Differentially private Bayesian learning

Antti Honkela[1]

Helsinki Institute for Information Technology HIIT,
Department of Computer Science &
Department of Mathematics and Statistics &
Department of Public Health,
University of Helsinki

Helsinki Machine Learning Seminar, 6 March 2017

[1]Joint work with Mrinal Das, Onur Dikmen, Mikko Heikkilä, Joonas Jälkö, Samuel Kaski, Yusuke Okimoto and Kana Shimizu

The need for privacy: Genomics

- Rapid increase in generation of new genomic data
  - Estimated 228 000 human genomes sequenced by 2014
- The genome is potentially highly sensitive
  - Personal and inherently identifiable (Gymrek *et al.*, Science 2013)
  - Irrevokable and irreplacable
  - Possible leaks affect also relatives and offspring
  - Even aggregate data can compromise privacy (Homer *et al.*, PLoS Genetics 2008)
- ... yet the information contained within can be very useful for personalised health care

Why simple methods fail

- We want to study average weight μ of students
- Assume Bob wants to keep his weight private (he is afraid he might be bullied)
- Privacy mechanism: allow release only for averages of more than 20 people
- Assume Bob's weight is x and the total weight of all other 25 students in his class is y
- $\mathrm{mean}(\text{Bob} \cap \text{class}) = \frac{x+y}{1+25}$
- Knowing the average weight of the rest of the class

$$\mathrm{mean}(\text{class}) = \frac{y}{25}$$

would completely destroy Bob's privacy:

$$x = (1+25)\mathrm{mean}(\text{Bob} \cap \text{class}) - 25\mathrm{mean}(\text{class})$$

Call for a better solution

We want a privacy framework that

- protects against adversaries with arbitrary side information;
- allows fine-grained control of the level of privacy; and
- composes nicely for use in analysis pipelines.

Differential privacy (DP) gives all this.

Differential privacy (Dwork, 2006)

Definition

An algorithm $\mathcal{M}$ operating on a data set $\mathcal{D}$ is said to be (ϵ, δ)-differentially private ((ϵ, δ)-DP) if for any two data sets $\mathcal{D}$ and $\mathcal{D}'$, differing only by one sample, the probabilities of obtaining any result S fulfil

$$\Pr(\mathcal{M}(\mathcal{D}) \in S) \leq e^{\epsilon} \Pr(\mathcal{M}(\mathcal{D}') \in S) + \delta.$$

When $\delta = 0$, we get ϵ-DP, also known as pure DP.

Laplace mechanism (Dwork, 2006)

Theorem

Let

$$\Delta f = \sup_{\|\mathcal{D}-\mathcal{D}'\|=1} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1.$$

If $\xi \sim \mathrm{Lap}(0, \lambda)$ *with* $\lambda = \Delta f/\epsilon$, *then* $\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + \xi$ *is* ϵ*-DP.*

Laplace mechanism (Dwork, 2006)

Theorem

Let

$$\Delta f = \sup_{\|\mathcal{D}-\mathcal{D}'\|=1} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1.$$

If $\xi \sim \mathrm{Lap}(0, \lambda)$ *with* $\lambda = \Delta f/\epsilon$*, then* $\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + \xi$ *is* ϵ*-DP.*

Proof.

$$\begin{aligned}
\frac{p(\mathcal{M}(\mathcal{D}) = c)}{p(\mathcal{M}(\mathcal{D}') = c)} &= \frac{p(\mathrm{Lap}(c - f(\mathcal{D});\ \lambda))}{p(\mathrm{Lap}(c - f(\mathcal{D}');\ \lambda))} \\
&= \frac{\exp(\|c - f(\mathcal{D})\|_1/\lambda)}{\exp(\|c - f(\mathcal{D}')\|_1/\lambda)} \leq \exp\left(\frac{\|f(\mathcal{D}) - f(\mathcal{D}')\|_1}{\lambda}\right) \\
&\leq \exp\left(\frac{\Delta f}{\lambda}\right) = \exp(\epsilon)
\end{aligned}$$

□

Differential privacy and Bob

Let's apply differential privacy to Bob's case.

Assuming the weights of each student are in the interval $[30\text{ kg}, 60\text{ kg}]$, the sensitivity of the mean over N students,

$$f(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is $\Delta f = \sup \|f(\mathcal{D}) - f(\mathcal{D}')\|_1 = 30/N$ kg.

Differential privacy and Bob

Applying the Laplace mechanism with
$\Delta f = \sup \|f(\mathcal{D}) - f(\mathcal{D}')\|_1 = 30/N$ kg we get:

$\epsilon = 1.0, N = 25$:

```
Exact mean: 43.78
Private mean: 43.62 44.19 44.57 45.77 44.52
Mean absolute error: 1.20
```

Differential privacy and Bob

Applying the Laplace mechanism with
$\Delta f = \sup \|f(\mathcal{D}) - f(\mathcal{D}')\|_1 = 30/N$ kg we get:

$\epsilon = 10.0, N = 25$:

```
Exact mean: 45.99
Private mean: 46.68 45.93 46.04 45.95 46.15
Mean absolute error: 0.12
```

Differential privacy and Bob

Applying the Laplace mechanism with
$\Delta f = \sup \|f(\mathcal{D}) - f(\mathcal{D}')\|_1 = 30/N$ kg we get:

$\epsilon = 0.1, N = 25$:

```
Exact mean: 45.34
Private mean: 35.09 28.66 46.87 54.29 43.25
Mean absolute error: 12.00
```

Differential privacy and Bob

Applying the Laplace mechanism with
$\Delta f = \sup \|f(\mathcal{D}) - f(\mathcal{D}')\|_1 = 30/N$ kg we get:

$\epsilon = 0.1, N = 250$:

```
Exact mean: 44.76
Private mean: 44.88 45.28 45.02 41.96 46.46
Mean absolute error: 1.20
```

Differential privacy and Bob

Applying the Laplace mechanism with
$\Delta f = \sup \|f(\mathcal{D}) - f(\mathcal{D}')\|_1 = 30/N$ kg we get:

$\epsilon = 1.0, N = 250$:

```
Exact mean: 45.22
Private mean: 45.29 45.23 45.35 45.35 45.23
Mean absolute error: 0.12
```

Attacking differential privacy

Let us now check the error in estimating the true weight.

Attacking differential privacy

Let us now check the error in estimating the true weight.

$\epsilon = 1.0, N = 25$:

```
Exact mean: 43.78
Private mean: 43.62 44.19 44.57 45.77 44.52
Mean absolute error: 1.20

Exact attack error: 0.00
Private attack error: 48.67 44.20 16.36 23.05 24.89
Mean absolute error: 45.04
```

Attacking differential privacy

Let us now check the error in estimating the true weight.

$\epsilon = 10.0, N = 25$:

```
Exact mean: 45.99
Private mean: 46.68 45.93 46.04 45.95 46.15
Mean absolute error: 0.12

Exact attack error: 0.00
Private attack error: 21.25 2.98 0.73 5.16 2.28
Mean absolute error: 4.49
```

Attacking differential privacy

Let us now check the error in estimating the true weight.

$\epsilon = 0.1, N = 25$:

```
Exact mean: 45.34
Private mean: 35.09 28.66 46.87 54.29 43.25
Mean absolute error: 12.00

Exact attack error: 0.00
Private attack error: 17.78 100.22 296.54 882.76 297.45
Mean absolute error: 447.56
```

Attacking differential privacy

Let us now check the error in estimating the true weight.

$\epsilon = 0.1, N = 250$:

```
Exact mean: 44.76
Private mean: 44.88 45.28 45.02 41.96 46.46
Mean absolute error: 1.20

Exact attack error: 0.00
Private attack error: 101.85 1030.99 107.32 961.58 1231.00
Mean absolute error: 450.60
```

Attacking differential privacy

Let us now check the error in estimating the true weight.

$\epsilon = 1.0, N = 250$:

```
Exact mean: 45.22
Private mean: 45.29 45.23 45.35 45.35 45.23
Mean absolute error: 0.12

Exact attack error: 0.00
Private attack error: 1.96 3.16 5.57 68.27 19.89
Mean absolute error: 45.18
```

Outline

Outline

Bayesian inference for conjugate exponential models

Consider an exponential family model

$$p(x \mid \eta) = h(x) \exp(\eta^T S(x) - A(\eta))$$

with a conjugate prior

$$p(\eta \mid \tau, n_0) = H(\tau, n_0) \exp(\tau^T \eta - n_0 A(\eta)).$$

(Examples: binomial, multinomial, Poisson, Gaussian)

Bayesian inference for conjugate exponential models

Consider an exponential family model

$$p(x \mid \eta) = h(x) \exp(\eta^T S(x) - A(\eta))$$

with a conjugate prior

$$p(\eta \mid \tau, n_0) = H(\tau, n_0) \exp(\tau^T \eta - n_0 A(\eta)).$$

(Examples: binomial, multinomial, Poisson, Gaussian)

Given a sample $\mathcal{D} = (x_1, \ldots, x_n)$, the likelihood is

$$p(\mathcal{D} \mid \eta) = \prod_i h(x_i) \exp\left(\eta^T \left(\sum_i S(x_i)\right) - n\, A(\eta)\right).$$

Bayesian inference for conjugate exponential models

Consider an exponential family model

$$p(x \mid \eta) = h(x)\exp(\eta^T S(x) - A(\eta))$$

with a conjugate prior

$$p(\eta \mid \tau, n_0) = H(\tau, n_0)\exp(\tau^T \eta - n_0 A(\eta)).$$

(Examples: binomial, multinomial, Poisson, Gaussian)

Given a sample $\mathcal{D} = (x_1, \ldots, x_n)$, the likelihood is

$$p(\mathcal{D} \mid \eta) = \prod_i h(x_i)\exp\left(\eta^T \left(\sum_i S(x_i)\right) - n\,A(\eta)\right).$$

Combining the prior and the likelihood yields the posterior

$$p(\eta \mid \tau, n_0, \mathcal{D}) \propto \exp\left(\left(\tau + \sum_i S(x_i)\right)^T \eta - (n_0 + n)A(\eta)\right)$$

Bayesian inference and mean parameters

It can be shown that the expectation of the mean of the parameter is

$$E[\mu \mid \tau, n_0] = \frac{\tau}{n_0}.$$

This implies that for the posterior expectation is

$$E[\mu \mid \tau, n_0, \mathcal{D}] = \frac{\tau + \sum_i S(x_i)}{n + n_0}.$$

Differentially private Bayesian inference

For exponential family models

$$p(\eta \mid \mathcal{D}, \dots) = p(\eta \mid \sum_i S(x_i), \dots),$$

i.e. all information about the data $\mathcal{D}$ is contained in the sum of sufficient statistics $\sum_i S(x_i)$.

This suggests a differentially private version where we apply the Laplace mechanism on the sum to obtain perturbed sufficient statistics

$$\mathcal{M}(\mathcal{D}) = \sum_i S(x_i) + \xi,$$

with $\xi \sim \mathrm{Lap}(\Delta S/\epsilon)$, and then proceed with the inference as usual (Foulds *et al.*, UAI 2016; Honkela *et al.*, 2016).

Consistency and efficiency

- Consistency: DP estimates of posterior mean parameters converge to the corresponding non-private values as $n \to \infty$

$$\begin{aligned}
\hat{\theta}_{\mathcal{M}} = \frac{\tau + \mathcal{M}(\mathcal{D})}{n + n_0} &= \frac{\tau + \sum_i S(x_i) + \xi}{n + n_0} \\
&= \frac{\tau + \sum_i S(x_i)}{n + n_0} + \frac{\xi}{n + n_0} \\
&\xrightarrow{p} \frac{\tau + \sum_i S(x_i)}{n + n_0} = \hat{\theta}_{NP}.
\end{aligned}$$

- Convergence rate $\mathcal{O}(1/n)$ is optimal for any DP mechanism, i.e. sufficient statistic perturbation is asymptotically efficient

Outline

Differentially private linear regression (Mrinal Das)

- Setting: inputs $\mathbf{x}_i \in \mathbb{R}^d$, prediction targets $y_i \in \mathbb{R}$
- Linear regression model:

$$
\begin{aligned}
y_i | \mathbf{x}_i &\sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \lambda) \\
\boldsymbol{\beta} &\sim N(0, \lambda_0 I)
\end{aligned}
$$

- Privacy requirement: the inferred parameters $\boldsymbol{\beta}$ should be differentially private with respect to the data $\mathbf{x}_i, y_i$

Bayesian linear regression and DP

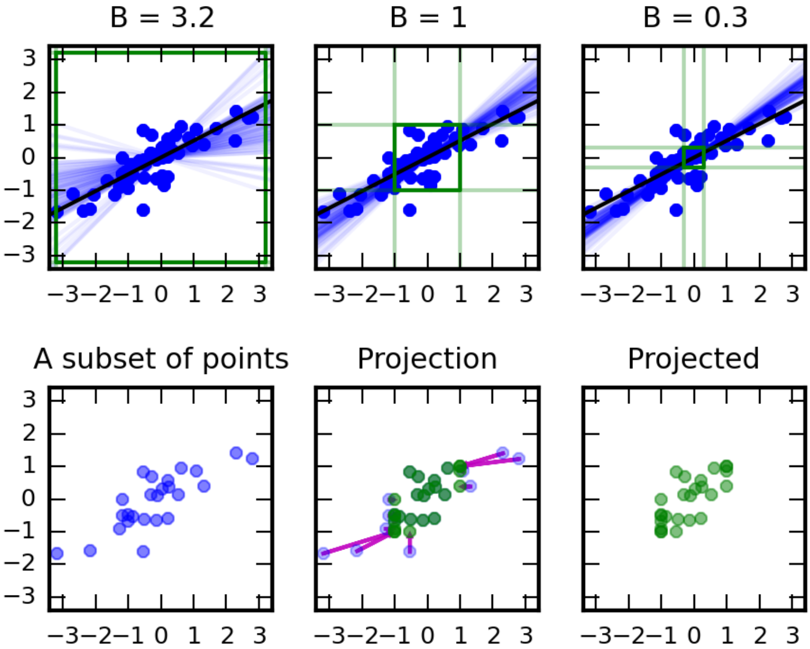
$$y_i | \mathbf{x}_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \lambda)$$

- Gaussian distribution, an exponential family
- Sufficient statistics: mean and covariance
  - Specifically $\mathrm{E}[\mathbf{x}_i y_i]$ and $\mathrm{E}[\mathbf{x}_i \mathbf{x}_i^T]$
- Fixed size, does not depend on number of samples
- DP inference: perturb $\mathrm{E}[\mathbf{x}_i y_i]$ with Laplace and $\mathrm{E}[\mathbf{x}_i \mathbf{x}_i^T]$ with Wishart noise, then perform inference as usual

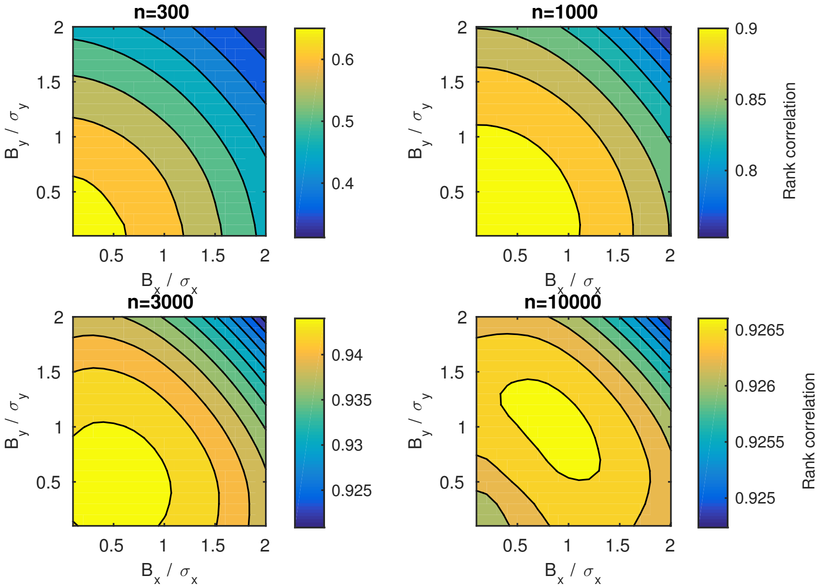
Efficient DP learning in practice

- Asymptotic efficiency is insufficient to guarantee practical efficiency
- High dimensional data needs more DP noise
  - More aggressive dimensionality reduction than usual often needed
- Further: a single outlier can impose huge bounds on the data
  - Need to inject a lot of noise in DP to mask it
  - The useful contribution such points have in learning is at best minimal

Clipping in action



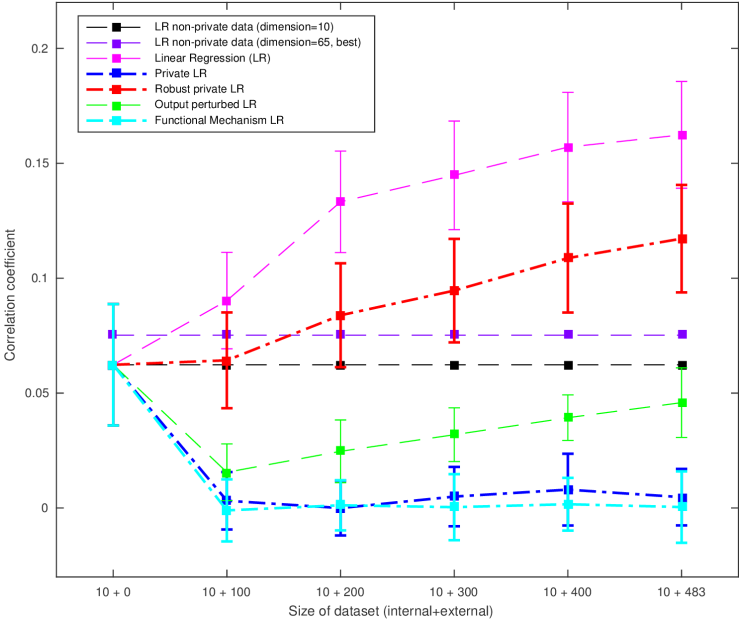
The effect of decreasing B_x, B_y



DP linear regression for drug sensitivity prediction

- Task: predict the sensitivity of cell lines to a cancer drug using gene expression data
- Data: Genomics of Drug Sensitivity in Cancer (GDSC) project gene expression data and sensitivity to 124 drugs
- Evaluation: rank correlation of predictions over cell lines
- Dimensionality reduction: use prior knowledge to select 65 most important cancer genes, ranked by observed number of mutations in an unrelated data set

DP linear regression for drug sensitivity prediction



Outline

DP for non-exponential-family models

- Sufficient statistic perturbation is efficient, but only applicable to exponential family models
- MCMC inference applicable to more general models, but current DP variants (Dimitrakakis *et al.*, ALT 2014; Wang *et al.*, ICML 2015) are inefficient and cumbersome
  - Require model-specific sensitivity derivations
  - Privacy guarantee conditional on convergence
  - Privacy cost linear in the number of samples drawn
- Variational inference offers a promising generic alternative

Variational inference

- True posterior $p(\boldsymbol{\theta}|\mathbf{x})$ is approximated with a variational distribution $q_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ that has a simpler form
- Optimal approximation obtained through minimising the Kullback–Leibler (KL) divergence between $q_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x})$
- Equivalently, maximising the *evidence lower bound* (ELBO)

$$
\begin{aligned}
\mathcal{L}(q_{\boldsymbol{\xi}}) &= E_{q_{\boldsymbol{\xi}}(\boldsymbol{\theta})}\left[\ln\left(\frac{p(\mathbf{x},\boldsymbol{\theta})}{q_{\boldsymbol{\xi}}(\boldsymbol{\theta})}\right)\right] \\
&= \sum_{i=1}^{N}\left(-\frac{1}{N}\mathrm{KL}(q_{\boldsymbol{\xi}}(\boldsymbol{\theta})\,||\,p(\boldsymbol{\theta})) + E_q\left[\ln p(x_i|\boldsymbol{\theta})\right]\right) \\
&\equiv \sum_{i=1}^{N}\mathcal{L}_i(q_{\boldsymbol{\xi}})
\end{aligned}
$$

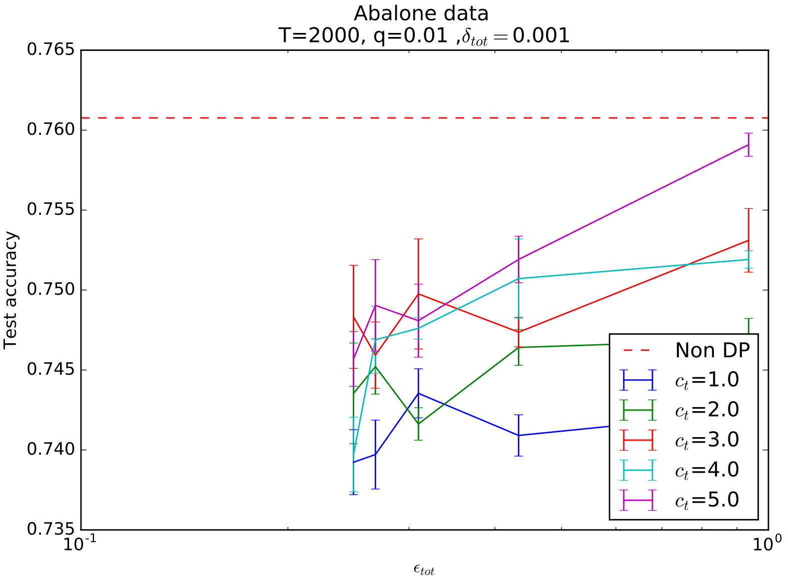
Doubly stochastic variational inference

- Modern approach to gradient-based inference
- Transform $\nabla E_q[\ldots]$ to $E_q[\nabla \ldots]$
- Use Monte Carlo to evaluate the expectation
- Optimise using stochastic gradient optimisation

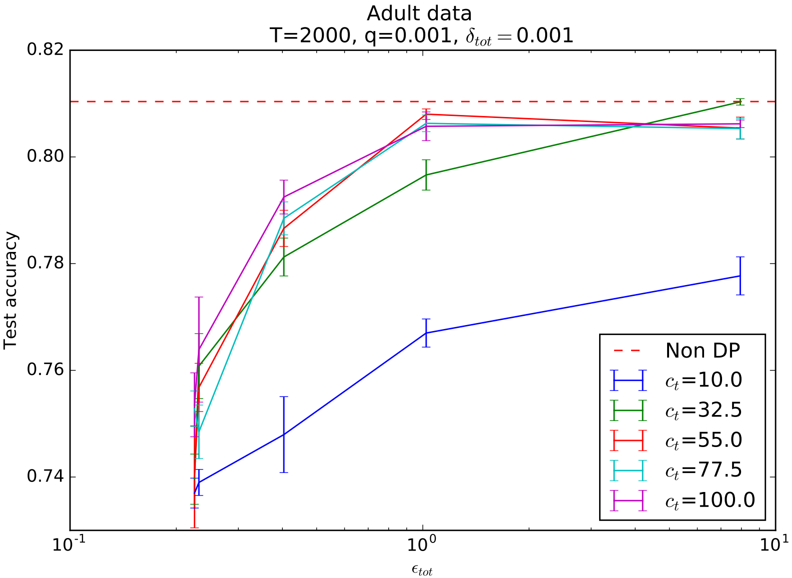
DP variational inference (Joonas Jälkö and Onur Dikmen)

- Each $g(x_i) = \nabla_{\boldsymbol{\xi}} \mathcal{L}_i(q_{\boldsymbol{\xi}})$ is clipped s.t. $||g(x_i)||_2 \leq c_t$ in order to calculate *gradient sensitivity*
- Subsampling with frequency q in order to use the *privacy amplification theorem*
- Gradient contributions from all data samples in the mini batch are summed and perturbed with Gaussian noise $\mathcal{N}(0, 4c_t^2\sigma_\delta^2\mathbf{I})$
- Total privacy cost can be computed from composition theorems

DP logistic regression results on UCI Abalone



DP logistic regression results on UCI Adult

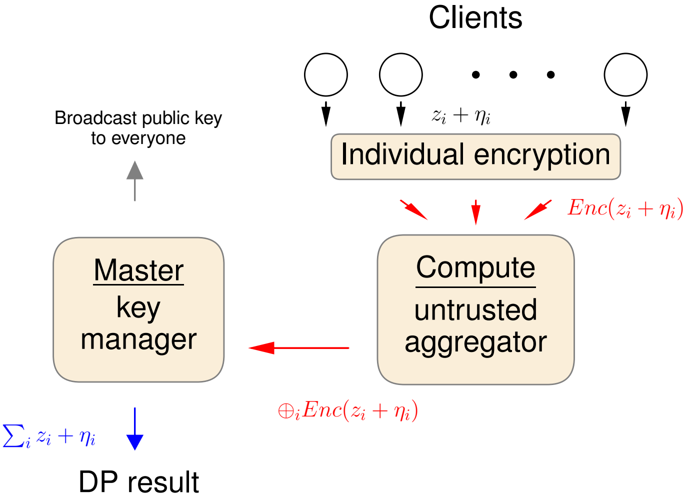


Outline

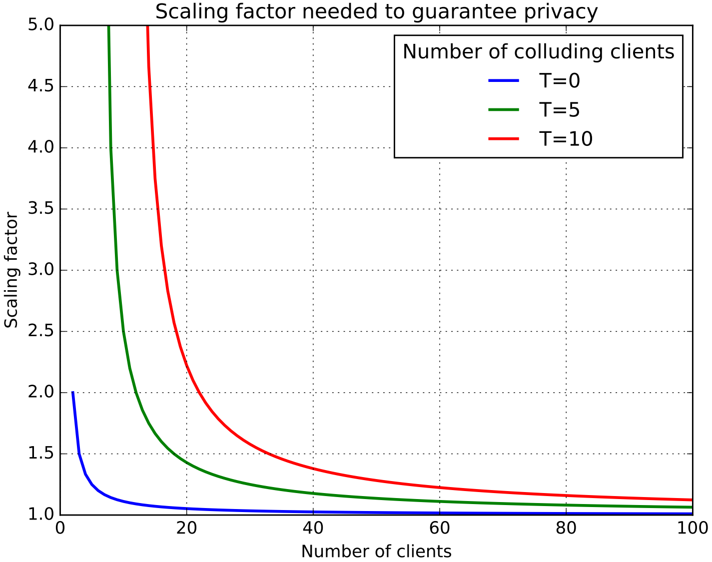
DP and distributed data (Mikko Heikkilä, Yusuke Okimoto and Kana Shimizu)

- Previous methods assume a trusted aggregator has access to all data, limiting their applicability
- Naive distributed approach needs to add noise proportional to the size of each local data set
- Secure multi-party computation with *homomorphic encryption* can be used to securely combine distributed data sets
- The Gaussian mechanism allows easy distributed generation of DP noise

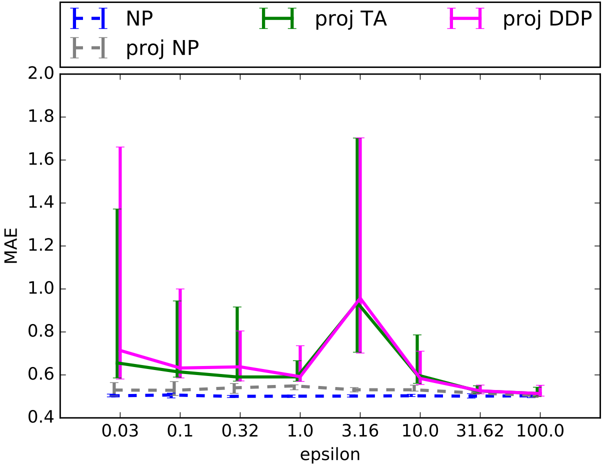
System diagram for distributed DP inference



Penalty for distributed inference

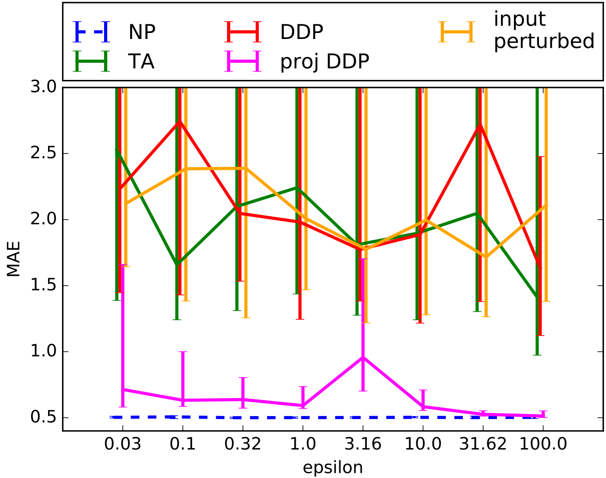


Linear regression results on UCI Wine Quality (white)



d=11, sample size=2000, repeats=40, $\delta = 0.0001$

Linear regression results on UCI Wine Quality (white)



d=11, sample size=2000, repeats=40, $\delta = 0.0001$

Conclusion

- DP as a strong privacy framework
- DP Bayesian inference through perturbing the sufficient statistics $S(x_i)$
- Asymptotically consistent and efficient
- For finite data: dimensionality reduction and clipping the data are essential to obtain better performance
- DP variational inference for more general models
- DP inference with distributed data

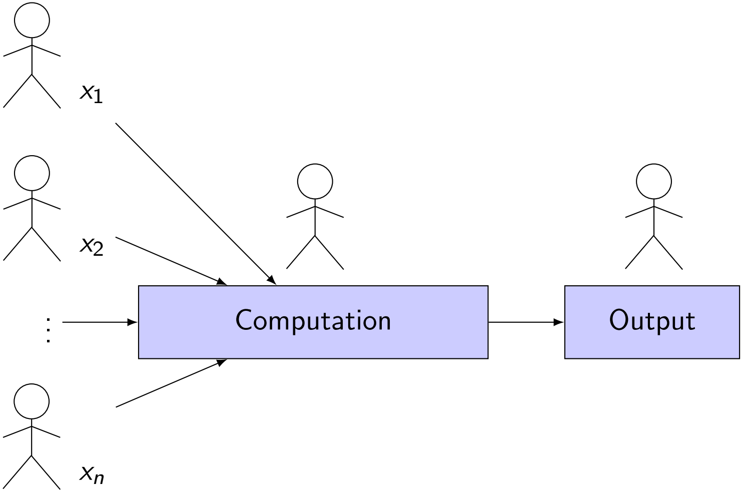
References

A. Honkela, M. Das, O. Dikmen, S. Kaski.
Efficient differentially private learning improves drug sensitivity prediction
arXiv:1606.02109 [stat.ML]

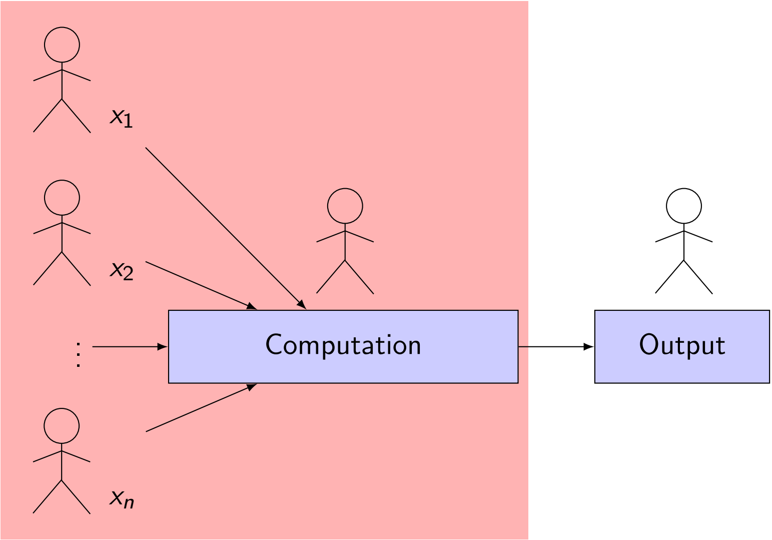
J. Jälkö, O. Dikmen, A. Honkela.
Differentially Private Variational Inference for Non-conjugate Models
arXiv:1610.08749 [stat.ML]

M. Heikkilä, Y. Okimoto, S. Kaski, K. Shimizu, A. Honkela
Differentially Private Bayesian Learning on Distributed Data
arXiv:1703.01106 [stat.ML]

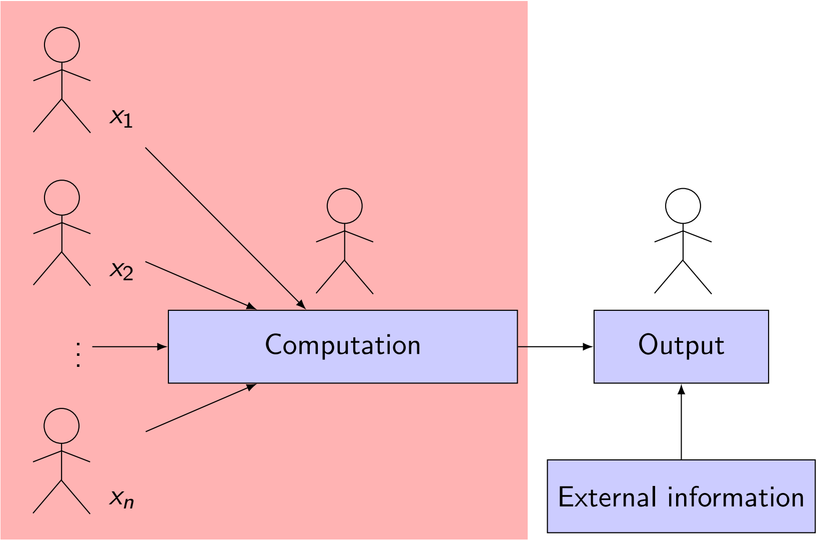
Privacy in machine learning



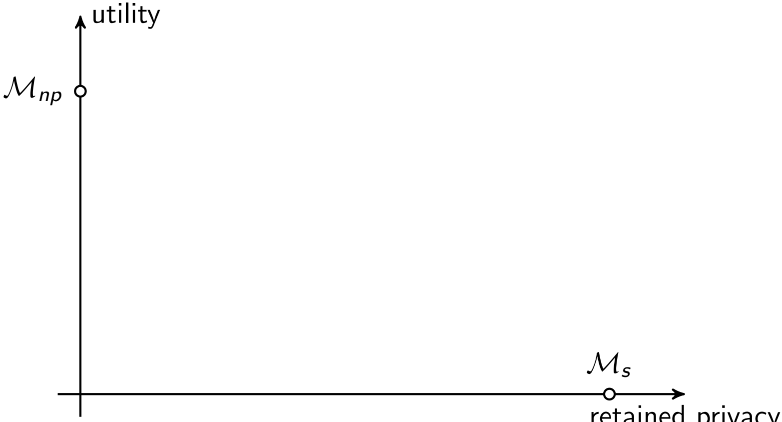
Privacy in machine learning



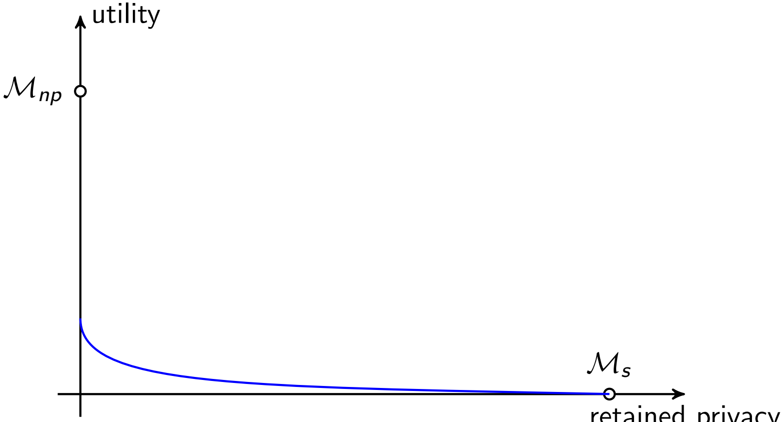
Privacy in machine learning



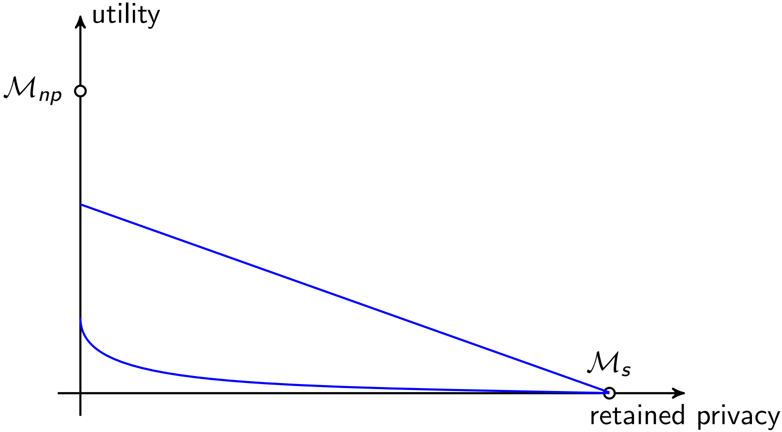
The privacy–utility tradeoff



The privacy–utility tradeoff



The privacy–utility tradeoff



The privacy–utility tradeoff