OWKIN

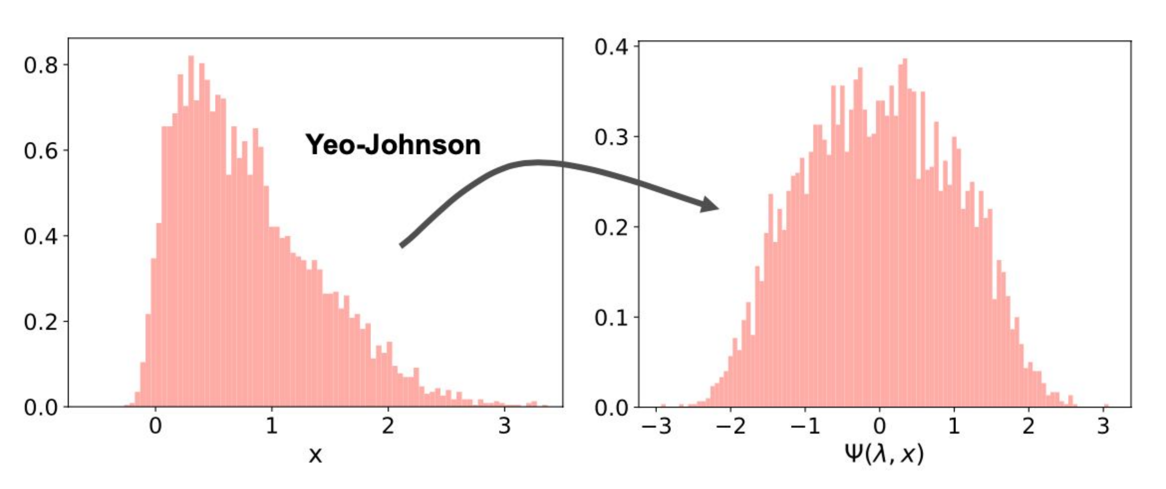
Tanguy Marchand

NeurIPS 2022

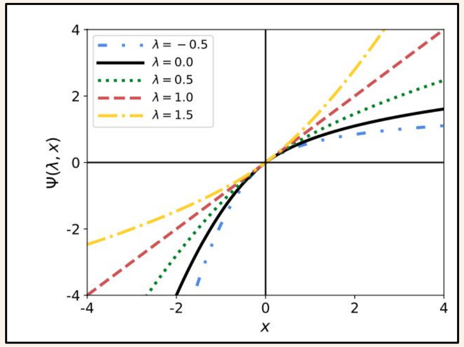
SecureFedYJ: a safe feature Gaussianization protocol for Federated Learning

Tanguy Marchand, Boris Muzellec, Constance Beguier, Jean Ogier du Terrail, Mathieu Andreux

The Yeo-Johnson (YJ) transformation



$$\Psi(\lambda, x) = \begin{cases} [(x+1)^{\lambda} - 1]/\lambda, & \text{if } x \geq 0, \lambda \neq 0, \\ \ln(x+1), & \text{if } x \geq 0, \lambda = 0, \\ -[(-x+1)^{2-\lambda} - 1]/(2-\lambda), & \text{if } x < 0, \lambda \neq 2, \\ -\ln(-x+1), & \text{if } x < 0, \lambda = 2. \end{cases}$$

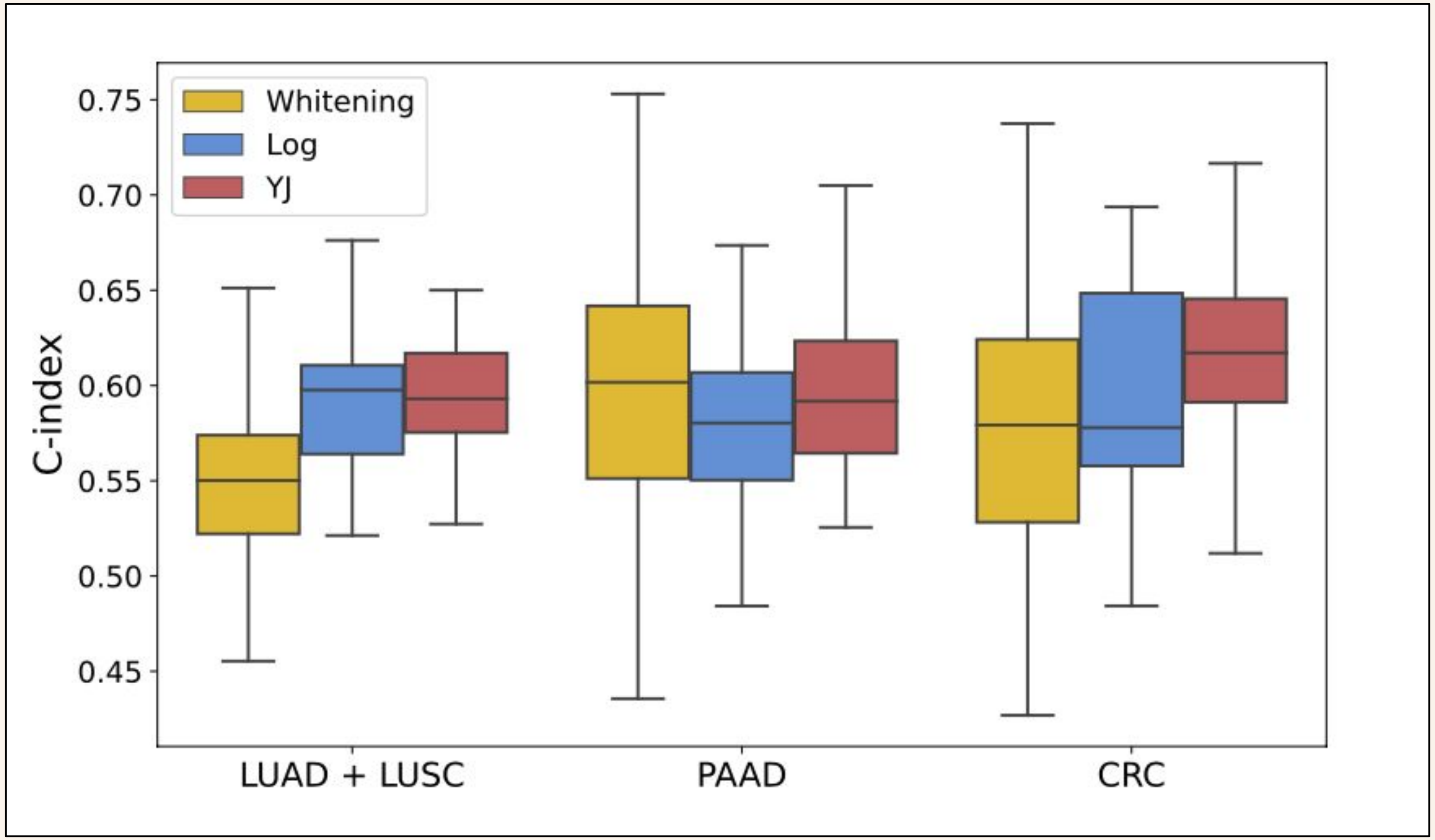


Optimal λ_* found by maximizing the log-likelihood:

$$\log \mathcal{L}_{\mathrm{YJ}}(\lambda) = -\frac{n}{2} \log(\sigma^2_{\Psi(\lambda, \{x_i\})}) + (\lambda - 1) \sum_{i=1}^{n} \mathrm{sgn}(x_i) \log(|x_i| + 1) - \frac{n}{2} \log(2\pi)$$

Brent minimization method

The Yeo-Johnson transformation: effect on survival models using TCGA gene expression raw counts



Cross-Silo Federated Learning

→ Datasets remain on each server

→ Each server train their own local model, and a central server, aggregates at regular steps all the models

→ Assumes that the loss is separable

→ Challenges:
- **heterogeneity**: ensure pooled-equivalence irrespective of data partition
- **Confidentiality**: Secure Multi-party computation

Research question

In cross-silo FL, as YJ log-likelihood is not **separable**, can we apply the Yeo-Johnson transformation:

→ and obtain a result identical to the case where all the data is pooled in the same server? (**heterogeneity**)
→ without leaking any information on the data from each center? (**confidentiality**)
→ using an algorithm that can be realistically applied in real-world FL project? (**communication efficiency**)

First theoretical contribution: the negative log-likelihood is Convex

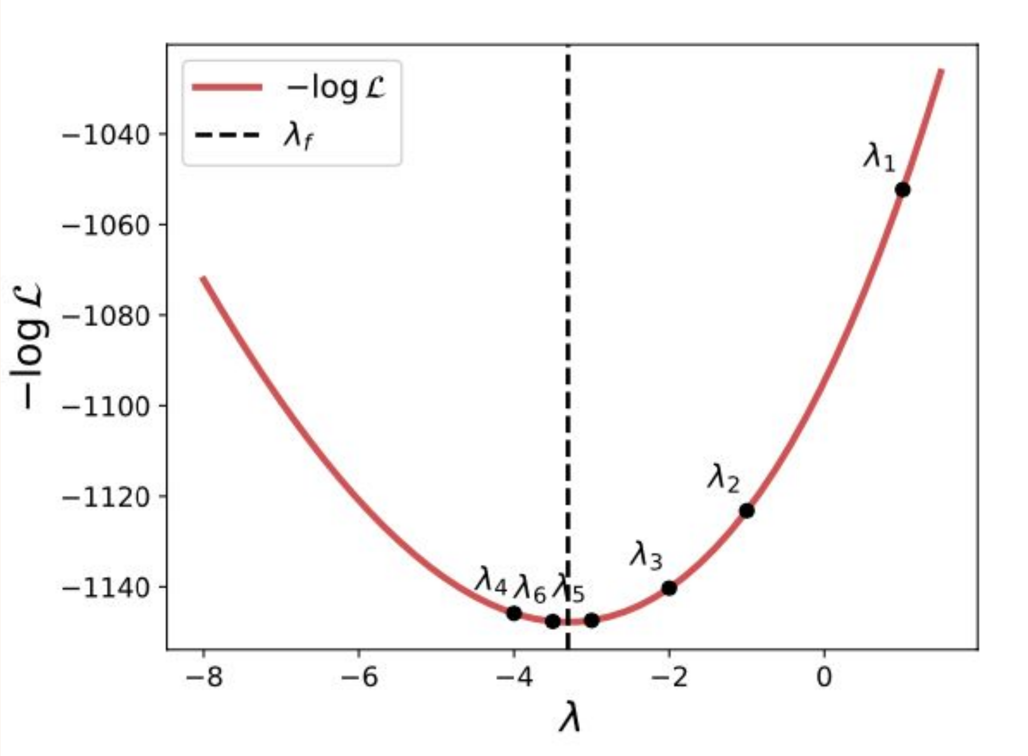
Proposition 3.1: The negative log-likelihood $\lambda \mapsto -\log \mathcal{L}_{\mathrm{YJ}}(\lambda)$ is strictly convex

1964		1995		2000		**Our work**
Box-Cox transformation introduced [1]	→	Box-Cox negative log-likelihood is convex [2]	→	YJ transformation is introduced [3]	→	We extend the proof of the Box-Cox convexity to YJ

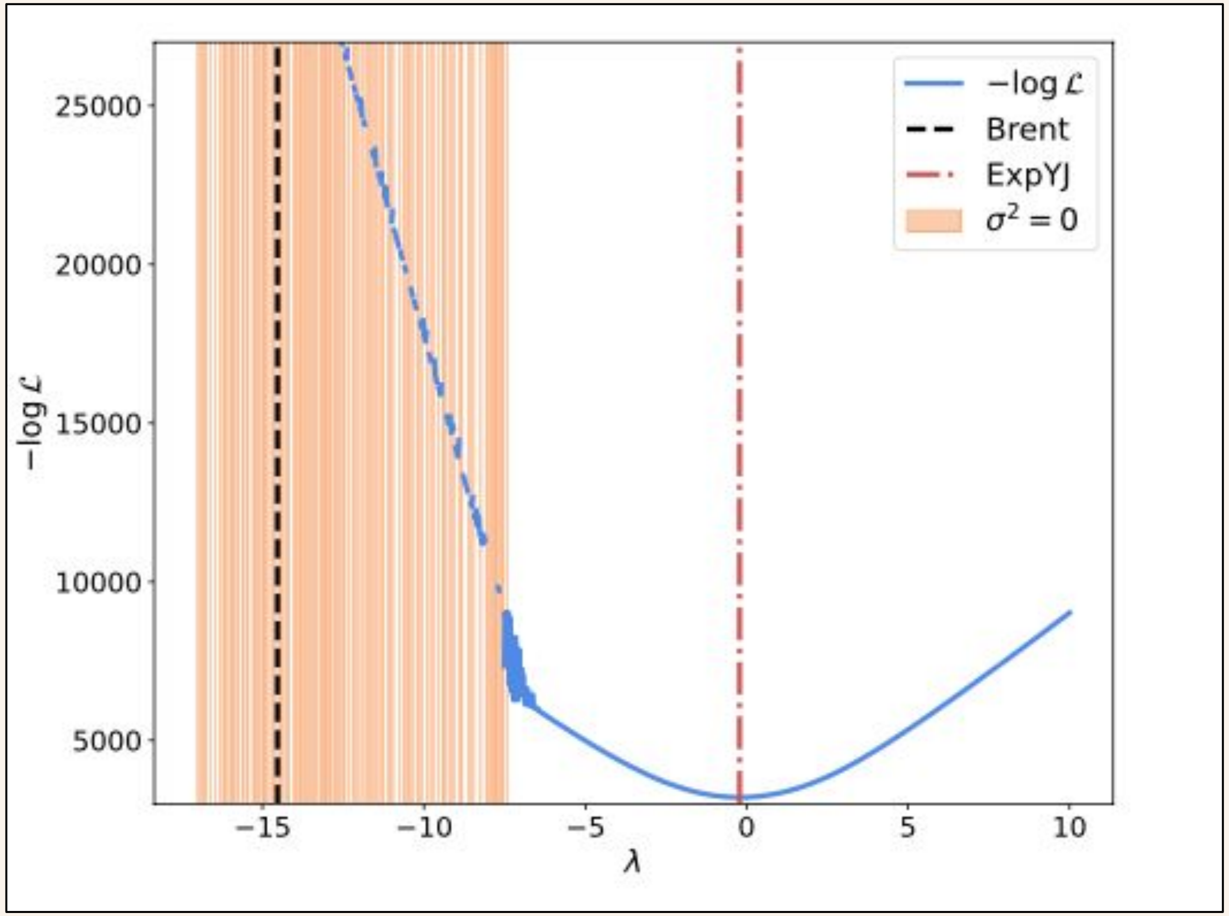
[1] George EP Box and David R Cox. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2):211–243, 1964.
[2] Elies Kouider and Hanfeng Chen. Concavity of Box-Cox log-likelihood function. Statistics and probability letters, 25(2):171–175, 1995.
[3] In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. Biometrika, 87(4):954–959, 2000.

ExpYJ: using an exponential search computing only the sign of the derivative of the negative log-likelihood

Exponential search:
1. Find an upper and lower bound
2. Perform a binary search

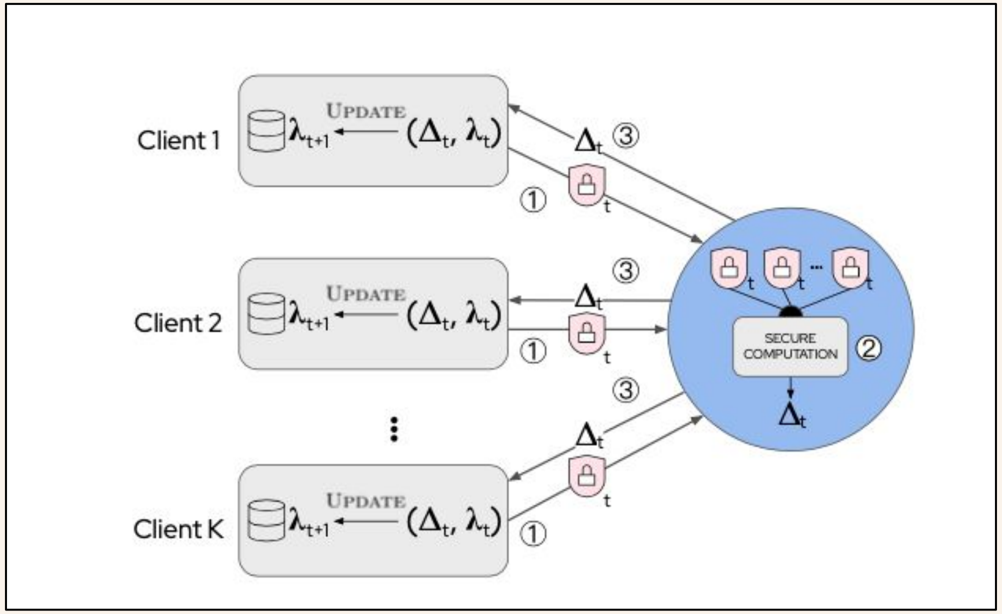


More stable than Brent minimization method



$$\log \mathcal{L}_{\mathrm{YJ}}(\lambda) = -\frac{n}{2}\log\left(\boxed{\sigma^2_{\Psi(\lambda,\{x_i\})}}\right) + (\lambda - 1)\sum_{i=1}^{n} \operatorname{sgn}(x_i)\log(|x_i| + 1) - \frac{n}{2}\log(2\pi)$$

SecureFedYJ: Secure Multiparty Computation (SMC) + expYJ



The sign of the derivative of the log-likelihood is computed using SMC at each step.

secureFedYJ:

→ is pooled-equivalent
→ is resilient to heterogeneity
→ does not leak information on the dataset: cf Prop 4.1 of our paper
→ can be realistically used in real-world FL project

Summary of contributions

- First proof that the YJ negative log-likelihood is convex

- **expYJ**, optimizing YJ using exponential search
  - as accurate as SOTA YJ method...
  - ...and even more stable !

- **secureFedYJ**
  - pooled-equivalent, and therefore resilient to heterogeneity
  - does not leak any further information than the final YJ parameters
  - can be realistically used in a real-world FL project