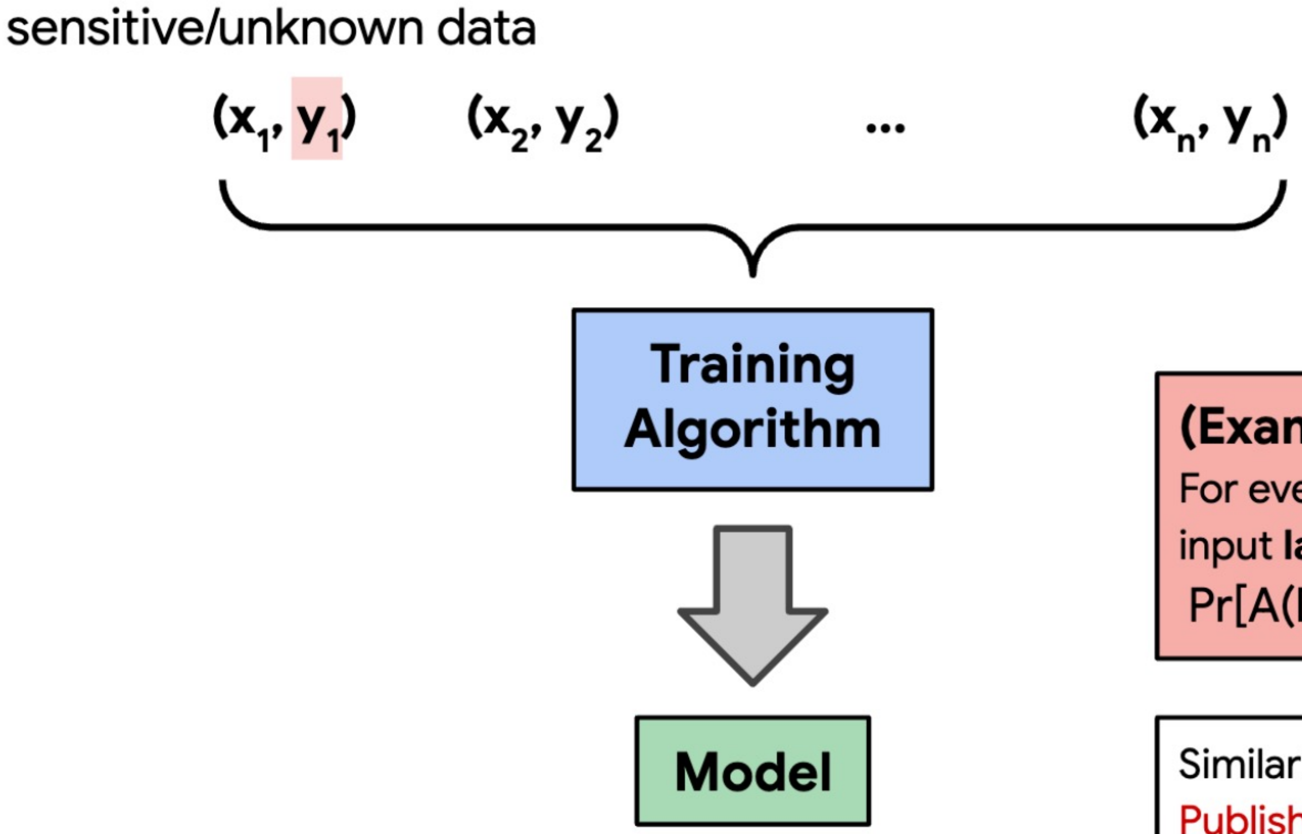
Complementary results on Private Conversion Measurement using label-DP

Maxime Vono, Senior Researcher, Criteo AI Lab
m.vono@criteo.com

Training with *Label* Differential Privacy



(Example-level) ε-Label-DP

For every datasets D, D' that differ by a single input **label** and every output model o,

$\Pr[A(D) = o] \leq e^{\varepsilon} \cdot \Pr[A(D') = o]$

Similarly, for Impression x Time, User x Publisher x Time, User x Advertiser x Time, and User x Time privacy units

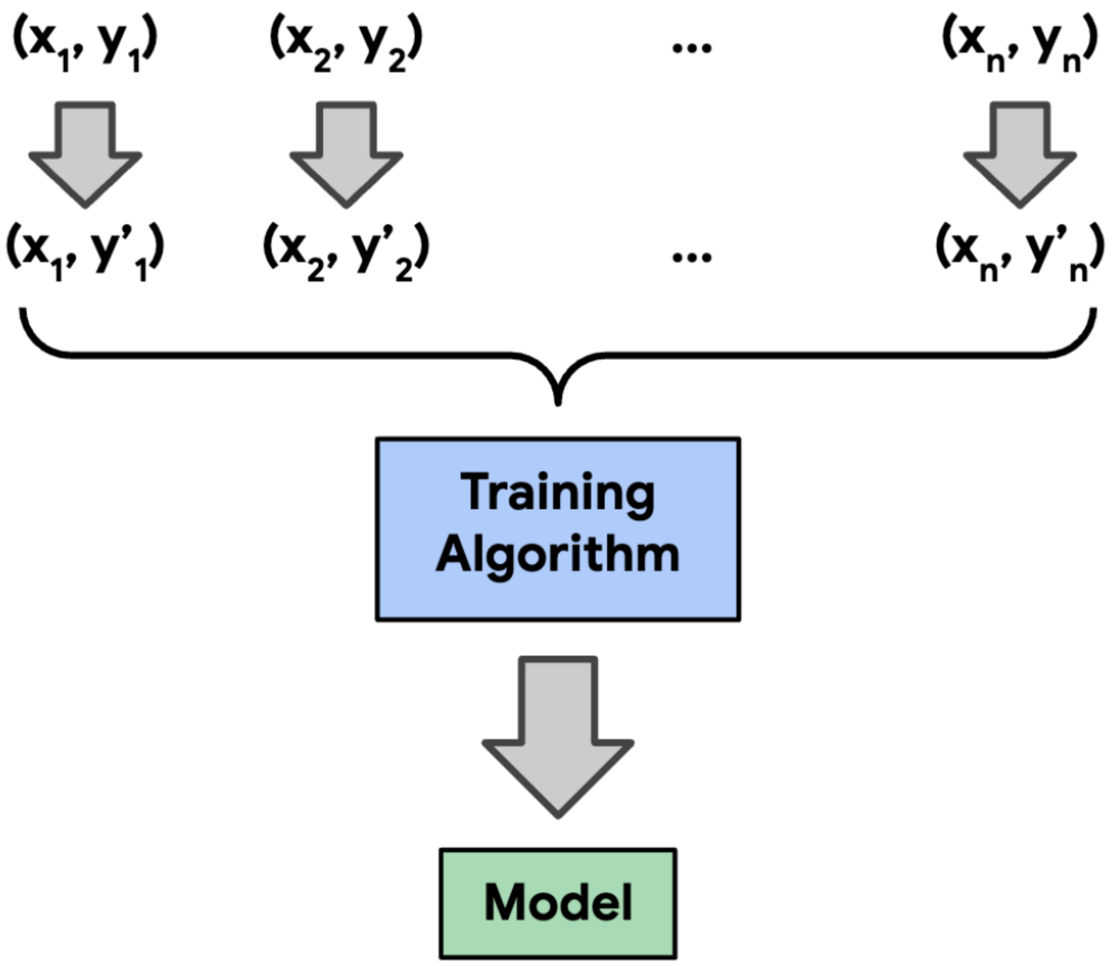
From Google's presentation, PATCG London 2023

Binary Randomized Response

RR_p *[Warner'65]*

Given $\mathbf{y_i}$ in {0, 1}:
$\mathbf{y'_i = y_i}$ with probability **1 - p**
Otherwise, $\mathbf{y'_i}$ is a random label in {0, 1}.

RR_p is ε-DP for $p = 2/(e^{\varepsilon}+1)$.



If y'_i is obtained by applying RR_p independently to y_i, then output model is ε-Label-DP.

From Google's presentation, PATCG London 2023

Handling Multiple Impressions Per Privacy Unit

Example: Consider User x Time privacy unit.

Cap number of impressions per user and time period to K (keeping K random impressions, or the K first impressions). Then, we have multiple options including:

1. For each user, set the privacy budget per impression to ε/K.
2. For each user i with $K_i \leq K$ impressions, set the privacy budget per impression to ε/K_i.

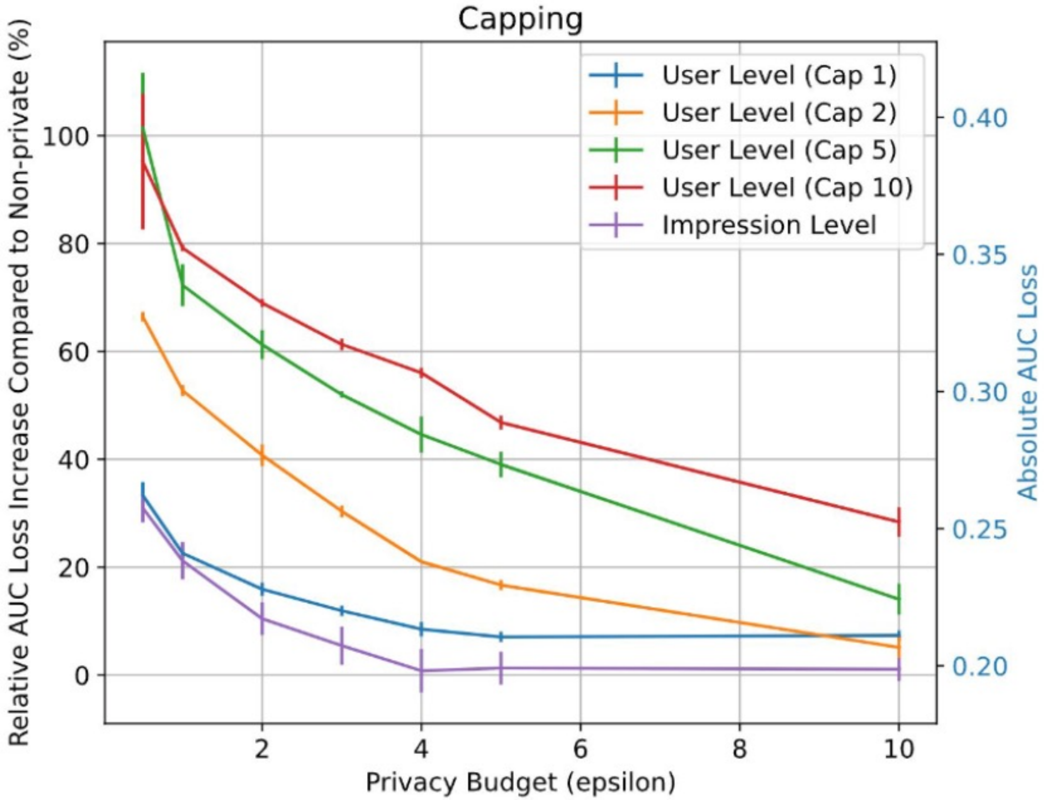
Both options satisfy ε-Label-DP for User x Time privacy unit.

Similar options hold for User x Advertiser x Time and User x Publisher x Time privacy units.

For Impression x Time privacy unit, no capping is needed. RR is applied privacy budget ε.

From Google's presentation, PATCG London 2023

Criteo Attribution Modeling for Bidding – Evaluation Results



Notes

- For Impression x Time privacy unit and ε = 4, relative AUC loss is 0.79%.
- For User x Time privacy unit with ε = 4, smallest relative AUC loss is 8.51%.
- For User x Time privacy unit, smaller loss is achieved by increasing caps as we increase ε.

From Google's presentation, PATCG London 2023

Complementary insights

- The AUC performance metric is not the most relevant for measuring the performance of bidding models
- Without surrogates, learning on noisy labels comes with poor performance in low privacy regime (e.g. epsilon < 3)
- Leveraging research works on noise-tolerant learning & learning on DP data, we can improve the performances of the learnt model via e.g. **debiasing the loss function, optimal transport, the use of “robust” losses, …**

Debiasing the loss function

Learning with Noisy Labels

Nagarajan Natarajan **Inderjit S. Dhillon** **Pradeep Ravikumar**
Department of Computer Science, University of Texas, Austin.
{naga86,inderjit,pradeepr}@cs.utexas.edu

Ambuj Tewari
Department of Statistics, University of Michigan, Ann Arbor.
tewaria@umich.edu

Lemma 1. *Let $\ell(t, y)$ be any bounded loss function. Then, if we define,*

$$\tilde{\ell}(t, y) := \frac{(1-\rho_{-y})\,\ell(t, y) - \rho_y\,\ell(t, -y)}{1 - \rho_{+1} - \boxed{\rho_{-1}}}$$

we have, for any t, y, $\mathbb{E}_{\tilde{y}}\left[\tilde{\ell}(t, \tilde{y})\right] = \ell(t, y)\,.$

= Proba (noisy label = 1 | true label = -1)

This surrogate loss might be not convex, even if the initial loss is !

For logistic regression, it is hopefully the case so optimisation is easier!

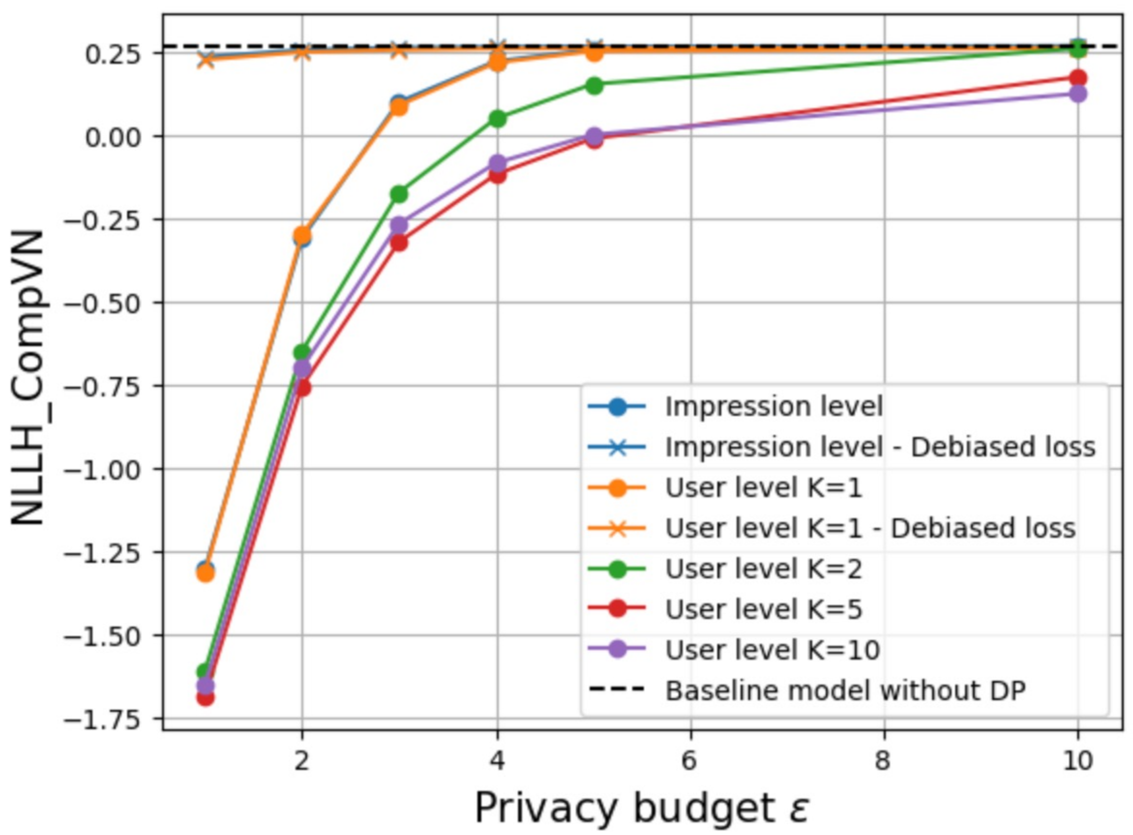
Empirical Results - Dataset

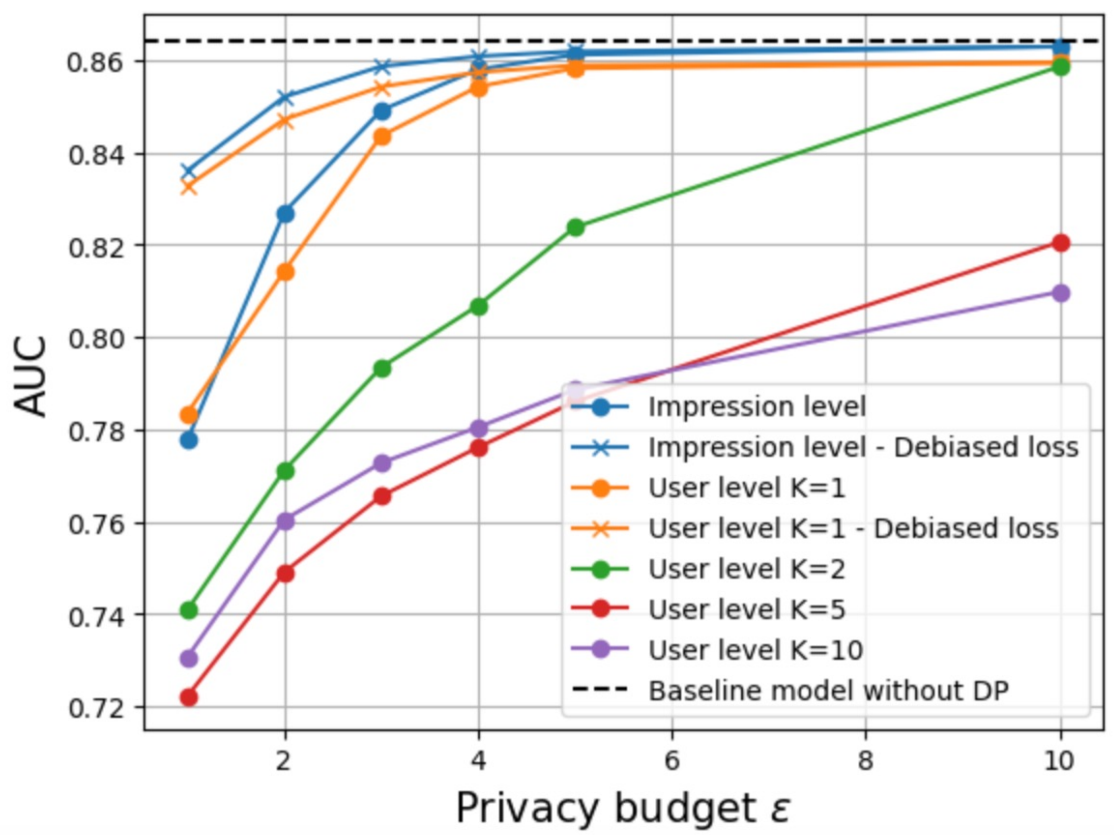
Criteo Attribution Modeling for Bidding Dataset

https://ailab.criteo.com/criteo-attribution-modeling-bidding-dataset/

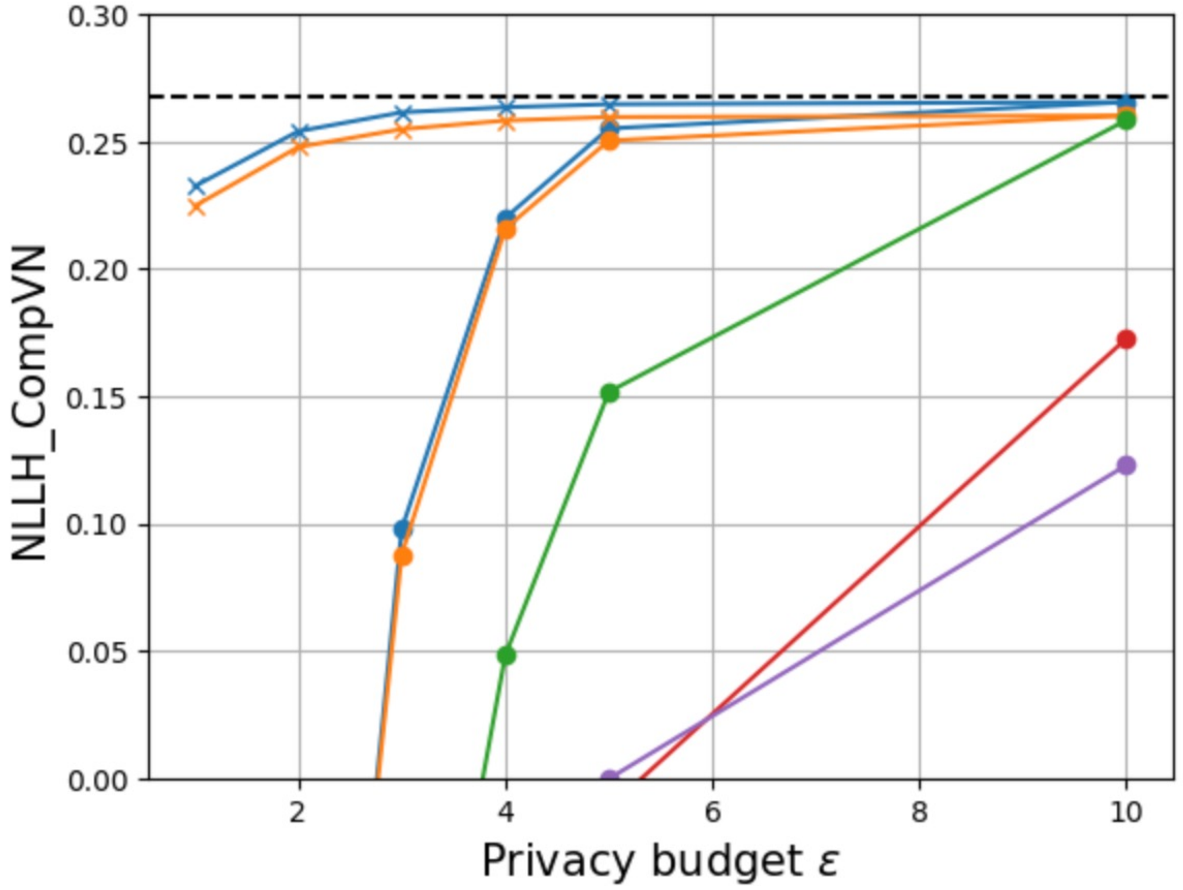
- Sample of 30 days of Criteo live traffic data.
- Each example corresponds to a click and contains:
  - **Features:** campaign ID, 9 contextual features, and the cost paid for the display.
  - **Label:** a 0/1 field indicating whether there was a conversion in the 30 days after the click and that is last-touch attributed to this click.
  - **User ID:** can be used to evaluate User x Time privacy unit.
- Number of rows is 5,947,563. Conversion rate (under last-touch attribution) is 6.74%.

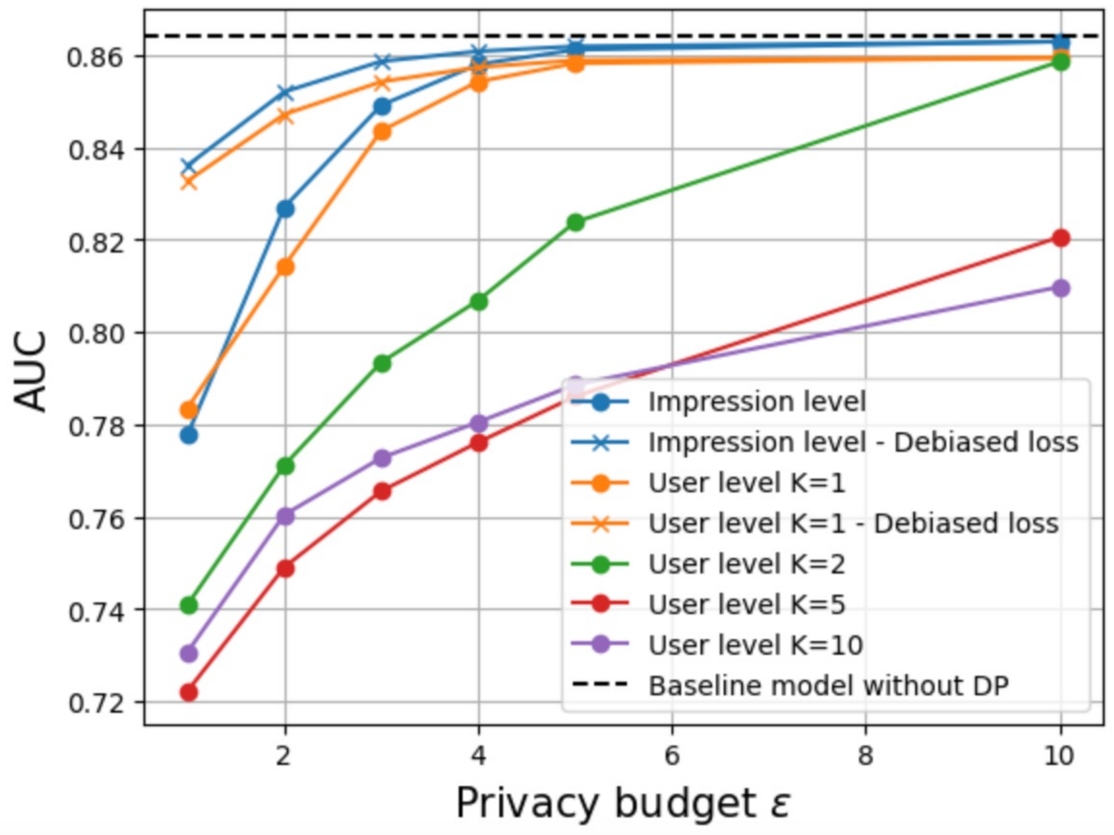
Empirical Results





Empirical Results





Next Steps

- Present to PATCG other alternatives to learn efficiently on noisy data (feature + label, or only label)
- Comparison with DP obtained with gradient perturbation (cf. DP-SGD)
- Comparison with DP obtained with WALR and other variants (see Meta’s upcoming presentation)