## A Variational Approach to Privacy and Fairness

#### Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund

KTH – Royal Institute of Technology ISE – Information Science and Engineering



January 22, 2021



## Problem #1: privacy leakage





## Problem #1: privacy leakage













### What do we want?

- 1. As **few** sensitive information **leakage** as possible.
- 2. As few influence on the task by the sensitive information as possible.
- 3. A decent usage of the data X.
  - Maitain sufficient information about X to perform tasks (e.g., task T).
- 4. **(Extra)** If the system is specific for *T*, leak as few of the information about *X* not used for *T* as possible.



### How can we do that?

- 1. Design a new system.
  - Robust against privacy attacks (e.g., DP training procedures).
  - Robust decision making.
- 2. Generate a representation Y of the data and keep our favourite ML System.
  - In particular, our proposal is:





## **Problem formalization**

#### In terms of **mutual information**:

Without a specific task T.



 $\underset{P_{Y|X}:I(X;Y|S)\geq r}{\operatorname{arg\,inf}}I(S;Y)$ 

Conditional Privacy Funnel

With a specific task T.



 $\underset{P_{Y|X}:I(T;Y|S)\geq r}{\operatorname{arg inf}}I(S;Y)+I(X;Y|S,T)$ 

#### Conditional Fairness Bottleneck



## **Problem relaxation**

We relax the constrained optimization problems through their Lagrangians.

Conditional Privacy Funnel (CPF)

$$\blacktriangleright \ \mathcal{L}_{CPF}(P_{Y|X},\lambda) = I(S;Y) - \lambda I(X;Y|S), \quad \lambda > 0$$

Conditional Fairness Bottleneck (CFB)

$$\blacktriangleright \mathcal{L}_{CFB}(P_{Y|X},\lambda) = I(S;Y) + I(X;Y|S,T) - \lambda I(T;Y|S), \quad \lambda > 0$$

We simplify the Lagrangians through an equivalence:

**Proposition:** Minimizing  $\mathcal{L}_{CPF}(P_{Y|X}, \lambda)$  is equivalent to minimizing

$$\mathcal{J}_{CPF}(P_{Y|X},\gamma) = I(X;Y) - \gamma I(X;Y|S), \quad \gamma = \lambda + 1$$

**Proposition**: Minimizing  $\mathcal{L}_{CFB}(P_{Y|X}, \lambda)$  is equivalent to minimizing

$$\blacktriangleright \mathcal{J}_{CFB}(P_{Y|X},\beta) = I(X;Y) - \beta I(T;Y|S), \quad \beta = \lambda + 1$$



# A solution via variational inference

We assume:

- A paramaterized encoder and marginal densities  $p_{Y|X,\theta}, q_{Y|\theta}$ .
- A variational decoder and inference densities  $q_{X|Y,S,\phi}, q_{T|Y,S,\phi}$ .

Then the solution of the Lagrangians is seeked by

A VAE-like method for the CPF.

A VIB-like method for the CFB.





## Caveats and "TO-DOs"

Caveat: The variational approach to the CPF and the CFB does not offer theoretical guarantees.

- It scales well, but needs an a-posteriori evaluation of the mutual information.
- **Caveat:** The encoding  $P_{Y|X}$  can leak information about the sensitive data S.
  - One needs to keep  $P_{Y|X}$  private and only release Y.
- TO-DO: Force the method to obtain a specific solution of the CPF or CFB with a single optimization.
  - Using similar techniques to the Convex Information Bottleneck Lagrangian.
- TO-DO: Find theoretical connections between the CPF and CFB and other measures of privacy/fairness.
  - It is hard to find connections with DP, since one uses on average measures and the other worst-case measures.
  - Some connections already found with **demographic parity**.



### Take-aways

- 1. The **privacy and fairness** problems are **similar** to each other.
- 2. The CPF and CFB model these problems as a constrained optimization involving information measures.
- 3. A variational Bayesian optimization of the Lagrangians of the CPF and CFB lead to a VAE/VIB-like optimization through gradient descent:
  - The encoder network is the same.
  - The decoder receives the protected data.
- 4. The proposed method achieves SoTA results on the fairness benchmarks and improves upon variational approaches to privacy.