}=t)]} + \underbrace{\mathbb{E}_{x\in\mathcal{X}, y\in\mathcal{Y}, t'\in\mathcal{T}}\left[y\frac{\mathbb{1}(t'=t)}{p(t' \mid x)}\right]}_{\mathbb{E}[y|\mathrm{do}(\mathcal{T}=t)]} - \underbrace{\mathbb{E}_{x\in\mathcal{X}, t'\in\mathcal{T}}\left[\mathbb{E}[y \mid x, t']\frac{\mathbb{1}(t'=t)}{p(t' \mid x)}\right]}_{\mathbb{E}[y|\mathrm{do}(\mathcal{T}=t)]} \\ &= \mathbb{E}_{x\in\mathcal{X}, y\in\mathcal{Y}, t'\in\mathcal{T}}\left[\mathbb{E}[y \mid x, t] + (y - \mathbb{E}[y \mid x, t'])\frac{\mathbb{1}(t'=t)}{p(t' \mid x)}\right],\end{aligned} \tag{5}$$

where $\mathbb{E}[y \mid x, t]$ equals to the prediction outputs $f(x, t)$ in non-parametric settings of DML such that:

$$\mathbb{E}[y \mid x, t] = \mathbb{E}[f(x, t) + u \mid x, t] = f(x, t) \quad (\because \mathbb{E}[u \mid x, t] = 0). \tag{6}$$

Finally, we can practically compute the interventional expectation as follows:

$$\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)] = \mathbb{E}_{\mathcal{D}_t}\left[f(x, t) + \frac{y - f(x, t)}{p(\mathcal{T}=t \mid x)}\right] \tag{7}$$

To validate its satisfactory of Neyman-Orthogonality, we introduce a concrete definition of it:

Theorem 1 ***Neyman-Orthogonality***
If two nuisance parameters $\eta = (f, g)$ satisfy the property of "Orthogonal Estimand" for an arbitrary function ϕ as follows:

$$d\phi(\eta_0)\{\eta - \eta_0\} = \nabla_{h=0}\phi(\eta_0 + (\eta - \eta_0)h) = 0,$$

where $d\phi$ denotes Gateaux derivative and $\eta_0 = (f_0, g_0)$ indicates the true nuisance parameters, then the function ϕ satisfies Neyman-Orthogonality.

Following it, we conduct Gateaux derivative of Eq. (7) with $\phi = \mathbb{E}[y \mid \text{do}(\mathcal{T} = t)]$ where it satisfies $p(\mathcal{T} = t \mid x) = p(t - g(x))$, of which formulation can be illustrated as:

$$\begin{aligned}
&\nabla_{h=0}\mathbb{E}[y \mid \text{do}(\mathcal{T} = t); \eta_0 + (\eta - \eta_0)h] \\
&\qquad = \nabla_{h=0}\mathbb{E}_{\mathcal{D}_t}\left[f_0 + (f - f_0)h + \frac{y - f_0 - (f - f_0)h}{p(t - g_0 - (g - g_0)h)}\right] \\
&\qquad = \mathbb{E}_{\mathcal{D}_t}\left[(f - f_0) + \frac{-(f - f_0)p(t - g_0) + (y - f_0)p'(t - g_0)(g - g_0)}{\{p(t - g_0)\}^2}\right] \\
&\qquad = \mathbb{E}_{\mathcal{D}_t}\left[(f - f_0) - \frac{(f - f_0)}{p(t - g_0)}\right] \quad (\because y \approx f_0) \\
&\qquad = \mathbb{E}_{\mathcal{D}_t}\left[(f - f_0) - (f - f_0)\right] \quad (\because t \approx g_0 \to p(t - g_0) = 1) \\
&\qquad = 0.
\end{aligned} \tag{8}$$

From this result, Eq. (7) satisfies Orthogonal Estimand, hence we demonstrate the interventional expectation satisfies Neyman-Orthogonality according to Theorem 1.

B Mathematical Verification of ADML

Beyond the intuition, we strictly verify the reasonable justification for identifiability of (a) our problem setup and (b) our causal parameter in ADML, because the original problem setup and the technique of estimating causal parameter in double machine learning (DML) are somewhat altered to fit adversarial problem setup and estimate adversarially causal parameter.

B.1 Indentifiability of Our Problem Setup

As described in section 3.2 at our manuscript, once we use Taylor expansion for scalar value function $f : \mathbb{R} \to \mathbb{R}$ and decompose it by its input component x and t as:

$$f(x + t) = f(x) + \sum_{i=1}^{\infty} \frac{f^{(i)}(x)}{i!} t^i \tag{9}$$

where $f^{(i)}$ indicates i-th order derivative function, we can express partially linear settings for a nuisance parameter f as follows:

$$y = \underbrace{f(x) + \theta\bar{t}}_{f(x+t)} + u, \quad (\mathbb{E}[u \mid x, t] = 0) \tag{10}$$

where $\bar{t}$ indicates Taylor-order matrix: $[t, t^2, \cdots]^T$ and θ represents Taylor-coefficient matrix $[\frac{f^{(1)}(x)}{1!}, \frac{f^{(2)}(x)}{2!}, \cdots]$. Here, we can ask the **(a) exact dimensions** of θ and $\bar{t}$, and wonder if Eq. (10) is a proper expression for the vector-valued functions f with **(b) simple multiplication** $\theta\bar{t}$, instead of expression with sequential vector/matrix multiplications such as $t^T \frac{f^{(2)}(x)}{2!} t$.

(a) Exact dimensions of θ and $\bar{t}$. Once we limit the Taylor-order to large enough n, the Taylor-coefficient matrix θ and Taylor-order matrix $\bar{t}$ can be written as: $[\frac{f^{(1)}(x)}{1!}, \frac{f^{(2)}(x)}{2!}, \cdots, \frac{f^{(n)}(x)}{n!}]$ and $[t, t^2, \cdots, t^n]^T$, respectively. When clean images $x \in \mathbb{R}^{hwc}$ and (one-hot) target classes $y \in \mathbb{R}^d$ are given with vector-valued function $f : \mathbb{R}^{hwc} \to \mathbb{R}^d$, it satisfies $\frac{f^{(i)}(x)}{i!} \in \mathbb{R}^{d \times (hwc)^i}$ and $t^i \in \mathbb{R}^{(hwc)^i}$ where we concretely explain what $t^i \in \mathbb{R}^{(hwc)^i}$ represents in next paragraph. In conclusion, the dimension of θ is $\mathbb{R}^{d \times hwc\frac{(hwc)^n - 1}{hwc - 1}}$, the dimension of $\bar{t}$ is $\mathbb{R}^{hwc\frac{(hwc)^n - 1}{hwc - 1}}$, and the dimension of their multiplication $\theta\bar{t}$ is $\mathbb{R}^d$.

(b) Simple multiplication: $\theta\bar{t}$. For better understanding, we show the following examples for first-order and second-order derivative of function $f : \mathbb{R}^{hwc=2} \rightarrow \mathbb{R}^{d=1}$ for the given $n = 2$ as:

$$\frac{f^{(1)}(x)}{1!} = \begin{bmatrix} a & b \end{bmatrix}, \quad \frac{f^{(2)}(x)}{2!} = \begin{bmatrix} c & d \\ e & f \end{bmatrix}, \quad t = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}. \tag{11}$$

Then, the following Taylor expansion of $f(x+t) \in \mathbb{R}$ with the above notations is normally used as:

$$f(x+t) = f(x) + \underbrace{\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}}_{\frac{f^{(1)}(x)}{1!}t} + \underbrace{\begin{bmatrix} t_1 & t_2 \end{bmatrix} \begin{bmatrix} c & d \\ e & f \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}}_{t^T \frac{f^{(2)}(x)}{2!}t}. \tag{12}$$

To explicitly decompose causal parameter θ in this Taylor expansion, we illustrate a more efficient way of representing Taylor expansion with simple multiplication despite vector-valued functions f, which can be written as follows:

$$\begin{aligned} f(x+t) &= f(x) + \underbrace{\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}}_{\frac{f^{(1)}(x)}{1!}t} + \underbrace{\begin{bmatrix} c & d & e & f \end{bmatrix} \begin{bmatrix} {t_1}^2 \\ t_1 t_2 \\ t_2 t_1 \\ {t_2}^2 \end{bmatrix}}_{\frac{f^{(2)}(x)}{2!}t^2} \\ &= f(x) + \underbrace{\begin{bmatrix} a & b & c & d & e & f \end{bmatrix}}_{\theta} \underbrace{\begin{bmatrix} t_1 \\ t_2 \\ {t_1}^2 \\ t_1 t_2 \\ t_2 t_1 \\ {t_2}^2 \end{bmatrix}}_{\bar{t}}. \end{aligned} \tag{13}$$

Here, the matrix dimension of $\frac{f^{(2)}(x)}{2!}$ is implicitly changed to be a flattened vector and t^2 indicates the vector of all multiplication combinations for the elements t_1 and t_2 in t. Once we increase the output dimension of function from $d = 1$ to $d = 2$ such that $f : \mathbb{R}^{hwc=2} \rightarrow \mathbb{R}^{d=2}$ for the given $n = 2$, we simply add one row vector to matrix θ in Eq. (13) as follows:

$$f(x+t) = f(x) + \underbrace{\begin{bmatrix} a_1 & b_1 & c_1 & d_1 & e_1 & f_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 & f_2 \end{bmatrix}}_{\theta} \underbrace{\begin{bmatrix} t_1 \\ t_2 \\ {t_1}^2 \\ t_1 t_2 \\ t_2 t_1 \\ {t_2}^2 \end{bmatrix}}_{\bar{t}}. \tag{14}$$

In this way, we can explicitly decompose the causal parameter θ from $f(x+t) \in \mathbb{R}^2$, thus we demonstrate using the simple multiplications is sufficient to represent Taylor expansion even with vector-valued function and show that Eq. (10) is the workable expression for our problem setup.

B.2 Indentifiability of Our Causal Parameter

As described in equation 8 at our manuscript, we have depicted the way of estimating causal parameter in our adversarial problem setup, again which can be written as follows:

$$\hat{\theta} = \mathbb{E}_{\mathcal{D}_t}\left[-\left(\frac{1}{p(\mathcal{T} = t \mid x)} - 1\right) \frac{\partial}{\partial t} f(x+t)\right]. \tag{15}$$

When we look at this equation more closely through Taylor expansion, Eq. (15) can be written as follows:

$$\hat{\theta} = \mathbb{E}_{\mathcal{D}_t}\left[-\left(\frac{1}{p(\mathcal{T} = t \mid x)} - 1\right) \theta \frac{\partial}{\partial t} \bar{t}\right]. \tag{16}$$

However, the dimension of θ is actually infinite, therefore we consider $\theta\frac{\partial}{\partial t}\bar{t}$ as a controllable part of causal parameter, because its dimension is as follows:

$$\theta \in \mathbb{R}^{d\times hwc\frac{(hwc)^n-1}{hwc-1}}, \quad \frac{\partial}{\partial t}\bar{t} \in \mathbb{R}^{hwc\frac{(hwc)^n-1}{hwc-1}\times hwc} \quad \Rightarrow \quad \theta\frac{\partial}{\partial t}\bar{t} \in \mathbb{R}^{d\times hwc}. \tag{17}$$

According to the controllable part: $\frac{\partial}{\partial t}f(x+t) = \theta\frac{\partial}{\partial t}\bar{t}$, we can finally estimate causal parameter θ of adversarial vulnerability, while replacing it with $\hat{\theta}$ from which we can catch how worst perturbations t affect network predictions for target classes y.

C Mild Assumption for $\mathbb{E}[u \mid x, t] = 0$

Once the conditional expected error of u for the given covariates x and treatments t equals to zero, then the expected error given covariates x also equals to zero, as follows:

$$\mathbb{E}[u \mid x, t] = 0 \quad \Rightarrow \quad \mathbb{E}[u \mid x] = \mathbb{E}_{t|x}\left[\mathbb{E}[u \mid x, t]\right] = 0. \tag{18}$$

Here, if we apply normalized weighted geometric mean (NWGM) approximation normally used to address computational issue in deep learning [11, 8, 4] to Eq. (18) for treatments t, the following formulation satisfies:

$$\underbrace{\mathbb{E}_{t|x}\left[\mathbb{E}[u \mid x, t]\right]}_{\mathbb{E}[u|x]} = 0 \quad \Rightarrow \quad \mathbb{E}[u \mid x, \mathbb{E}[t \mid x]] = \mathbb{E}[u \mid x, g(x)] = 0. \tag{19}$$

Consequently, we can replace a strong assumption of $\mathbb{E}[u|x,t] = 0$ with a mild assumption of $\mathbb{E}[u|x, g(x)] = 0$ by using $\mathbb{E}[u|x] = 0$.

Alternatively, we can explain the mild assumption from a different perspective of both partially linear and non-parametric settings. For partially linear settings, specifically, the expected outcomes y for the given covariates x can be written in Robinson-style [7] as follows:

$$\begin{aligned}\mathbb{E}[y \mid x] &= \mathbb{E}[f(x) + \theta\bar{t} + u \mid x] \\ &= f(x) + \theta\mathbb{E}[\bar{t} \mid x] \quad (\because \mathbb{E}[u \mid x] = 0) \\ &= f(x) + \theta\bar{g}(x) = f(x + g(x)) \quad (\because \text{Taylor Expansion}),\end{aligned} \tag{20}$$

where we suppose $\bar{t} = \bar{g}(x) + \bar{v}$ such that $\mathbb{E}[\bar{t} \mid x] = \bar{g}(x)$. To the next, non-parametric settings have a disparate viewpoint of $f(x,t) = f(x+t)$, which can be written as follows:

$$\begin{aligned}\mathbb{E}[y \mid x] &= \mathbb{E}[f(x,t) + u \mid x] \\ &= \mathbb{E}_{t|x}[f(x+t)] \approx f(x+t).\end{aligned} \tag{21}$$

Here, we can feasibly approximate $\mathbb{E}_{t|x}[f(x+t)] \approx f(x+t)$ because we newly generate one minibatch of worst perturbation t corresponding to one minibatch of x in every training iteration, and we do not generate multiple minibatches of worst perturbation t with a fixed minibatch of x. Incorporating Eq. (20) and Eq. (21), we can finally conclude the following formulations:

$$f(x + g(x)) \approx f(x+t) \quad \Rightarrow \quad \mathbb{E}[y - f(x+g(x)) \mid x] \approx \mathbb{E}[\underbrace{y - f(x+t)}_{u} \mid x] = 0. \tag{22}$$

For that reason, a mild assumption of $\mathbb{E}[u \mid x, g(x)] = 0$ can be also satisfied by using $\mathbb{E}[u \mid x] = 0$.

D Realization of Causal Parameter in ADML

We apply the concept of "additive noise" in adversarial problem setup to the interventional expectation Eq. (7) satisfying Neyman-Orthogonality, which can be written as follows:

$$\mathbb{E}[y \mid \mathrm{do}(\mathcal{T} = t)] = \mathbb{E}_{\mathcal{D}_t}\left[f(x+t) + \frac{y - f(x+t)}{p(\mathcal{T} = t \mid x)}\right]. \tag{23}$$

Based on the equation 5 at our manuscript, we can compute causal parameter with high-dimensional treatments t on non-parametric settings by differentiating Eq. (23) with respect to t as:

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}[y \mid \text{do}(\mathcal{T}=t)] &= \frac{\partial}{\partial t}\mathbb{E}_{\mathcal{D}_t}\left[f(x+t)+\frac{y-f(x+t)}{p(\mathcal{T}=t\mid x)}\right] \\ &= \mathbb{E}_{\mathcal{D}_t}\left[\frac{\partial}{\partial t}f(x+t)+\frac{-\frac{\partial}{\partial t}f(x+t)}{p(\mathcal{T}=t\mid x)}\right] \\ &= \mathbb{E}_{\mathcal{D}_t}\left[-\left(\frac{1}{p(\mathcal{T}=t\mid x)}-1\right)\frac{\partial}{\partial t}f(x+t)\right].\end{aligned} \tag{24}$$

In this way, it enables us to numerically estimate causal parameter in ADML.

E Approximation of $p(\mathcal{T}=t\mid x)$

To practically handle $p(\mathcal{T}=t\mid x)$ to calculate causal parameter θ in ADML, we need to approximate its probability with marginalization as follows:

$$\begin{aligned}p(\mathcal{T}=t\mid x) &= \sum_{y'\in\mathcal{Y}} p(y',\mathcal{T}=t\mid x) \\ &= \sum_{y'\in\mathcal{Y}} p(\mathcal{T}=t\mid x,y')p(y'\mid x) \\ &= \sum_{y'\in\mathcal{Y}} p(\mathcal{T}=t\mid x,y')\sum_{t'\in\mathcal{T}} p(y',t'\mid x) \\ &= \sum_{y'\in\mathcal{Y}} p(\mathcal{T}=t\mid x,y')\sum_{t'\in\mathcal{T}} p(y'\mid x,t')p(t'\mid x) \\ &= \sum_{y'\in\mathcal{Y}} p(\mathcal{T}=t\mid x,y')\mathbb{E}_{t'|x}[p(y'\mid x,t')]\end{aligned} \tag{25}$$

where $p(\mathcal{T}=t\mid x,y')$ illustrates the probability of owning worst perturbations in our hand for the given clean images x and specific classes y'. Here, we use the sharpening technique for $p(\mathcal{T}=t\mid x,y')$ to obtain a one-hot vector, where the position of the activated value represents the attacked class y_a for which the worst perturbation t changes the network prediction from y. By using the technique, $p(\mathcal{T}=t\mid x,y')$ can be written as follows:

$$p(\mathcal{T}=t\mid x,y') = \begin{cases} 1 & (y'=y_a) \\ 0 & (y'\neq y_a)\end{cases} \tag{26}$$

The reason why it is possibly fitted to our problem setup is that we newly generate one minibatch of worst perturbation t corresponding to one minibatch of x in every training iteration and we discard the perturbation not harming network predictions at each training iteration so that we deal only with the collection of perturbation breaking network predictions. Therefore, we can approximate the distribution of worst perturbation $p(\mathcal{T}=t|x)$ with the expected attacked confidence as follows:

$$p(\mathcal{T}=t\mid x)\approx\mathbb{E}_{t'|x}[p(y_a\mid x,t')]. \tag{27}$$

Similar to Eq. (21), we feasibly compute $\mathbb{E}_{t'|x}[p(y_a|x,t')]\approx f_{j^*}(x+t)$ such that $j^*=\arg\max_j f_j(x+t)$, where j is a class index.

F Building Objective Function for ADML

To harmonize the interventional expectation with deep learning, we redesign it to build the loss function used in deep learning as follows:

$$\begin{aligned}
\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)] &= \mathbb{E}_{\mathcal{D}_t}\left[f(x+t) + \frac{y - f(x+t)}{p(\mathcal{T}=t \mid x)}\right] \\
&= \mathbb{E}_{\mathcal{D}_t}\left[f(x+t) - y + y + \frac{y - f(x+t)}{p(\mathcal{T}=t \mid x)}\right] \\
&= \mathbb{E}_{\mathcal{D}_t}\left[\left(\frac{1}{p(\mathcal{T}=t \mid x)} - 1\right)\{y - f(x+t)\} + y\right] \\
&\approx \mathbb{E}_{\mathcal{D}_t}\left[\left(\frac{1}{p(\mathcal{T}=t \mid x)} - 1\right)\mathcal{L}_{\mathrm{CE}}(f(x+t), y) + y\right]
\end{aligned} \tag{28}$$

Through the above equation, we can calculate $\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=0)]$ as:

$$\begin{aligned}
\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=0)] &\approx \mathbb{E}_{\mathcal{D}_0}\left[\left(\underbrace{\cancel{\frac{1}{p(\mathcal{T}=0 \mid x)}}}_{\textit{ignored}} -1\right)\mathcal{L}_{\mathrm{CE}}(f(x), y) + y\right] \\
&\approx \mathbb{E}_{\mathcal{D}_0}\left[-\mathcal{L}_{\mathrm{CE}}(f(x), y) + y\right],
\end{aligned} \tag{29}$$

where we posit $p(\mathcal{T}=0 \mid x) = 0$ because $\mathcal{T}=0$ cannot exist among generated perturbations by an adversarial attack of PGD which is a perturbation generator g. Accordingly, we *ignore its inverse*: $1/p(\mathcal{T}=0 \mid x)$ and calculate the other terms only. Then, we can figure out the difference between $\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)]$ and $\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=0)]$ to approximate the partial derivative of $\frac{\partial}{\partial t}\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)]$ as follows:

$$\mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=t)] - \mathbb{E}[y \mid \mathrm{do}(\mathcal{T}=0)] = \mathbb{E}_{\mathcal{D}_t}\left[\tau\mathcal{L}_{\mathrm{CE}}(f(x+t), y)\right] + \mathbb{E}_{\mathcal{D}_0}\left[\mathcal{L}_{\mathrm{CE}}(f(x), y)\right], \tag{30}$$

where balancing ratio τ represents $\frac{1}{p(\mathcal{T}=t|x)} - 1$. In such manner, we can reasonably estimate causal parameters of adversarial vulnerability reconciling with deep learning.

G Power of Sample-Splitting and Cross-Fitting

The effectiveness of ADML stems from the utilization of causal parameter based on sample-splitting plus cross-fitting in two distinct parts with split samples: (i) Mild assumption, (ii) Mitigating causal parameter, as described in Algorithm 1 (Line 4 and 8, each) at our main paper. Performing (i) represents previous AT-based defenses loss, while performing (ii) reweights adversarial loss function to prioritize more vulnerable samples. These split samples possibly cross checks whether to fit the two parts without sharing data samples, thereby preventing models from the excessive adaptation of the AT-based defense loss alone and avoiding overfitting to specific samples or classes.

From these aspects, interestingly, ADML can no longer be considered as a variation of AT. Instead, it is designed to address biased predictions induced by overfitting [1]. Therefore, we need ADML to prevent excessive focus on optimizing the AT-based defense loss, as there exists the risk of overfitting without ADML in Fig. 1 at our main paper. This distinction highlights our unique contributions in addressing the limitations of the existing works.

Sample-splitting ratio. We conduct an ablation study for the sample-splitting ratio on adversarial robustness and its consequential outcomes are presented in the following table for CIFAR-10, where $\mathcal{S}_{\mathcal{D}_1}$ denotes the sample-splitting ratio for Line 4 (Mild assumption) in Algorithm 1, and $\mathcal{S}_{\mathcal{D}_2}$ denotes the ratio for Line 8 (Causal parameter estimation). From the following table, we observe that the highest adversarial robustness is achieved when using half sample-splitting ratio. Therefore, we choose half for our experiments to ensure the best robustness.

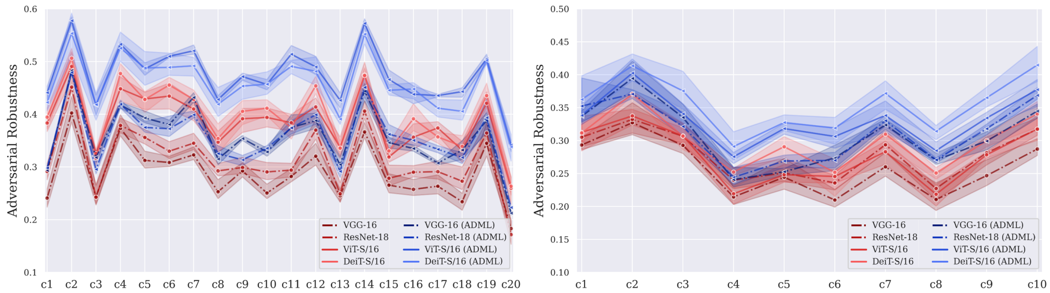
Time complexity. Additionally, we count the training time per epoch as a measure of empirical time complexity. In Algorithm 1, we compute the gradient only once at each iteration in Line 10, which is consistent with previous AT-based defense methods. Therefore, ADML ensures a fair comparison without any extra cross-fitting burden. The slight time differences in the following table arise from computing independent two losses described in Line 4 and 8, which could be easily addressed by integrating parallel processing with deep learning library. Hence, We believe these differences do not pose computational challenges and do not bring in unfair settings.

	VGG-16							*ResNet-18*						
$\mathcal{S}_{\mathcal{D}_1}$	0	0.1	0.3	**0.5**	0.7	0.9	1	0	0.1	0.3	**0.5**	0.7	0.9	1
$\mathcal{S}_{\mathcal{D}_2}$	1	0.9	0.7	**0.5**	0.3	0.1	0	1	0.9	0.7	**0.5**	0.3	0.1	0
Clean	79.4	79.4	80.3	**80.9**	80.7	78.3	78.8	84.0	84.3	84.3	**84.5**	84.0	83.6	83.1
PGD	49.8	56.1	60.3	**61.7**	59.4	53.2	48.1	52.3	54.0	58.2	**60.8**	57.5	53.9	51.9
CW_∞	48.5	54.5	58.9	**59.8**	58.9	52.5	46.8	51.9	53.6	57.5	**58.5**	56.8	52.8	50.8
DLR	47.5	52.1	**54.8**	**54.8**	54.7	49.7	46.4	50.4	52.0	55.1	**56.2**	54.5	51.6	49.7
AA	46.8	51.2	53.9	**54.1**	54.0	49.0	46.0	49.8	51.2	54.5	**55.2**	53.8	51.0	49.1
Avg (%)	54.4	58.7	61.6	**62.3**	61.5	56.5	53.2	57.7	59.0	61.9	**63.0**	61.3	58.6	57.0
Time (m)	0.50	0.53	0.53	0.53	0.53	0.53	0.50	0.67	0.75	0.75	0.75	0.75	0.75	0.67

Table 1: Measuring adversarial robustness across sample-splitting ratio from zero to one for each training batch and measuring the training time per epoch across these sample-splitting ratios.

H Distribution of Adversarial Robustness in Larger Datasets

We have validated integrated distribution of four adversarial defense baselines [5, 12, 9, 10] and their corresponding combinatorial ADML models with CIFAR-10 [2] in figure 4 at our manuscript. To provide further demonstrations on the larger datasets with various distribution of common object classes, we expand the infographics into CIFAR-100 [2] and Tiny-ImageNet [3]. For better visualization from disorganization, we cluster each dataset into several groups. Note that, each cluster is the 10% of randomly selected labels for the total number of classes, thus 10 and 20 clusters for CIFAR-100 and Tiny-ImageNet, respectively. The above visualization in Fig. 1 shows averaged adversarial robustness for each cluster. As in the figure, we can clearly observe the disparity of the robustness between baselines and their corresponding ADML models. To sum up, we corroborate the effectiveness of our proposed method to improve the robustness, by presenting the way of computing causal parameter in adversarial examples and mitigating its causal effects.



(a) CIFAR-100

(b) Tiny-ImageNet

Figure 1: Distribution of adversarial robustness across whole classes on larger datasets: CIFAR-100 and Tiny-ImageNet. Four AT-based defense methods: AT [5], TRADES [12], MART [9], AWP [10] are integrated on each architecture. In x-axis, c_k indicates the average robustness for each cluster.

References

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," 2018. 7

[2] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. 8

[3] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, 2015. 8

[4] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao, and J. Zhao, "Show, deconfound and tell: Image captioning with causal inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 041–18 050. 5

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb 8

[6] J. Pearl, *Causality*. Cambridge university press, 2009. 1

[7] P. M. Robinson, "Root-n-consistent semiparametric regression," *Econometrica: Journal of the Econometric Society*, pp. 931–954, 1988. 5

[8] T. Wang, J. Huang, H. Zhang, and Q. Sun, "Visual commonsense r-cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 760–10 770. 5

[9] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rklOg6EFwS 8

[10] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2958–2969, 2020. 8

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015. 5

[12] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 09–15 Jun 2019, pp. 7472–7482. 8