

# Unrolled proximal algorithms for estimating the reproduction number of Covid-19 pandemic

Master internship in machine learning applied to signal processing

**Supervision and contact:** Barbara Pascal, [barbara.pascal@cnrs.fr](mailto:barbara.pascal@cnrs.fr), CNRS, Researcher, Sébastien Bourguignon, [sebastien.bourguignon@ec-nantes](mailto:sebastien.bourguignon@ec-nantes.fr), Centrale Nantes, Professor.

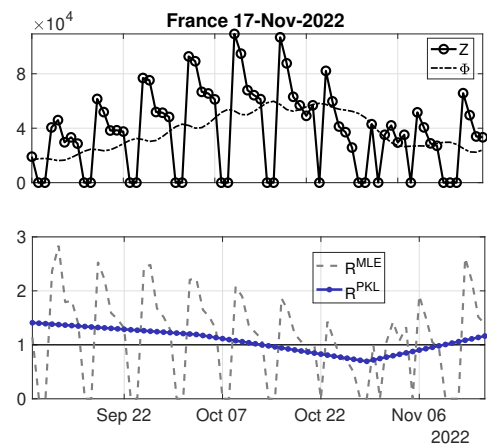
**Application:** Send a CV, master grades, references and motivations.

**Location:** Laboratoire des Sciences du Numérique de Nantes (LS2N), École Centrale Nantes.

**Duration and dates:** 4 to 6 months in 2025.

**Context:** Epidemics, striking heavily our societies both on the sanitary, economic and societal sides, are a major public health issue. The crisis triggered by the Covid-19 pandemic emphasized the need for accurate, robust and real-time tools for assessing epidemic intensity.

In practice, health authorities collect daily new infection counts  $Z_t$  at days  $t = 1, \dots, T$ . As an example, daily Covid-19 infections in France during ten weeks is displayed on the right figure, top plot. Yet, direct observation of these infections counts is not sufficient to precisely assess the intensity of Covid-19 pandemic in real-time for two main reasons: first, infections constitute a consequence of the virus spread, but quantifying the *instantaneous* dynamics of the epidemics requires to resort to more advanced epidemiological indicators; second, the reported counts are corrupted by *severe administrative noise*, leading for example to drastically underestimated counts during week-ends and days-off, which severely hampers direct quantification of the epidemic intensity.

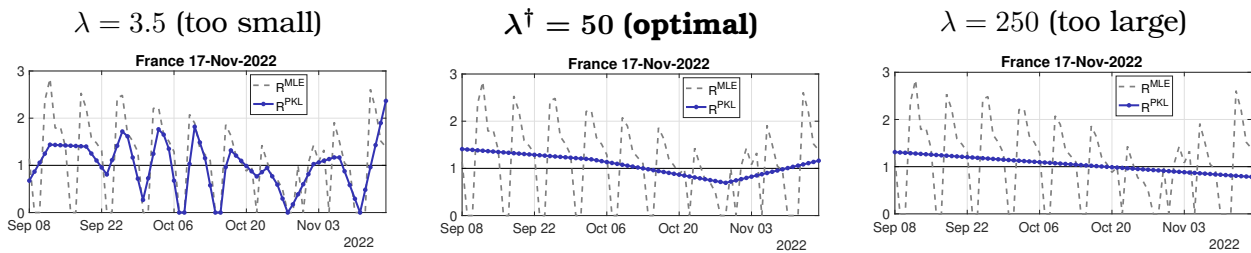


A key epidemic indicator is the basic reproduction number  $R_0$  of an epidemic, defined as the average number of secondary infections caused by one standard contagious individual [1]. Relaxed into a daily indicator  $R_t$  at day  $t$  it is called the *effective reproduction number*, providing one of the most widely used tools to monitor the intensity of virus propagation in a population: when  $R_t > 1$  (resp.  $R_t < 1$ ) the number of cases is growing (resp. decreasing) exponentially [2, 3]. Direct estimation of  $R_t$  from the low quality reported counts exemplified in the top plot yields the Maximum Likelihood Estimate (MLE) [2], represented as the dashed gray curve in the bottom plot, which follows closely the underestimation during week-ends and days-off observed in the reported counts, not reflecting the effective virus spread. To design an accurate and robust to low quality data estimate of  $R_t$ , a state-of-the-art epidemiological model, proposed in [3], has been combined with advanced signal processing tools to design a variational estimator [4]

$$\hat{\mathbf{R}}^{\text{PKL}} = \underset{\mathbf{R} \in \mathbb{R}_+^T}{\operatorname{argmin}} -\log \mathcal{L}(\mathbf{Z}, \mathbf{R}) + \lambda \mathcal{T}(\mathbf{R}) \quad (1)$$

balancing the fidelity to the model, through a Poisson likelihood  $\mathcal{L}$ , and regularity constraints, through the regularization term  $\mathcal{T}$ . This *Penalized Kullback-Leibler* (PKL) estimate, solid blue curve in the bottom plot, varies continuously and independently of the large fluctuations caused by misreported counts. Not only this variational estimate behaves far more realistically from an epidemiological point of view, but also, it appears robust to administrative noise.

**Challenge:** A major bottleneck to the practical use of the Penalized Kullback-Leibler estimate of Equation (1) is the necessary fine-tuning of the *regularization parameter*  $\lambda > 0$  controlling



the amount of regularity enforced: if too small, the estimate is not robust to noise in reported counts, if too large there is a risk of oversmoothing, as illustrated on the figure above.

**Objectives:** The main goal of this internship is to implement an *unrolled* version of the Chambolle-Pock scheme [4, Algorithm 1], which consists in reframing the variational estimate (1) into a trainable neural network, leveraging recent connections established between proximal algorithms and deep learning [5, 6, 7]. The regularization parameter  $\lambda$  is then interpreted as a weight of the unrolled network to be learned from a *training dataset*.

The accuracy of a neural network-based estimator depends to a large extent on the availability of a large and diverse training dataset. The second objective of the internship is thus to leverage the synthetic infection counts generation procedure proposed in [8] to build an annotated dataset containing pairs of infection counts  $\mathbf{Z}$  and associated ground truth reproduction number  $\mathbf{R}$ , on which to learn the weights of the unrolled network.

Finally, the developed model-informed deep network will be applied to real Covid-19 data, made available by the Johns Hopkins University <https://coronavirus.jhu.edu/>.

**Prerequisite:** The recruited intern is expected to be at ease with the basic concepts of statistics, machine learning and convex optimization. Knowledge in convex nonsmooth optimization, with a special focus on proximity operators, would be a plus. Good programming skills in Python are required, in particular in the use of PyTorch environment.

## References

- [1] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.
- [2] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS one*, 2(8):e758, 2007.
- [3] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512, 2013.
- [4] P. Abry, N. Pustelnik, S. Roux, P. Jensen, P. Flandrin, R. Gribonval, C.-G. Lucas, É. Guichard, P. Borgnat, and N. Garnier. Spatial and temporal regularization to estimate COVID-19 reproduction number  $R(t)$ : Promoting piecewise smoothness via convex optimization. *PLOS One*, 15(8):e0237901, 2020.
- [5] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th Int. Conf. Machine Learning*, pages 399–406, 2010.
- [6] P. L. Combettes and J.-C. Pesquet. Deep neural network structures solving variational inequalities. *Set-Valued and Variational Analysis*, 28:491–518, 2020.
- [7] V. Monga, Y. Li, and Y. C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.
- [8] J. Du, B. Pascal, and P. Abry. Compared performance of Covid19 reproduction number estimators based on realistic synthetic data. Grenoble, France, Aug.28 - Sept. 1 2023. URL <https://hal.science/hal-04032614v2/document>.