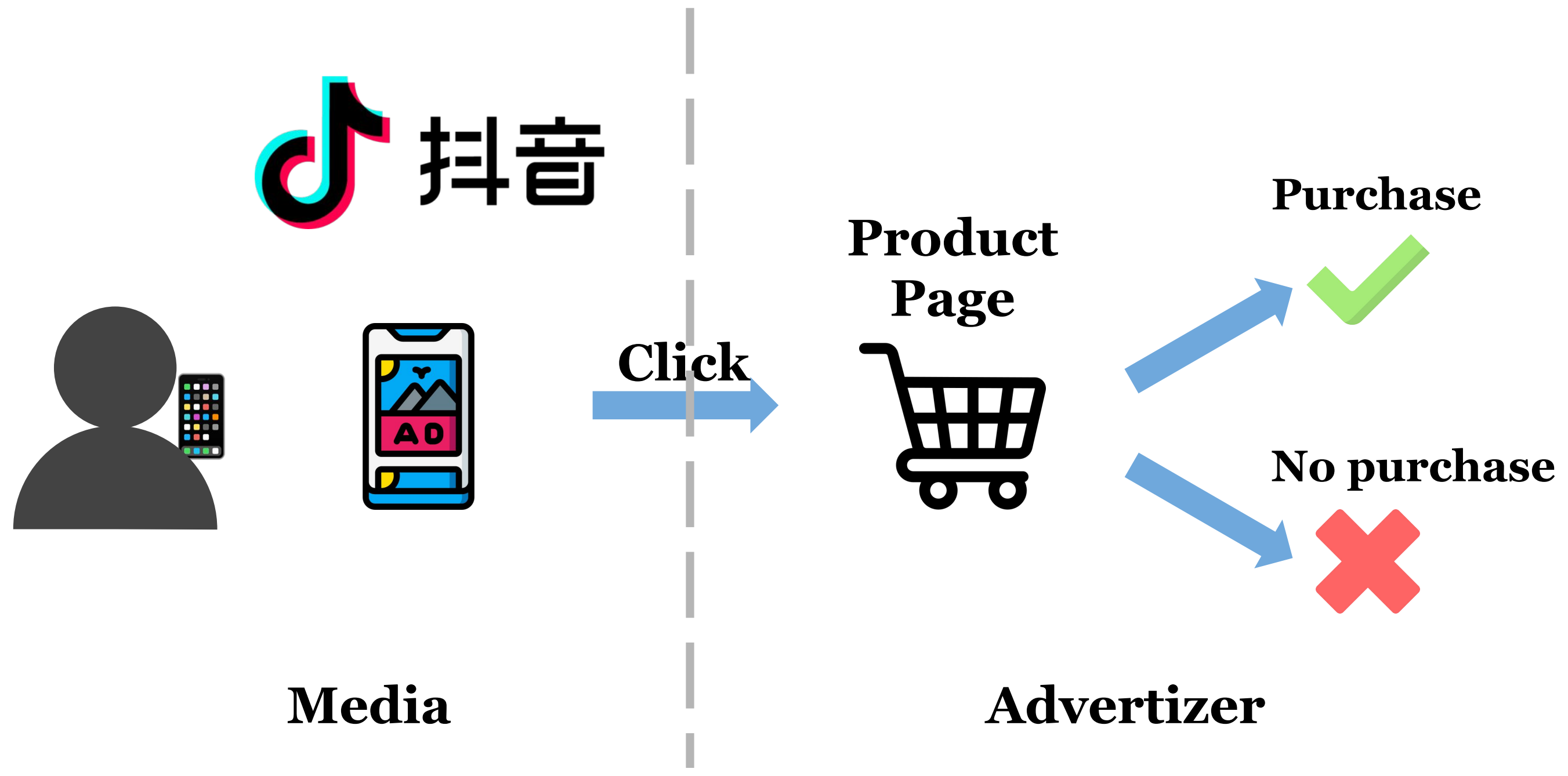
Label Leakage and Protection in Two-party Split Learning

Oscar Li, Jiankai Sun, Weihao Gao, Hongyi Zhang, Xin Yang, Junyuan Xie, Chong Wang

- **Our finding:** Data leakage is a real problem not only in horizontal FL but also in **vertical FL**. Simple method exists to uncover the **label information** through communicated gradients in two-party split learning.
- **Our solution**: A theoretically justified random gradient perturbation method to protect against a large class of label-uncovering methods.

Two party vertical FL example



Two party split learning abstraction

$f: \mathscr{X} \to \mathbb{R}^d$

$L = \log(1+\exp(-\ell)) + (1-y)\ell$

$h: \mathscr{R}^d \to \mathbb{R}$

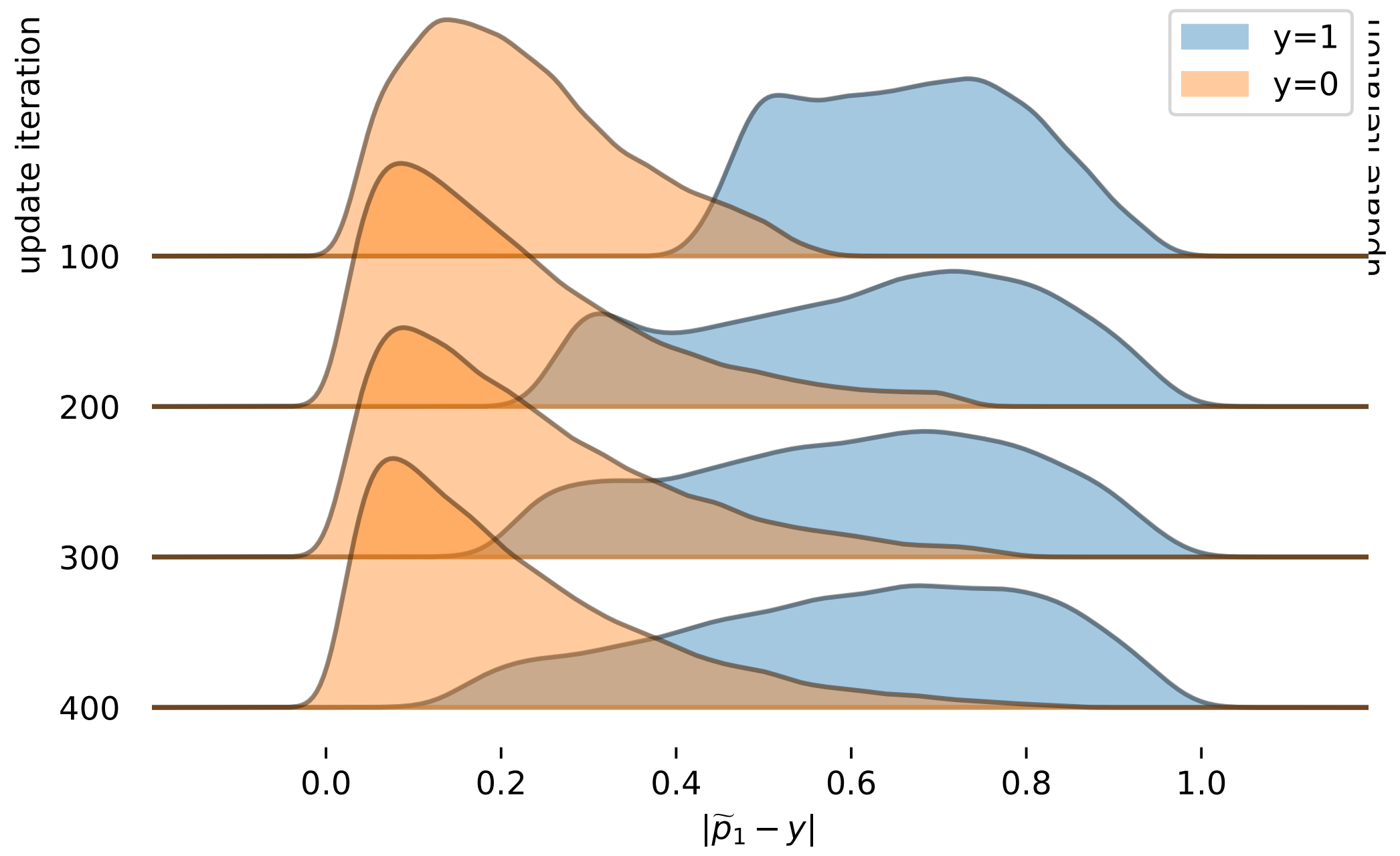
$\widetilde{p}_1 = 1/(1+\exp(-\ell))$

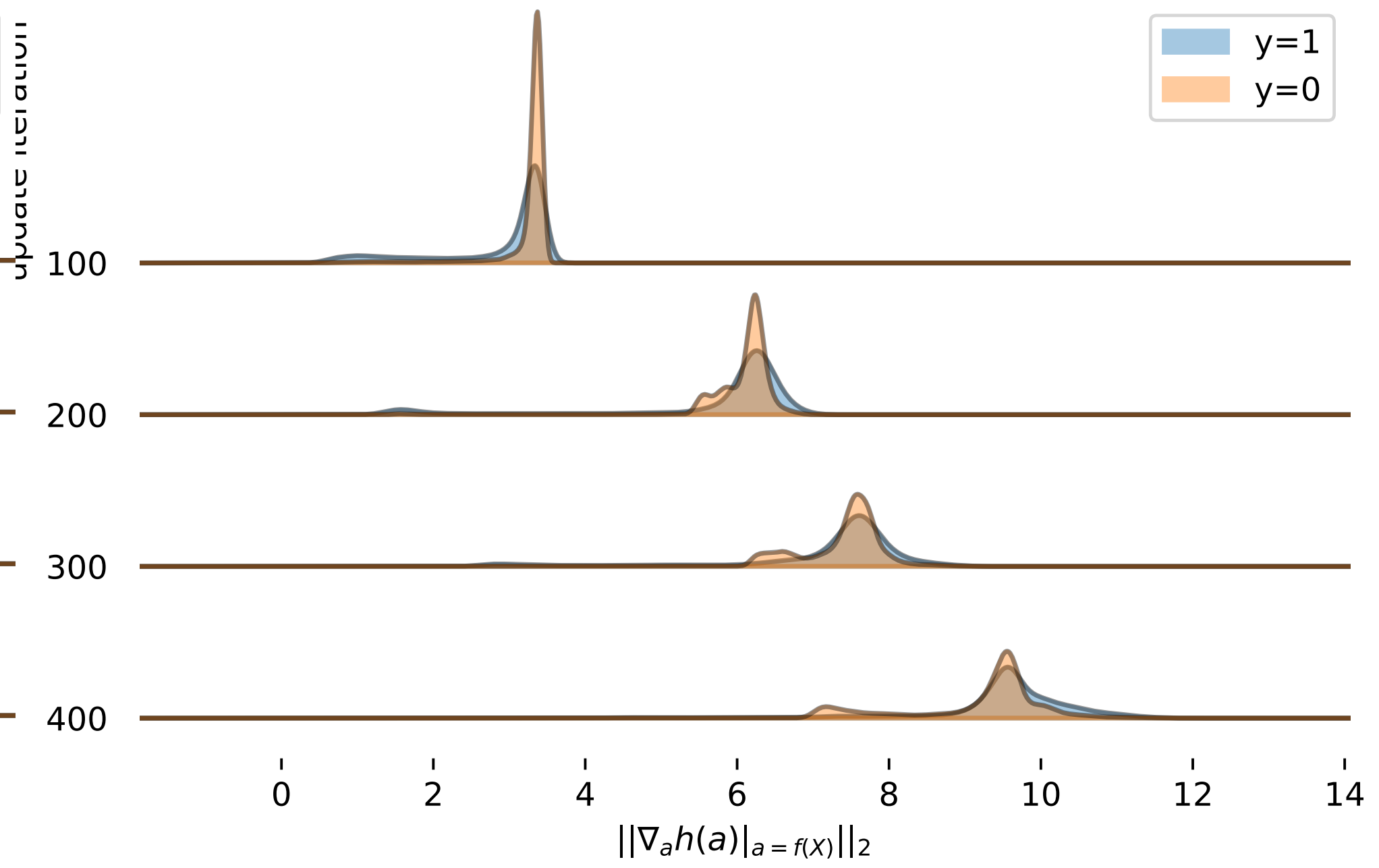
$$\begin{array}{lccccccccc} & \multicolumn{3}{c}{\text{Passive party}} & & \multicolumn{5}{c}{\text{Active party}} \\ \text{Forward:} & X & \xrightarrow{f} & f(X) & \overset{\text{comm.}}{\Longrightarrow} & f(X) & \xrightarrow{h} & \ell = h(f(X)) & \rightarrow & \overset{y \downarrow}{L} \\ \text{Backward:} & & & g & \overset{\text{comm.}}{\Longleftarrow} & g & \leftarrow & \frac{dL}{dl} = (\widetilde{p}_1 - y) & \leftarrow & 1 \\ & & & \downarrow & & & & \downarrow & & \\ & & & f\text{'s param} & & & & h\text{'s param} & & \end{array}$$

Communicated gradient $g \triangleq \nabla_{f(X)} L = (\widetilde{p}_1 - y) \nabla_a h(a) \Big|_{a=f(X)} \in \mathbb{R}^d$

A simple label-uncovering method through gradient norm

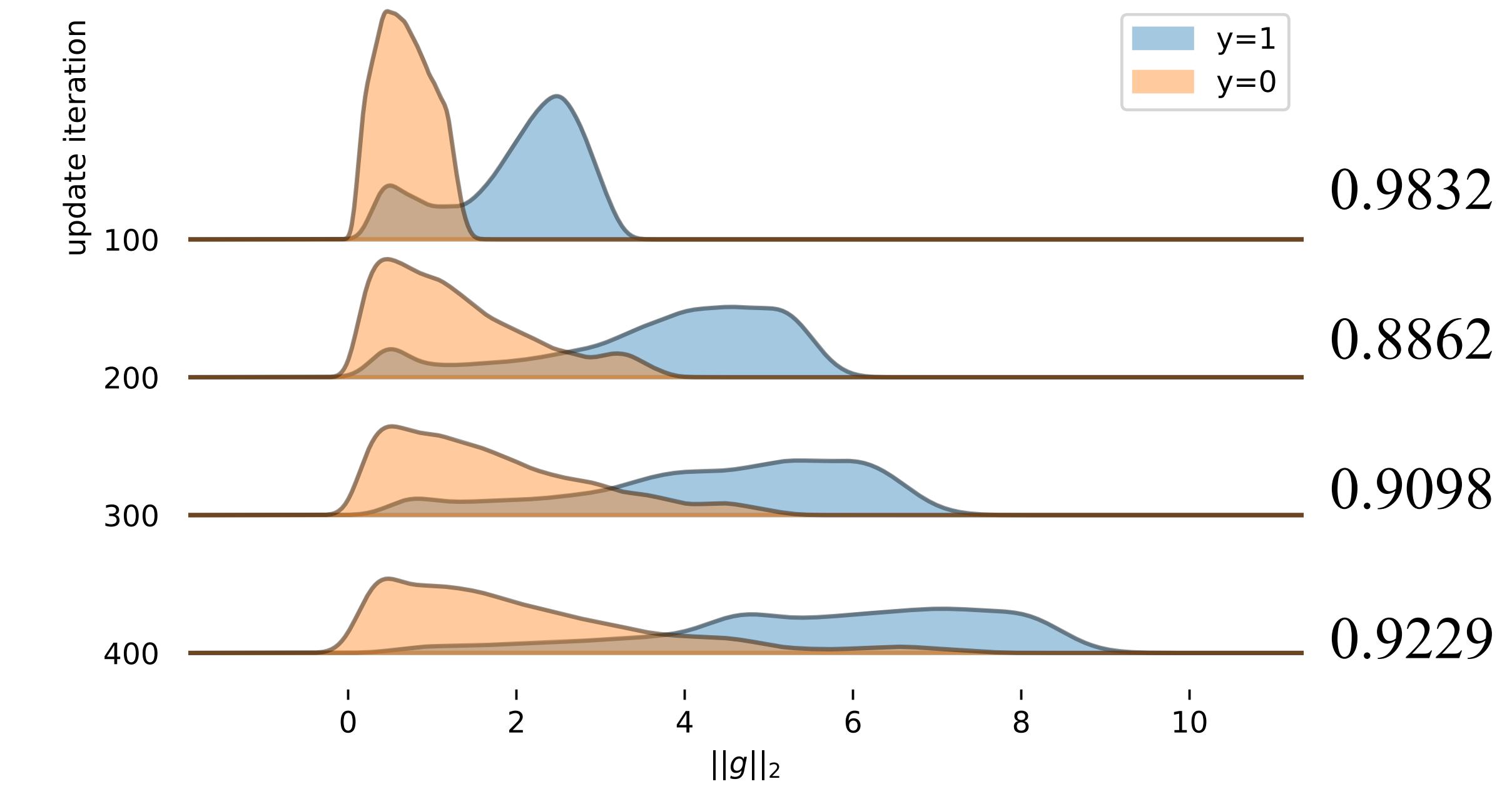
$$\|g\|_2 = |\widetilde{p}_1 - y| \cdot \|\nabla_a h(a)|_{a=f(X)}\|_2$$





Observation 1: less confidence about positive examples ($y = 1$)

Observation 2: same norm magnitude for both classes $\|\nabla_a h(a)|_{a=f(X)}\|_2$



Quantitative measure of the label leak through $\|g\|_2$

Leak AUC: AUC using $\|g\|_2$ to predict y

Protection through randomness

- $\widetilde{g}^{(1)} = g^{(1)} + n^{(1)} \quad (\widetilde{P}^{(1)}); \qquad \mathbb{E}[n^{(1)}] = \mathbf{0}$
- $\widetilde{g}^{(0)} = g^{(0)} + n^{(0)} \quad (\widetilde{P}^{(0)}); \qquad \mathbb{E}[n^{(0)}] = \mathbf{0}$

Formulating protection objective

- General class of label recovering function: $\{1_A, A \subseteq \mathbb{R}^d\}$
- A labelling function 1_A's *detection error:*

$$\frac{1}{2}(\text{FNR} + \text{FPR}) = \frac{1}{2}[\mathbb{P}(1_A(\widetilde{g}^{(1)}) = 0) + \mathbb{P}(1_A(\widetilde{g}^{(0)}) = 1)]$$
$$= \frac{1}{2}(\widetilde{P}^{(1)}(A^C) + \widetilde{P}^{(0)}(A))$$

- Maximize worst-case adversarial passive party's detection error:

$$\max_{\widetilde{P}^{(1)}, \widetilde{P}^{(0)}} \min_A \frac{1}{2}(\widetilde{P}^{(1)}(A^C) + \widetilde{P}^{(0)}(A)) = \max_{\widetilde{P}^{(1)}, \widetilde{P}^{(0)}} \frac{1}{2}(1 - \text{TV}(\widetilde{P}^{(1)}, \widetilde{P}^{(0)}))$$

Pinsker + Jensen inequality: $\text{TV}(P, Q) \le \frac{1}{2}\sqrt{\text{KL}(P\|Q) + \text{KL}(Q\|P)}$

$$\min_{\widetilde{P}^{(1)}, \widetilde{P}^{(0)}} \text{KL}(\widetilde{P}^{(1)} \,\|\, \widetilde{P}^{(0)}) + \text{KL}(\widetilde{P}^{(0)} \,\|\, \widetilde{P}^{(1)})$$
$$\text{s.t.} \quad p \cdot \text{tr}(\text{Cov}[n^{(1)}]) + (1-p) \cdot \text{tr}(\text{Cov}[n^{(0)}]) \le P$$

Optimizing the objective with assumptions

$$g^{(0)} \sim \mathcal{N}(\bar{g}^{(0)}, vI_{d\times d}) \quad g^{(1)} \sim \mathcal{N}(\bar{g}^{(1)}, vI_{d\times d})$$
$$n^{(0)} \sim \mathcal{N}(0, \Sigma_0) \quad n^{(1)} \sim \mathcal{N}(0, \Sigma_1)$$
$$\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$$

Analytical Solution

$$\Sigma_1^* = \frac{\lambda_1^{(1)*} - \lambda_2^{(1)*}}{\|\Delta g\|_2^2}(\Delta g)(\Delta g)^T + \lambda_2^{(1)*} I_d$$
$$\Sigma_0^* = \frac{\lambda_1^{(0)*} - \lambda_2^{(0)*}}{\|\Delta g\|_2^2}(\Delta g)(\Delta g)^T + \lambda_2^{(0)*} I_d$$

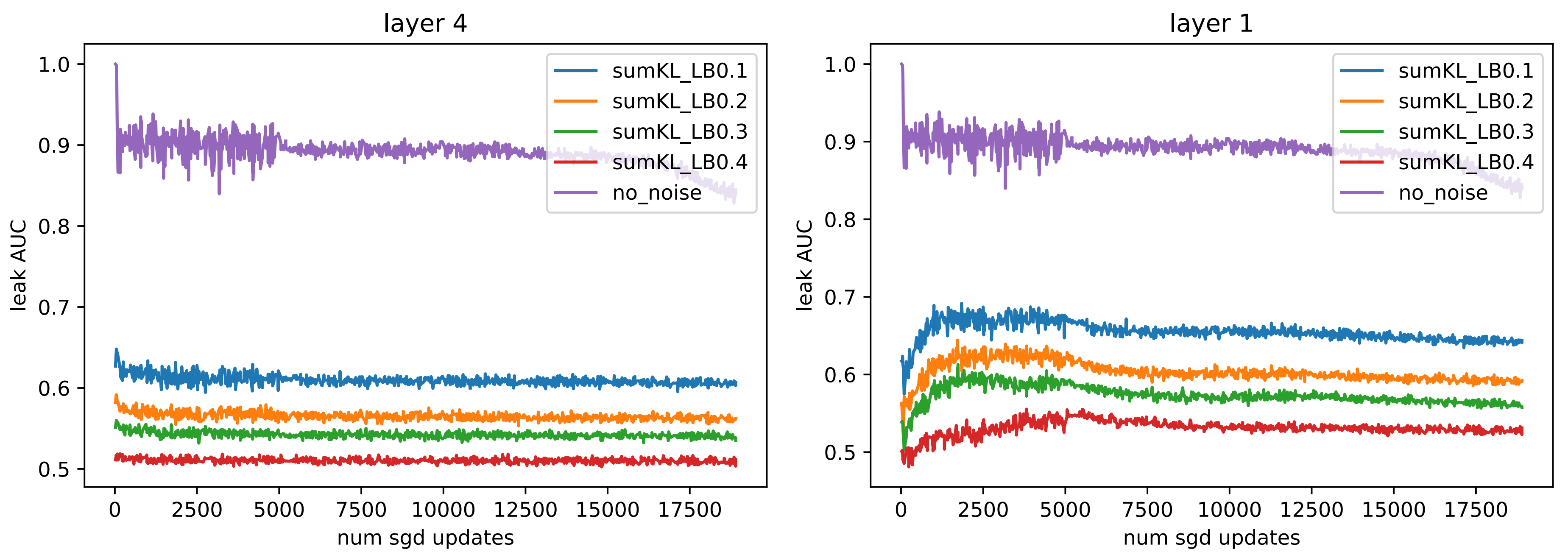
where $\Delta g \triangleq \bar{g}^{(1)} - \bar{g}^{(0)}$.

with $\lambda_1^{(0)*}, \lambda_2^{(0)*}, \lambda_1^{(1)*}, \lambda_2^{(1)*}$ the optimal solution to a 4-variable optimization problem (see paper).

Guarantee:

$$\text{sumKL}^* \le (2 - 4L)^2$$
$$\Downarrow$$
$$\text{detection error } \ge L$$

Experiments (Criteo)



Model: Wide & Deep (embedding layer + 4 128-unit layer MLP)

Isotropic baseline:

$$\mathcal{N}(0, \frac{25 \cdot \max_{i=1}^B \|g_i\|_2^2}{d} I_{d\times d})$$

	no noise	iso	sumKL
train loss	0.4067	0.4700	0.4672
test AUC	0.8062	0.7921	0.7949
	max leak AUC		
layer1	1.0000	0.5536	0.5554
layer2	1.0000	0.5652	0.5583
layer3	1.0000	0.6089	0.5710
layer4	1.0000	0.5129	0.5188