# A Variational Approach to Privacy and Fairness

Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund

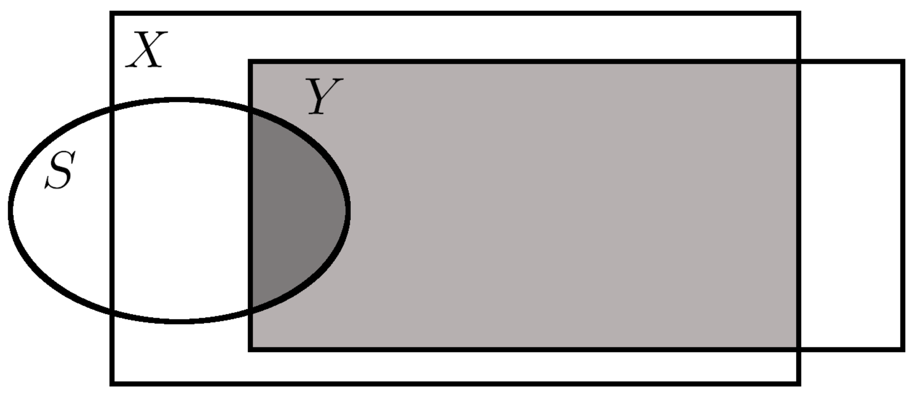
## 1. Problem formulation

### Conditional Privacy Funnel (CPF)

- Data we want to disclose: $X$
- Data we want to protect: $S$
- Privacy protecting mapping: $P_{Y|X}$
- Data we disclose: $Y$

The desired mapping is as follows:

$$\inf_{P_{Y|X}} \{\underline{I(S;Y)}\} \quad s.t. \quad \underline{I(X;Y|S)} \geq r$$

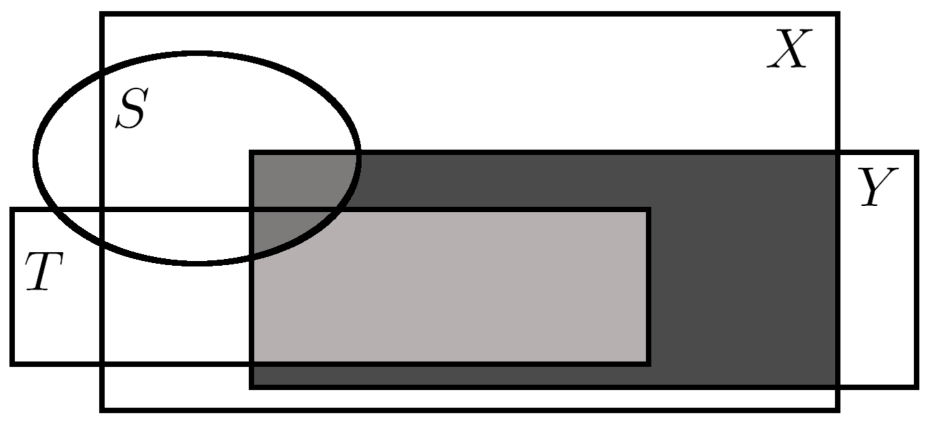



### Conditional Fairness Bottleneck (CFB)

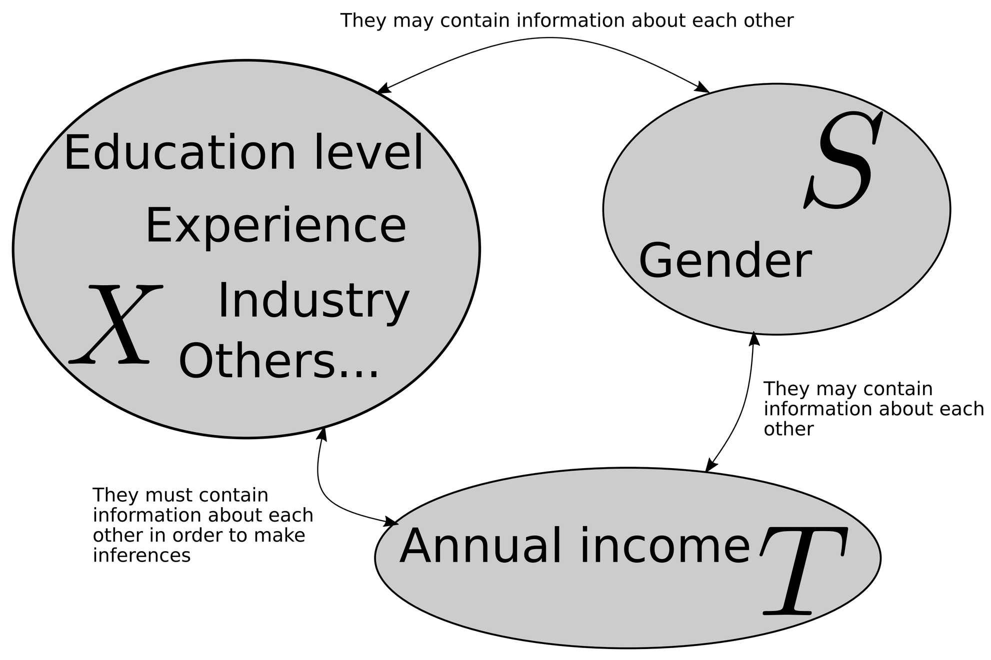
- Data we want to infer: $T$
- Data we want to use for inference: $X$
- Data we want to protect: $S$
- Fairness preserving mapping: $P_{Y|X}$
- Data we use for inference: $Y$

The desired mapping is as follows:

$$\inf_{P_{Y|X}} \{\underline{I(S;Y)} + \underline{I(X;Y|S,T)}\} \quad s.t. \quad \underline{I(T;Y|S)} \geq r$$




## 2. An example




We aim for a representation $Y$ of $X$ that:
- Reveals as few information about the gender as possible. (Privacy)
- Can infer the anual income without containing information about the gender. (Fairness)

## 3. Proposed approach

We optimize the Lagrangians:

$$\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda) = I(S;Y) - \lambda I(X;Y|S)$$
$$\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda) = I(S;Y) + I(X;Y|S,T) - \lambda I(T;Y|S)$$

where $\lambda > 0$.

**Proposition**: Minimizing the above Lagrangians is equivalent to minimizing

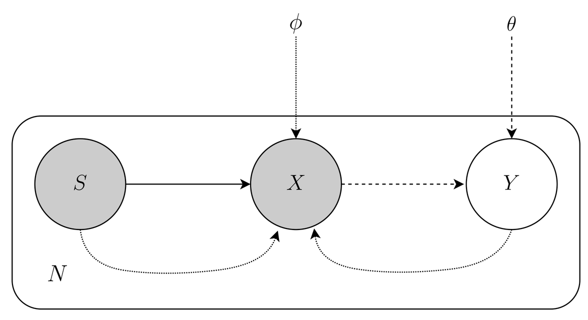
$$\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma) = I(X;Y) - \gamma I(X;Y|S)$$
$$\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta) = I(X;Y) - \beta I(T;Y|S)$$

where $\gamma = \lambda + 1$ and $\beta = \lambda + 1$.

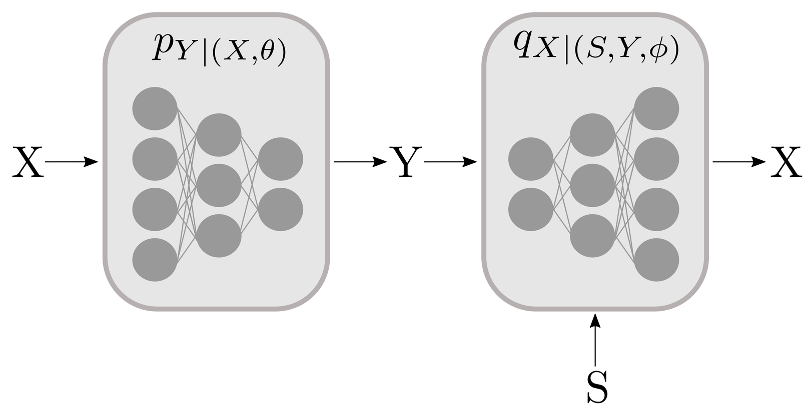
We consider a parametrized encoding density $p_{Y|(X,\theta)}$ and introduce the variational densities:
- Generative (CPF): $q_{X|(S,Y,\phi)}$
- Inference (CFB): $q_{T|(S,Y,\phi)}$
- Marginal (both): $q_{Y|\theta}$

### Variational CPF

The resulting graphical model is

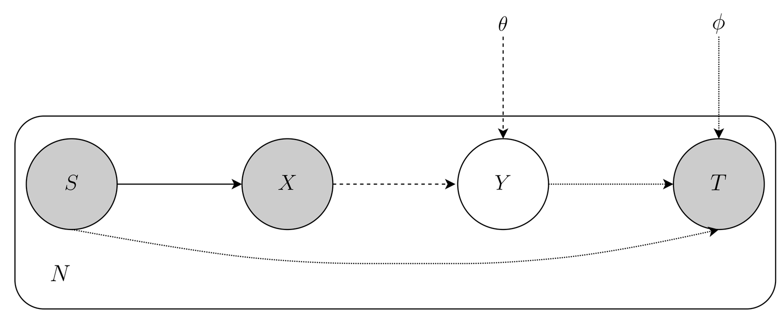



Which leads to a VAE-like architecture

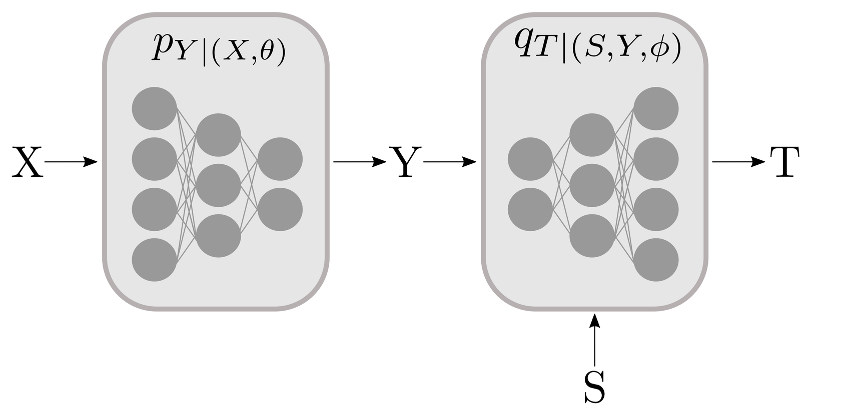



### Variational CFB

The resulting graphical model is

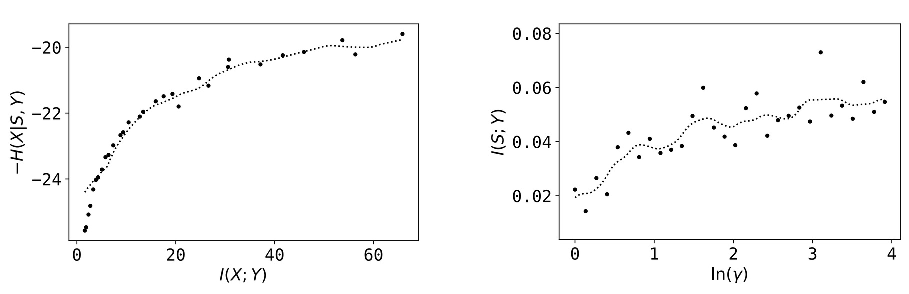



Which leads to a VIB-like architecture




## 4. Results

### Privacy on the Adult dataset

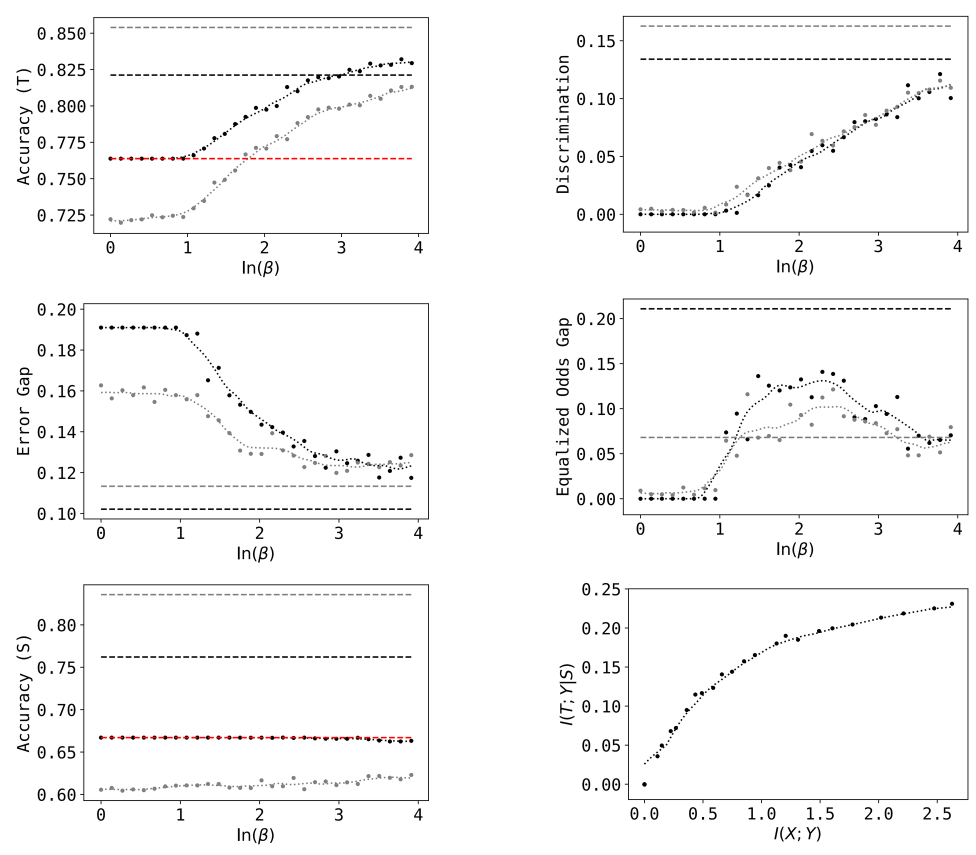


- $\gamma$ controls the trade off between utility and privacy.

| Methods | Accuracy (S) | $I(S;Y)$ |
|---|---|---|
| Ours | 0.60 – 0.64 | 0.01 - 0.08 |
| PPVAE | 0.79 – 0.93 | 0.29 – 0.63 |
| VFAE | 0.81 – 0.95 | 0.28 – 0.44 |

- better results than current SoTA variational methods for privacy.

### Fairness on the Adult dataset



- $\beta$ controls the trade off between utility and fairness.
- similar results than FFVAE or CFAIR.

## 5. 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.

Twitter: @BorjaRodrguezG2
E-mail: borjarg@kth.se