Epidemic monitoring:

Estimation of the reproduction number of Covid19

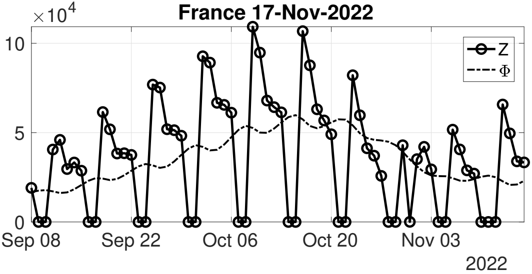
DATASIM

January 25th 2023

Barbara Pascal

Plots of Section III are reproduced with courtesy of N. Pustelnik and J.-C. Pesquet.

Data: counts of daily new infections

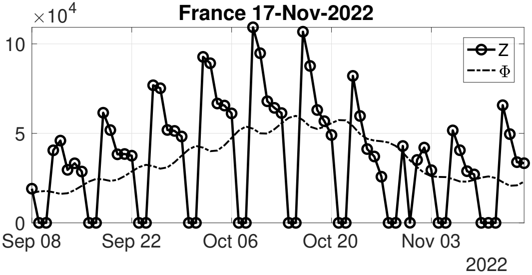


data from National Health Agencies collected by Johns Hopkins University

$\Longrightarrow$ number of cases not informative enough: need to capture the **dynamics**

Motivation and context: pandemic surveillance

Data: counts of daily new infections



data from National Health Agencies collected by Johns Hopkins University

$\Longrightarrow$ number of cases not informative enough: need to capture the **dynamics**

Goal: design adapted counter measures and evaluate their effectiveness

$\longrightarrow$ efficient monitoring tools — *epidemiological model,*

$\longrightarrow$ robust to low quality of the data — *managing erroneous counts,*

$\longrightarrow$ **(bonus)** accompanied by reliable confidence level — *credibility intervals.*

Outline

I. Epidemic modeling (Cori et al., 2013, *Am. Journal of Epidemiology*)

II. Reproduction number estimation (Pascal et al., 2022, *Trans. Sig. Process.*)

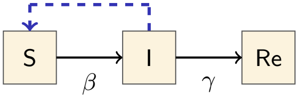
- **A)** maximum likelihood principle
- **B)** variational approaches

III. Nonsmooth convex optimization (Boyd et al., 2004, *Cambridge University Press*)

- **A)** basic tools and concepts
- **B)** algorithms

IV. Conclusion & Perspectives

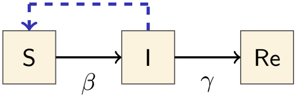
Susceptible-Infected-Recovered (SIR), among *compartmental models*



– ODE: $\frac{\mathrm{d}S_t}{\mathrm{d}t} = -\beta S_t I_t$, $\frac{\mathrm{d}I_t}{\mathrm{d}t} = \beta S_t I_t - \gamma I_t$, $\frac{\mathrm{d}Re_t}{\mathrm{d}t} = \gamma I_t$

– Stochastic model: likelihood maximization to infer β, γ

Susceptible-Infected-Recovered (SIR), among *compartmental models*



– ODE: $\frac{\mathrm{d}S_t}{\mathrm{d}t} = -\beta S_t I_t$, $\frac{\mathrm{d}I_t}{\mathrm{d}t} = \beta S_t I_t - \gamma I_t$, $\frac{\mathrm{d}Re_t}{\mathrm{d}t} = \gamma I_t$

– Stochastic model: likelihood maximization to infer β, γ

Limitations:

- refinement needed to get socially realistic model
- quadratic increase of the number of parameters
- Bayesian framework: heavy computational burden
- need consolidated and accurate datasets

✗ not adapted to real-time monitoring of Covid19 pandemic

Definition. The reproduction number associated to an epidemic is

"the averaged number of secondary cases generated by a typical infectious individual"

(Cori et al., 2013, *Am. Journal of Epidemiology*; Liu et al., 2018, *PNAS*)

Definition. The reproduction number associated to an epidemic is

"the averaged number of secondary cases generated by a typical infectious individual"

(Cori et al., 2013, *Am. Journal of Epidemiology*; Liu et al., 2018, *PNAS*)

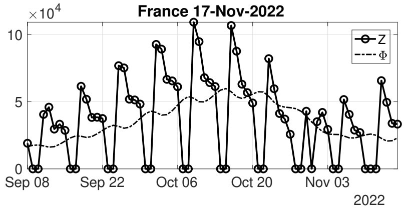
Interpretation. At day t

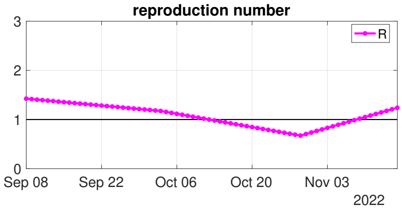
$\mathrm{R}_t > 1$ the virus propagates at exponential speed,

$\mathrm{R}_t < 1$ the epidemic shrinks with an exponential decay,

$\mathrm{R}_t = 1$ the epidemic is stable.

$\Longrightarrow$ one single indicator accounting for the overall pandemic mechanism

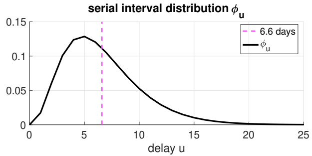




Principle: Z_t new infections at day t

$$\mathbb{E}\left[\mathrm{Z}_t\right] = \mathrm{R}_t \Phi_t, \quad \Phi_t = \sum_{u=1}^{\tau_\Phi} \phi_u \mathrm{Z}_{t-u}$$

with Φ_t global "infectiousness" in the population



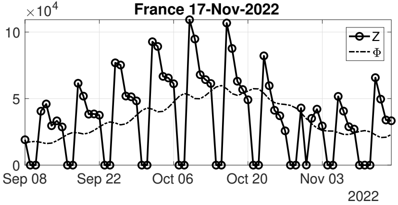
$\{\phi_u\}_{u=1}^{\tau_\Phi}$ distribution of delay between onset of symptoms in primary and secondary cases

Gamma distribution truncated at 25 days, of mean 6.6 days and standard deviation 3.5 days

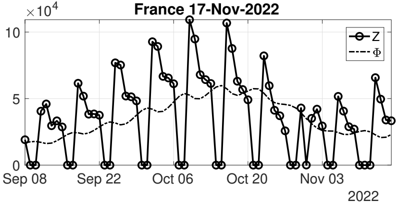
Data: daily counts $\mathbf{Z} = (Z_1, \ldots, Z_T)$

Model: Poisson distribution

$$\mathbb{P}(\mathsf{Z}_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, \mathsf{R}_t) = \frac{(\mathsf{R}_t \Phi_t)^{\mathsf{Z}_t} \mathrm{e}^{-\mathsf{R}_t \Phi_t}}{\mathsf{Z}_t!}$$



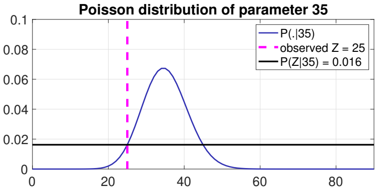
Data: daily counts $\mathbf{Z} = (Z_1, \ldots, Z_T)$

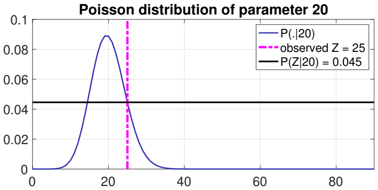


Model: Poisson distribution

$$\mathbb{P}(Z_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, R_t) = \frac{(R_t \Phi_t)^{Z_t} e^{-R_t \Phi_t}}{Z_t!}$$

Maximum Likelihood Principle: If one observes a given Z_t, how to infer R_t?





observation $Z_t = 25 \quad \Rightarrow \quad \widehat{R}_t^{\mathrm{MLE}} \Phi_t =$

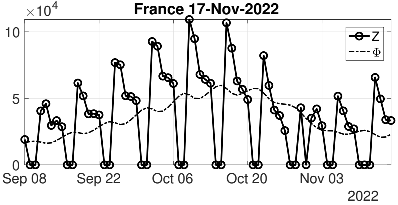
II. Reproduction number estimation

maximum likelihood principle

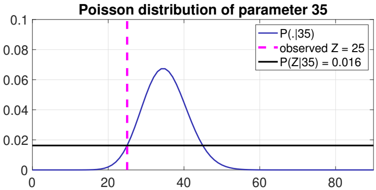
Data: daily counts $\mathbf{Z} = (Z_1, \ldots, Z_T)$

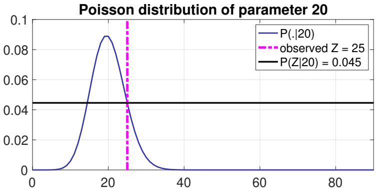
Model: Poisson distribution

$$\mathbb{P}(\mathrm{Z}_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, \mathrm{R}_t) = \frac{(\mathrm{R}_t \Phi_t)^{\mathrm{Z}_t} \mathrm{e}^{-\mathrm{R}_t \Phi_t}}{\mathrm{Z}_t!}$$



Maximum Likelihood Principle: If one observes a given Z_t, how to infer R_t?





observation $\mathrm{Z}_t = 25 \quad \Rightarrow \quad \widehat{\mathrm{R}}_t^{\mathrm{MLE}} \Phi_t = 25$

Maximum Likelihood Estimator. $\widehat{\mathsf{R}}_t^{\mathrm{MLE}} := \underset{\mathsf{R}_t}{\operatorname{argmax}}\, \mathbb{P}(\mathsf{Z}_t|\mathbf{Z}_{t-\tau_\Phi:t-1}, \mathsf{R}_t)$

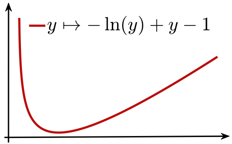
$$\begin{aligned}
\ln\left(\mathbb{P}(\mathsf{Z}_t|\mathbf{Z}_{t-\tau_\Phi:t-1}, \mathsf{R}_t)\right) &= \mathsf{Z}_t \ln(\mathsf{R}_t\Phi_t) - \mathsf{R}_t\Phi_t - \ln(\mathsf{Z}_t!) \\
&\underset{\mathsf{Z}_t \gg 1}{\simeq} \mathsf{Z}_t \ln(\mathsf{R}_t\Phi_t) - \mathsf{R}_t\Phi_t - \mathsf{Z}_t\ln(\mathsf{Z}_t) + \mathsf{Z}_t \\
&\underset{\text{(def.)}}{=} -\mathrm{d}_{\mathrm{KL}}(\mathsf{Z}_t|\mathsf{R}_t\Phi_t)
\end{aligned}$$

Maximum Likelihood Estimator. $\widehat{\mathsf{R}}_t^{\mathsf{MLE}} := \underset{\mathsf{R}_t}{\operatorname{argmax}}\ \mathbb{P}(\mathsf{Z}_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, \mathsf{R}_t)$

$$\begin{aligned}\ln\left(\mathbb{P}(\mathsf{Z}_t|\mathbf{Z}_{t-\tau_\Phi:t-1}, \mathsf{R}_t)\right) &= \mathsf{Z}_t \ln(\mathsf{R}_t\Phi_t) - \mathsf{R}_t\Phi_t - \ln(\mathsf{Z}_t!)\\ &\underset{\mathsf{Z}_t \gg 1}{\simeq} \mathsf{Z}_t \ln(\mathsf{R}_t\Phi_t) - \mathsf{R}_t\Phi_t - \mathsf{Z}_t\ln(\mathsf{Z}_t) + \mathsf{Z}_t\\ &\underset{\text{(def.)}}{=} -\mathsf{d}_{\mathsf{KL}}(\mathsf{Z}_t|\mathsf{R}_t\Phi_t)\end{aligned}$$

Definition. (Kullback-Leibler divergence)

$$\mathsf{d}_{\mathsf{KL}}(\mathsf{Z}|\mathsf{p}) = \begin{cases} \mathsf{Z}\ln(\mathsf{Z}/\mathsf{p}) + \mathsf{p} - \mathsf{Z} & \text{if } \mathsf{Z} > 0 \,\&\, \mathsf{p} > 0\\ \mathsf{p} & \text{if } \mathsf{Z} = 0 \,\&\, \mathsf{p} \geq 0\\ \infty & \text{otherwise.}\end{cases}$$

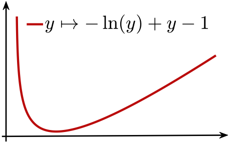


Maximum Likelihood Estimator. $\widehat{\mathrm{R}}_t^{\mathrm{MLE}} := \underset{\mathrm{R}_t}{\mathrm{argmax}}\ \mathbb{P}(\mathrm{Z}_t|\mathbf{Z}_{t-\tau_\Phi:t-1}, \mathrm{R}_t)$

$$\begin{aligned}\ln\left(\mathbb{P}(\mathrm{Z}_t|\mathbf{Z}_{t-\tau_\Phi:t-1}, \mathrm{R}_t)\right) &= \mathrm{Z}_t \ln(\mathrm{R}_t\Phi_t) - \mathrm{R}_t\Phi_t - \ln(\mathrm{Z}_t!)\\ &\underset{\mathrm{Z}_t\gg 1}{\simeq} \mathrm{Z}_t \ln(\mathrm{R}_t\Phi_t) - \mathrm{R}_t\Phi_t - \mathrm{Z}_t\ln(\mathrm{Z}_t) + \mathrm{Z}_t\\ &\underset{\text{(def.)}}{=} -\mathrm{d}_{\mathrm{KL}}(\mathrm{Z}_t|\mathrm{R}_t\Phi_t)\end{aligned}$$

Definition. (Kullback-Leibler divergence)

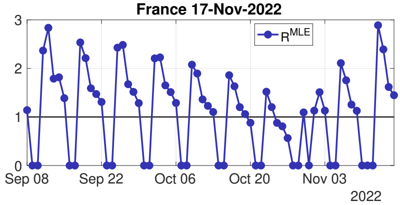
$$\mathrm{d}_{\mathrm{KL}}(\mathrm{Z}|\mathrm{p}) = \begin{cases} \mathrm{Z}\ln(\mathrm{Z}/\mathrm{p}) + \mathrm{p} - \mathrm{Z} & \text{if } \mathrm{Z} > 0 \,\&\, \mathrm{p} > 0\\ \mathrm{p} & \text{if } \mathrm{Z} = 0 \,\&\, \mathrm{p} \geq 0\\ \infty & \text{otherwise.}\end{cases}$$



$$\boxed{\widehat{\mathrm{R}}_t^{\mathrm{MLE}} = \underset{\mathrm{R}_t}{\mathrm{argmin}}\ \mathrm{d}_{\mathrm{KL}}(\mathrm{Z}_t|\mathrm{R}_t\Phi_t) = \mathrm{Z}_t/\Phi_t = \mathrm{Z}_t/\sum_{u=1}^{\tau_\Phi}\phi_u \mathrm{Z}_{t-u} \quad \text{ratio of moving averages}}$$

$$\widehat{\mathrm{R}}_t^{\mathrm{MLE}} = \underset{\mathrm{R}_t}{\mathrm{argmin}}\ \mathrm{d}_{\mathrm{KL}}(\mathrm{Z}_t | \mathrm{R}_t \Phi_t) = \mathrm{Z}_t / \Phi_t = \mathrm{Z}_t / \sum_{u=1}^{\tau_\Phi} \phi_u \mathrm{Z}_{t-u} \quad \text{ratio of moving averages}$$

Estimation.



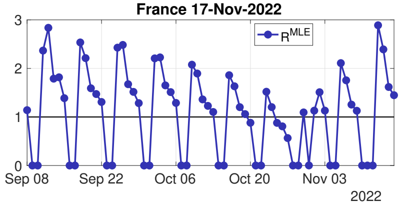
- ♣ huge variability along time/ no local trend
- ♣ not robust to pseudo-periodicity/ misreported counts

II. Reproduction number estimation

maximum likelihood principle

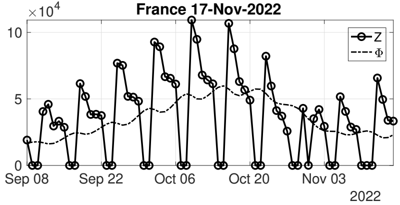
$$\widehat{\mathrm{R}}_t^{\mathrm{MLE}} = \underset{\mathrm{R}_t}{\operatorname{argmin}}\ \mathrm{d}_{\mathrm{KL}}(\mathrm{Z}_t|\mathrm{R}_t\Phi_t) = \mathrm{Z}_t/\Phi_t = \mathrm{Z}_t/\sum_{u=1}^{\tau_\Phi}\phi_u \mathrm{Z}_{t-u} \quad \text{ratio of moving averages}$$

Estimation.



- ♣ huge variability along time/ no local trend
- ♣ not robust to pseudo-periodicity/ misreported counts

Explanation.



New infection counts **Z** are corrupted by

- missing samples,
- non meaningful negative counts,
- retrospected cumulated counts,
- pseudo-seasonality effects.

State-of-the-art in epidemiology. Smoothing over a temporal window

$$\widehat{\mathrm{R}}_{t,s}^{\mathrm{MLE}}, \text{ with } s = 7 \text{ days}$$

(Cori et al., 2013, *Am. Journal of Epidemiology*)

$\Longrightarrow$ not able to detect rapid surge, nor fast decrease following sanitary restrictions

State-of-the-art in epidemiology. Smoothing over a temporal window

$$\widehat{\mathrm{R}}^{\mathrm{MLE}}_{t,s}, \text{ with } s = 7 \text{ days}$$

(Cori et al., 2013, *Am. Journal of Epidemiology*)

$\Longrightarrow$ not able to detect rapid surge, nor fast decrease following sanitary restrictions

---

Penalized likelihood. Regularization through nonlinear filtering

$$\widehat{\mathbf{R}}^{\mathrm{PKL}} = \underset{\mathbf{R}\in\mathbb{R}^T}{\operatorname{argmin}} \sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}}\left(\mathrm{Z}_t \,|\mathrm{R}_t\Phi_t\right) + \lambda_{\mathrm{R}}\mathcal{P}(\mathbf{R}) \quad \text{(penalized Kullback-Leibler)}$$

with $\mathcal{P}(\mathbf{R})$ favoring some temporal regularity (Abry et al., 2020, *PlosOne*)

State-of-the-art in epidemiology. Smoothing over a temporal window

$$\widehat{\mathrm{R}}^{\mathrm{MLE}}_{t,s}, \text{ with } s = 7 \text{ days}$$

(Cori et al., 2013, *Am. Journal of Epidemiology*)

$\Longrightarrow$ not able to detect rapid surge, nor fast decrease following sanitary restrictions

---

Penalized likelihood. Regularization through nonlinear filtering

$$\widehat{\mathbf{R}}^{\mathrm{PKL}} = \underset{\mathbf{R}\in\mathbb{R}^T}{\operatorname{argmin}} \sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}}\left(\mathrm{Z}_t \,|\mathrm{R}_t\Phi_t\right) + \lambda_{\mathrm{R}}\mathcal{P}(\mathbf{R}) \quad \text{(penalized Kullback-Leibler)}$$

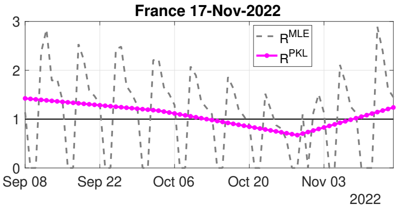
with $\mathcal{P}(\mathbf{R})$ favoring some temporal regularity (Abry et al., 2020, *PlosOne*)

$$\mathcal{P}(\mathbf{R}) = \|\mathbf{D}_2\mathbf{R}\|_1$$

$$(\mathbf{D}_2\mathbf{R})_t = R_{t+1} - 2R_t + R_{t-1}$$

2nd order derivative & ℓ_1-norm

$\Longrightarrow$ piecewise linearity

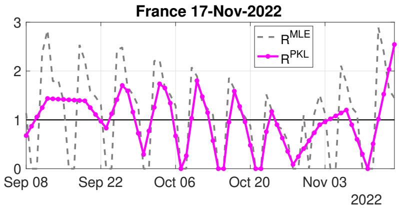


captures global **trend**, more **regular** than MLE, detect **ruptures**

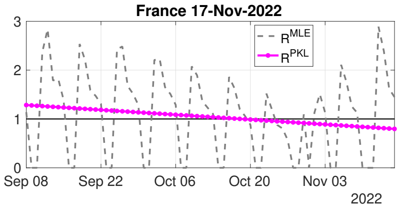
Penalized likelihood.

$$\widehat{\mathbf{R}}^{\mathrm{PKL}} = \underset{\mathbf{R}\in\mathbb{R}^T}{\mathrm{argmin}} \sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}}\left(\mathrm{Z}_t \,|\mathrm{R}_t\Phi_t\right) + \lambda_{\mathrm{R}}\mathcal{P}(\mathbf{R}) \quad \text{(penalized Kullback-Leibler)}$$

Balance between data-fidelity and temporal regularity.



small λ_{R}



large λ_{R}

Data. Daily reported counts $\mathbf{Z} = (Z_1, \dots, Z_T)$

Model. Poisson distribution $\mathbb{P}(Z_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, R_t) = \dfrac{(R_t \Phi_t)^{Z_t} e^{-(R_t \Phi_t)}}{Z_t!}$

Data. Daily reported counts $\mathbf{Z} = (Z_1, \ldots, Z_T)$

Model. Poisson distribution $\mathbb{P}(Z_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, R_t) = \dfrac{(R_t \Phi_t)^{Z_t} e^{-(R_t \Phi_t)}}{Z_t!}$

Penalized Kullback-Leibler divergence

$$\widehat{\mathbf{R}}^{\mathrm{PKL}} = \underset{\mathbf{R} \in \mathbb{R}^T}{\operatorname{argmin}} \sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}}\left(Z_t \,|\, R_t \Phi_t\right) + \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 \Longrightarrow \text{estimates } \underline{\text{piecewise linear}}\ R_t$$

Data. Daily reported counts $\mathbf{Z} = (Z_1, \ldots, Z_T)$

Model. Poisson distribution $\mathbb{P}(Z_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, R_t) = \dfrac{(R_t \Phi_t)^{Z_t} e^{-(R_t \Phi_t)}}{Z_t!}$

Penalized Kullback-Leibler divergence

$$\widehat{\mathbf{R}}^{\text{PKL}} = \underset{\mathbf{R} \in \mathbb{R}^T}{\operatorname{argmin}} \sum_{t=1}^{T} d_{\text{KL}} \left(Z_t \,|\, R_t \Phi_t \right) + \lambda_{\text{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 \Longrightarrow \text{estimates } \underline{\text{piecewise linear}}\ R_t$$

properties of the objective function:

- sum of convex functions composed with linear operators $\Longrightarrow$ globally convex;
- feasible domain: $\{\text{if } Z_t > 0,\ R_t \Phi_t > 0, \text{ else } R_t \Phi_t \geq 0\}$;
- $p_t \mapsto d_{\text{KL}} \left(Z_t \,|\, p_t \right)$ is strictly-convex.

Data. Daily reported counts $\mathbf{Z} = (Z_1, \ldots, Z_T)$

Model. Poisson distribution $\mathbb{P}(Z_t | \mathbf{Z}_{t-\tau_\Phi:t-1}, R_t) = \dfrac{(R_t \Phi_t)^{Z_t} e^{-(R_t \Phi_t)}}{Z_t!}$

Penalized Kullback-Leibler divergence

$$\widehat{\mathbf{R}}^{\mathrm{PKL}} = \underset{\mathbf{R} \in \mathbb{R}^T}{\mathrm{argmin}} \sum_{t=1}^{T} d_{\mathrm{KL}}\left(Z_t \,|\, R_t \Phi_t\right) + \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 \implies \text{estimates piecewise linear } R_t$$

properties of the objective function:

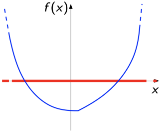
- sum of convex functions composed with linear operators $\Longrightarrow$ globally convex;
- feasible domain: $\{\text{if } Z_t > 0,\ R_t \Phi_t > 0, \text{ else } R_t \Phi_t \geq 0\}$;
- $p_t \mapsto d_{\mathrm{KL}}\left(Z_t \,|\, p_t\right)$ is strictly-convex.

Theorem (Pascal et al., 2022, *Trans. Sig. Process.*)

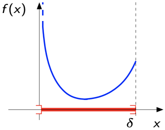
\+ The minimization problem has at least one solution $\widehat{\mathbf{R}}^{\mathrm{PKL}}$.

\+ The estimated time-varying Poisson intensity $\widehat{p}_t^{\mathrm{PKL}} = \widehat{R}_t^{\mathrm{PKL}} \Phi_t$ is unique.

Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, the domain of f is $\operatorname{dom} f = \{\boldsymbol{x} \in \mathbb{R}^T \mid f(x) < \infty\}$



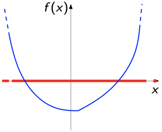
$\operatorname{dom} f = \mathbb{R}$



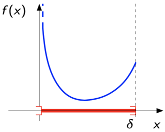
$\operatorname{dom} f =]0, \delta]$

If $\operatorname{dom} f \neq \emptyset$, f is said to be proper.

Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, the domain of f is $\operatorname{dom} f = \{\boldsymbol{x} \in \mathbb{R}^T \mid f(x) < \infty\}$



$\operatorname{dom} f = \mathbb{R}$



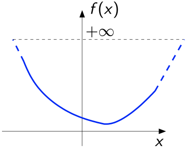
$\operatorname{dom} f =]0, \delta]$

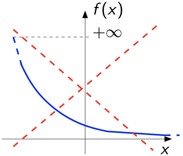
If $\operatorname{dom} f \neq \emptyset$, f is said to be proper.

Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$. If

$$\lim_{\|\boldsymbol{x}\|_2 \to \infty} f(\boldsymbol{x}) = \infty$$

then f is said to be coercive.





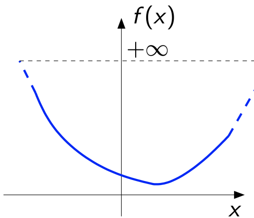
Theorem. If $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ is proper, continuous on $\operatorname{dom} f$, coercive then

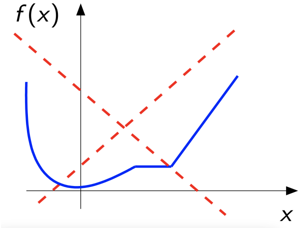
$$\operatorname{Argmin} f = \{\boldsymbol{x} \in \operatorname{dom} f \mid f(\boldsymbol{x}) = \inf\ f\}$$

is nonempty. If f is convex, then $\operatorname{Argmin} f$ is convex.

Theorem. If $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ is proper, C^1 on $\operatorname{dom} f$, coercive, and convex

$$\widehat{\boldsymbol{x}} \in \operatorname{Argmin} f \quad \Leftrightarrow \quad \nabla f(\widehat{\boldsymbol{x}}) = 0$$



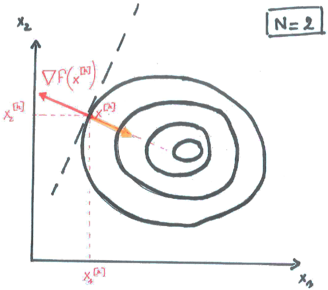


Gradient descent algorithm.

$f : \mathbb{R}^T \to \mathbb{R}$, continuously differentiable

for $k = 1, 2 \ldots$ **do**

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma \nabla f(\boldsymbol{x}^{[k]})$$

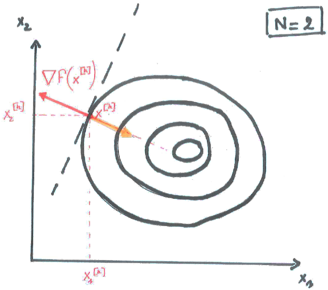


Gradient descent algorithm.

$f : \mathbb{R}^T \to \mathbb{R}$, continuously differentiable

for $k = 1, 2 \dots$ **do**

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma \nabla f(\boldsymbol{x}^{[k]})$$



Definition. Let $f : \mathbb{R}^T \to \mathbb{R}$, continuously differentiable, and $\beta > 0$. If

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^T, \quad \|\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v})\|_2 \leq \beta \|\boldsymbol{u} - \boldsymbol{v}\|_2$$

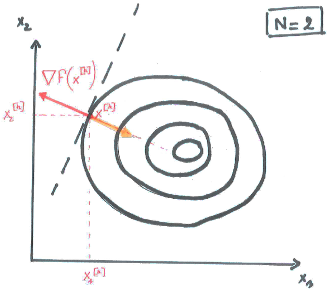
f is said to be $\underline{\beta\text{-smooth}}$, i.e., f has a β-Lipschitz gradient.

Gradient descent algorithm.

$f : \mathbb{R}^T \to \mathbb{R}$, continuously differentiable

for $k = 1, 2 \ldots$ **do**

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma \nabla f(\boldsymbol{x}^{[k]})$$



Definition. Let $f : \mathbb{R}^T \to \mathbb{R}$, continuously differentiable, and $\beta > 0$. If

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^T, \quad \|\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v})\|_2 \leq \beta \|\boldsymbol{u} - \boldsymbol{v}\|_2$$

f is said to be $\underline{\beta\text{-smooth}}$, i.e., f has a β-Lipschitz gradient.

Theorem. If $f : \mathbb{R}^T \to \mathbb{R}$ is convex, coercive, C^1, and β-smooth, with $\beta > 0$, then

$$\exists \widehat{\boldsymbol{x}} \in \mathbb{R}^T, \quad \lim_{k \to \infty} \boldsymbol{x}^{[k]} = \widehat{\boldsymbol{x}} \quad \text{with} \quad \nabla f(\widehat{\boldsymbol{x}}) = 0.$$

Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper, the subdifferential of f at $\boldsymbol{x}$ is

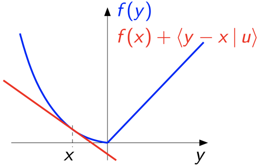
$$\partial f(\boldsymbol{x}) = \{\boldsymbol{u} \in \mathbb{R}^T \mid \forall \boldsymbol{y} \in \mathbb{R}^T, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle\}$$

$\boldsymbol{u} \in \partial f(\boldsymbol{x})$ is a subgradient of f at $\boldsymbol{x}$.

Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper, the <u>subdifferential</u> of f at $\boldsymbol{x}$ is

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{u} \in \mathbb{R}^T \mid \forall \boldsymbol{y} \in \mathbb{R}^T, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle\}$$

$\boldsymbol{u} \in \partial f(\boldsymbol{x})$ is a <u>subgradient</u> of f at $\boldsymbol{x}$.



$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$$

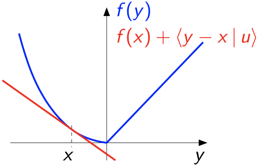
III. Nonsmooth convex optimization

basic tools and concepts

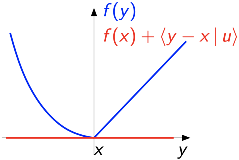
Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper, the <u>subdifferential</u> of f at $\boldsymbol{x}$ is

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{u} \in \mathbb{R}^T \mid \forall \boldsymbol{y} \in \mathbb{R}^T, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle\}$$

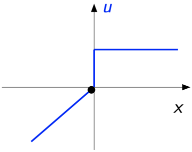
$\boldsymbol{u} \in \partial f(\boldsymbol{x})$ is a <u>subgradient</u> of f at $\boldsymbol{x}$.



$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$



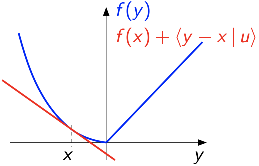
$\partial f(\boldsymbol{0}) = [0, 1]$



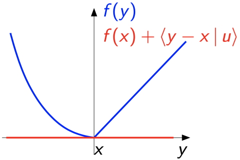
Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper, the subdifferential of f at $\boldsymbol{x}$ is

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{u} \in \mathbb{R}^T \mid \forall \boldsymbol{y} \in \mathbb{R}^T, \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle\}$$

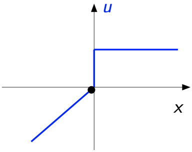
$\boldsymbol{u} \in \partial f(\boldsymbol{x})$ is a subgradient of f at $\boldsymbol{x}$.



$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$



$\partial f(\boldsymbol{0}) = [0, 1]$



Theorem. (Fermat's rule) Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ a proper function

$$\widehat{\boldsymbol{x}} \in \operatorname{Argmin} f \quad \Leftrightarrow \quad 0 \in \partial f(\widehat{\boldsymbol{x}}).$$

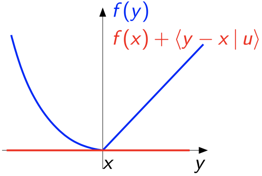
Subgradient descent algorithm.

$f : \mathbb{R}^T \to \mathbb{R}$, convex, continuous

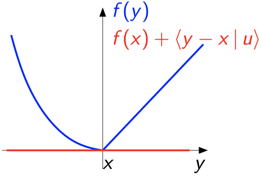
for $k = 1, 2 \ldots$ **do**

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma_k \boldsymbol{u}^{[k]}, \quad \boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k]})$$

explicit scheme: $\boldsymbol{x}^{[k+1]}$ derived from $\boldsymbol{x}^{[k]}$



Subgradient descent algorithm.



$f : \mathbb{R}^T \to \mathbb{R}$, convex, continuous

for $k = 1, 2 \ldots$ **do**

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma_k \boldsymbol{u}^{[k]}, \quad \boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k]})$$

explicit scheme: $\boldsymbol{x}^{[k+1]}$ derived from $\boldsymbol{x}^{[k]}$

Properties. For $(\boldsymbol{x}^{[k]})_{k \in \mathbb{N}}$ to converge:

- need a vanishing sequence $(\gamma_k)_{k \in \mathbb{N}}$: $\gamma_k \underset{k \to \infty}{\to} 0$;
- large number of iterations due to slow dynamics.

Explanation. $\partial f : \mathbb{R}^T \to 2^{\mathbb{R}^T}$ set-valued

Numerically instability because of ambiguity in the choice of $\boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k]})$.

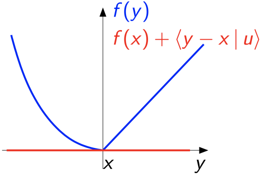
Subgradient descent algorithm.

$f : \mathbb{R}^T \to \mathbb{R}$, convex, continuous

for $k = 1, 2 \dots$ **do**

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma_k \boldsymbol{u}^{[k]}, \quad \boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k]})$$

explicit scheme: $\boldsymbol{x}^{[k+1]}$ derived from $\boldsymbol{x}^{[k]}$



Properties. For $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ to converge:

- need a vanishing sequence $(\gamma_k)_{k\in\mathbb{N}}$: $\gamma_k \underset{k\to\infty}{\to} 0$;
- large number of iterations due to slow dynamics.

Explanation. $\partial f : \mathbb{R}^T \to 2^{\mathbb{R}^T}$ set-valued

Numerically instability because of ambiguity in the choice of $\boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k]})$.

Solution. Turn to an implicit scheme

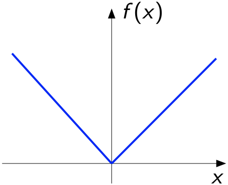
$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma_k \boldsymbol{u}^{[k]}, \quad \boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k+1]}) \Rightarrow \text{how to compute } \boldsymbol{x}^{[k+1]}?$$

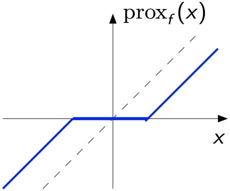
Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper, convex, continuous, $\gamma > 0$

$$\mathrm{prox}_{\gamma f}(\boldsymbol{x}) := \underset{\boldsymbol{y} \in \mathbb{R}^T}{\mathrm{argmin}} \, \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \gamma f(\boldsymbol{y})$$

is the proximity operator of γf at point $\boldsymbol{x}$.

Example.



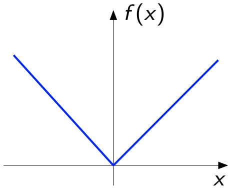


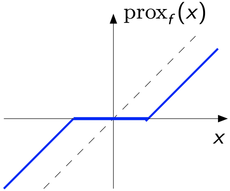
Definition. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper, convex, continuous, $\gamma > 0$

$$\mathrm{prox}_{\gamma f}(\boldsymbol{x}) := \underset{\boldsymbol{y}\in\mathbb{R}^T}{\mathrm{argmin}}\, \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \gamma f(\boldsymbol{y})$$

is the proximity operator of γf at point $\boldsymbol{x}$.

Example.





Theorem. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ a proper, convex, continuous function

$$\boldsymbol{p} = \mathrm{prox}_{\gamma f}(\boldsymbol{x}) \quad \Leftrightarrow \quad \boldsymbol{x} \in \boldsymbol{p} + \partial f(\boldsymbol{p})$$

Implicit scheme.

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma_k \boldsymbol{u}^{[k]}, \quad \boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k+1]}) \Rightarrow \text{how to compute } \boldsymbol{x}^{[k+1]}?$$

Theorem. Let $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ a proper, convex, continuous function

$$\boldsymbol{p} = \mathrm{prox}_f(\boldsymbol{x}) \quad \Leftrightarrow \quad \boldsymbol{x} \in \boldsymbol{p} + \partial f(\boldsymbol{p})$$

Solution. Apply the theorem in the $\Leftarrow$ sense with $\boldsymbol{x} = \boldsymbol{x}^{[k]}$ and $\boldsymbol{p} = \boldsymbol{x}^{[k+1]}$

$$\boldsymbol{x}^{[k]} = \boldsymbol{x}^{[k+1]} + \gamma_k \boldsymbol{u}^{[k]}, \quad \boldsymbol{u}^{[k]} \in \partial f(\boldsymbol{x}^{[k+1]})$$

Proximal point algorithm. $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$, proper , convex, continuous

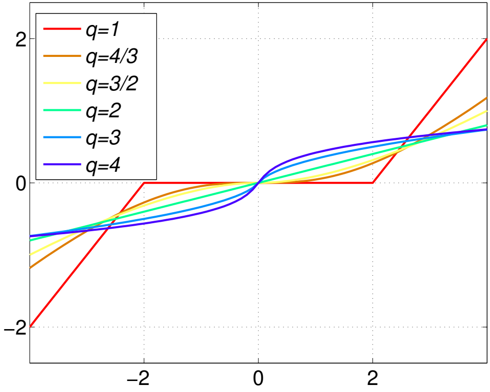
for $k = 1, 2 \ldots$ **do**

$$\boldsymbol{x}^{[k+1]} = \mathrm{prox}_{\gamma f}(\boldsymbol{x}^{[k]})$$

Theorem. For any $\gamma > 0$, $\left(\boldsymbol{x}^{[k]}\right)_{k \in \mathbb{N}}$ converges toward some $\widehat{\boldsymbol{x}} \in \mathrm{Argmin}\, f$.

Power q function with $q \geq 1$. Let $\eta > 0$, $q \in [1, +\infty[$

$$f : \mathbb{R} \to \mathbb{R} \cup \{\infty\},\ x \mapsto \eta |x|^q$$



many more explicit proximal operators at `http://proximity-operator.net/`

Property. If $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ is separable, i.e.,

$$\forall \boldsymbol{x} \in \mathbf{R}^T, \quad f(\boldsymbol{x}) = \sum_{t=1}^{T} f_t(x_t), \quad \text{with } f_t \text{ proper, convex, continuous}$$

then the proximal operator can be computed component-wise and

$$\boldsymbol{p} = \operatorname{prox}_{\gamma f}(\boldsymbol{x}) \quad \Leftrightarrow \quad \forall t = 1, \dots, T, \quad p_t = \operatorname{prox}_{\gamma f_t}(x_t).$$

Property. If $f : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ is separable, i.e.,

$$\forall \boldsymbol{x} \in \mathbf{R}^T, \quad f(\boldsymbol{x}) = \sum_{t=1}^{T} f_t(x_t), \quad \text{with } f_t \text{ proper, convex, continuous}$$

then the proximal operator can be computed component-wise and

$$\boldsymbol{p} = \mathrm{prox}_{\gamma f}(\boldsymbol{x}) \quad \Leftrightarrow \quad \forall t = 1, \dots, T, \quad p_t = \mathrm{prox}_{\gamma f_t}(x_t).$$

Problematic. $f, g : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ convex, proper, continuous

$$\underset{\boldsymbol{x} \in \mathbb{R}^T}{\text{minimize}} \;\; f(\boldsymbol{x}) + g(\boldsymbol{x}).$$

$\Rightarrow$ compute prox_{f+g}: in general intractable!

Problematic. $f, g : \mathbb{R}^T \to \mathbb{R} \cup \{\infty\}$ convex, proper, continuous

$$\underset{\boldsymbol{x} \in \mathbb{R}^T}{\text{minimize}} \; f(\boldsymbol{x}) + g(\boldsymbol{x})$$

Hypotheses. f is continuously differentiable and β-smooth, with $\beta > 0$.
g is proximable, i.e., $\text{prox}_{\gamma g}$ has an explicit formula.

Forward-backward algorithm. or “Proximal-gradient"

for $k = 1, 2 \ldots$ **do**

$$\boldsymbol{x}^{[k+1]} = \text{prox}_{\gamma g}(\boldsymbol{x}^{[k]} - \gamma \nabla f(\boldsymbol{x}^{[k]}))$$

explicit-implicit scheme: $\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \gamma \nabla f(\boldsymbol{x}^{[k]}) - \gamma \boldsymbol{u}^{[k]}, \; \boldsymbol{u}^{[k]} \in \partial g(\boldsymbol{x}^{[k+1]})$

Theorem. If $\gamma \in]0, 2/\beta[$, $\left(\boldsymbol{x}^{[k]}\right)_{k \in \mathbb{N}}$ converges toward some $\widehat{\boldsymbol{x}} \in \text{Argmin}\, f + g$.

$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}}\left(\mathrm{Z}_t \,|\, \mathrm{R}_t \Phi_t\right) + \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1$$

- each term of the functional is convex;
- ℓ_1-norm and indicative functions $\Longrightarrow$ nonsmooth; ✗ gradient descent
- gradient of $\mathrm{p}_t \mapsto \mathrm{d}_{\mathrm{KL}}\left(\mathrm{Z}_t \,|\, \mathrm{p}_t\right)$ is not Lipschitzian; ✗ forward-backward
- linear operator $\mathbf{D}_2 \Longrightarrow$ no explicit form for $\mathrm{prox}_{\|\mathbf{D}_2\cdot\|_1}$ ♣ need splitting

$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \; \sum_{t=1}^{T} \mathrm{d_{KL}}\left(\mathrm{Z}_t \,|\, \mathrm{R}_t \boldsymbol{\Phi}_t\right) + \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1$$

- each term of the functional is convex;
- ℓ_1-norm and indicative functions $\Longrightarrow$ nonsmooth; ✗ gradient descent
- gradient of $\mathrm{p}_t \mapsto \mathrm{d_{KL}}\left(\mathrm{Z}_t \,|\, \mathrm{p}_t\right)$ is not Lipschitzian; ✗ forward-backward
- linear operator $\mathbf{D}_2 \Longrightarrow$ no explicit form for $\mathrm{prox}_{\|\mathbf{D}_2 \cdot\|_1}$ ♣ need splitting

$$\Longleftrightarrow \underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \quad f(\mathbf{R}|\mathbf{Z}) + h(\mathbf{D}_2\mathbf{R}), \quad \mathbf{D}_2 \text{ linear; } f, h \text{ proximable}$$

$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \sum_{t=1}^{T} \mathrm{d_{KL}}\left(\mathsf{Z}_t \,|\, \mathsf{R}_t \Phi_t\right) + \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1$$

- each term of the functional is convex;
- ℓ_1-norm and indicative functions $\Longrightarrow$ nonsmooth; ✗ gradient descent
- gradient of $\mathsf{p}_t \mapsto \mathrm{d_{KL}}\left(\mathsf{Z}_t \,|\, \mathsf{p}_t\right)$ is not Lipschitzian; ✗ forward-backward
- linear operator $\mathbf{D}_2 \Longrightarrow$ no explicit form for $\mathrm{prox}_{\|\mathbf{D}_2 \cdot\|_1}$ ♣ need splitting

$$\Longleftrightarrow \underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \quad f(\mathbf{R}|\mathbf{Z}) + h(\mathbf{D}_2\mathbf{R}), \quad \mathbf{D}_2 \text{ linear; } f, h \text{ proximable}$$

Primal-dual algorithm (Chambolle et al., 2011, *Int. Conf. Comput. Vis.*)

for $k = 1, 2 \dots$ **do**

$$\mathbf{Q}^{[k+1]} = \mathrm{prox}_{\sigma h^*}(\mathbf{Q}^{[k]} + \sigma \mathbf{D}_2 \overline{\mathbf{R}}^{[k]} \qquad \text{dual}$$

$$\mathbf{R}^{[k+1]} = \mathrm{prox}_{\tau f(\cdot|\mathbf{Z})}(\mathbf{R}^{[k+1]} - \tau \mathbf{D}_2^* \mathbf{Q}^{[k+1]}) \qquad \text{primal}$$

$$\overline{\mathbf{R}}^{[k+1]} = 2\mathbf{R}^{[k+1]} - \mathbf{R}^{[k]} \qquad \text{auxiliary}$$

Theorem. If $\tau\sigma\|\mathbf{D}_2\|_{\mathrm{op}}^2 < 1$, $\left(\mathbf{R}^{[k]}\right)_{k\in\mathbb{N}}$ converges toward $\widehat{\mathbf{R}}^{\mathrm{PKL}}$.

New infection counts per county: $\mathbf{Z} = \left\{ \mathrm{Z}_t^{(d)},\, d \in [1, D],\, t \in [1, T] \right\}$

$\Rightarrow$ multivariate time-varying reproduction number $\mathrm{R}_t^{(d)}$

New infection counts per county: $\mathbf{Z} = \left\{ \mathsf{Z}_t^{(d)},\ d \in [1, D],\ t \in [1, T] \right\}$

$\Rightarrow$ multivariate time-varying reproduction number $\mathsf{R}_t^{(d)}$

Multivariate extended penalized Kullback-Leibler

$$\widehat{\mathbf{R}} = \underset{\mathbf{R} \in \mathbb{R}^{D \times T}}{\operatorname{argmin}} \sum_{d=1}^{D} \sum_{t=1}^{T} \mathsf{d}_{\mathsf{KL}} \left(\mathsf{Z}_t^{(d)} \middle| \mathsf{R}_t^{(d)} \Phi_t^{(d)} \right) + \lambda_{\mathsf{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 + \lambda_{\text{space}} \|\mathbf{G}\mathbf{R}\|_1$$

$\Longrightarrow$ $\|\mathbf{G}\mathbf{R}\|_1$ favors **piecewise constancy** in space

IV. Conclusion & Perspectives

New infection counts per county: $\mathbf{Z} = \left\{ \mathrm{Z}_t^{(d)},\ d \in [1, D],\ t \in [1, T] \right\}$

$\Rightarrow$ multivariate time-varying reproduction number $\mathrm{R}_t^{(d)}$

Multivariate extended penalized Kullback-Leibler

$$\widehat{\mathbf{R}} = \underset{\mathbf{R} \in \mathbb{R}^{D \times T}}{\operatorname{argmin}} \sum_{d=1}^{D} \sum_{t=1}^{T} \mathrm{d}_{\mathrm{KL}} \left(\mathrm{Z}_t^{(d)} \middle| \mathrm{R}_t^{(d)} \Phi_t^{(d)} \right) + \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 + {\color{magenta}\lambda_{\text{space}} \|\mathbf{GR}\|_1}$$

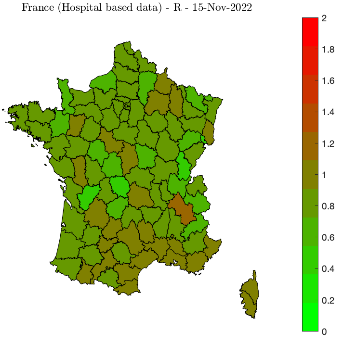
$\Longrightarrow$ ${\color{magenta}\|\mathbf{GR}\|_1}$ favors **piecewise constancy** in space

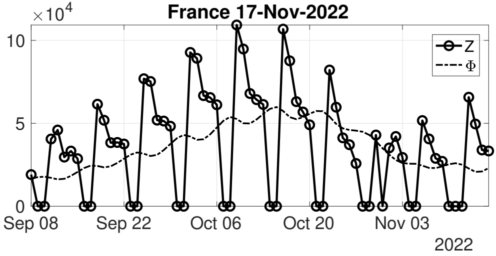
<ins>Graph Total Variation</ins>

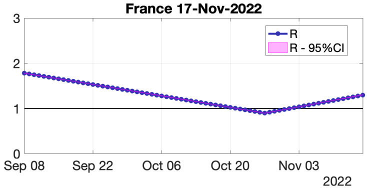
$${\color{magenta}\|\mathbf{GR}\|_1} = \sum_{t=1}^{T} \sum_{d_1 \sim d_2} \left| \mathrm{R}_t^{(d_1)} - \mathrm{R}_t^{(d_2)} \right|$$

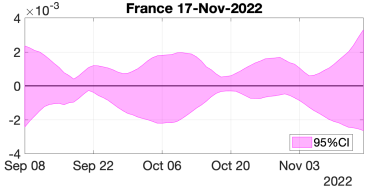
sum over neighboring counties

here: $d_1 \sim d_2 \Leftrightarrow$ share terrestrial border

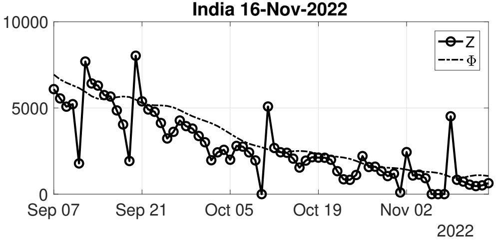


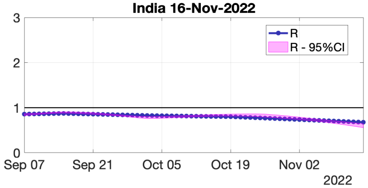

France 17-Nov-2022
×10^4
10
5
0
Z
Φ
Sep 08
Sep 22
Oct 06
Oct 20
Nov 03
2022

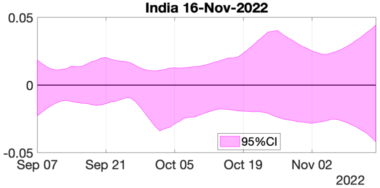

France 17-Nov-2022
3
2
1
0
R
R - 95%CI
Sep 08
Sep 22
Oct 06
Oct 20
Nov 03
2022


France 17-Nov-2022
×10^-3
4
2
0
-2
-4
95%CI
Sep 08
Sep 22
Oct 06
Oct 20
Nov 03
2022

Worldwide Covid19 monitoring

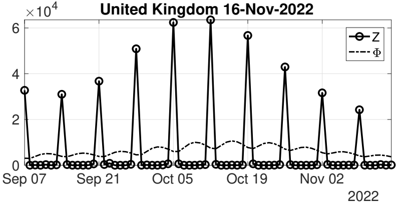






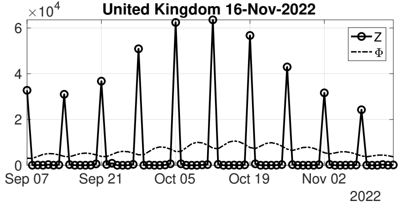
Why not United Kingdom?

Why not United Kingdom?



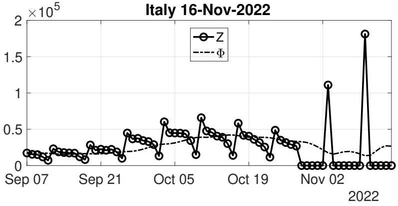
rate of erroneous counts: $6/7$!

Why not United Kingdom?



rate of erroneous counts: 6/7!

And Italy?



seems to adopt the same reporting rate ...

$\Longrightarrow$ call for new tools, robust to very scarce data

Bayesian framework for credibility interval estimation

Pointwise estimate of parameter $\boldsymbol{\theta} = \mathbf{R}$ from observations $\mathbf{Z}$

$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \quad f(\boldsymbol{\theta}|\mathbf{Z}) + h(\mathbf{A}\boldsymbol{\theta})$$

(Pascal et al., 2022, *Trans. Sig. Process.*)

Pointwise estimate of parameter $\boldsymbol{\theta} = \mathbf{R}$ from observations $\mathbf{Z}$

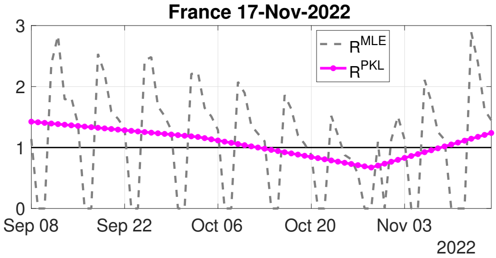
$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \quad f(\boldsymbol{\theta}|\mathbf{Z}) + h(\mathbf{A}\boldsymbol{\theta}) \quad \text{(Pascal et al., 2022, \textit{Trans. Sig. Process.})}$$

Q: what is the value of R today? **R:** solve the minimization problem and output $\widehat{\mathrm{R}}_T$.

Pointwise estimate of parameter $\boldsymbol{\theta} = \mathbf{R}$ from observations $\mathbf{Z}$

$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \quad f(\boldsymbol{\theta}|\mathbf{Z}) + h(\mathbf{A}\boldsymbol{\theta}) \quad \text{(Pascal et al., 2022, \textit{Trans. Sig. Process.})}$$

Q: what is the value of R today? **R:** solve the minimization problem and output $\widehat{\mathrm{R}}_T$.



$\widehat{\mathrm{R}}_T = 1.2955$

Bayesian framework for credibility interval estimation

Pointwise estimate of parameter $\boldsymbol{\theta} = \mathbf{R}$ from observations $\mathbf{Z}$

$$\underset{\mathbf{R}\in\mathbb{R}^T}{\text{minimize}} \quad f(\boldsymbol{\theta}|\mathbf{Z}) + h(\mathbf{A}\boldsymbol{\theta})$$

(Pascal et al., 2022, *Trans. Sig. Process.*)

Bayesian reformulation: interpret $\widehat{\mathbf{R}}^{\mathrm{PL}}$ as the Maximum A Posteriori of

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta}|\mathbf{Z}) - h(\mathbf{A}\boldsymbol{\theta}))$$

- $\exp(-f(\boldsymbol{\theta}|\mathbf{Z})) \sim$ likelihood of the observation
- $\exp(-h(\mathbf{A}\boldsymbol{\theta})) \sim$ prior on the parameter of interest

Bayesian framework for credibility interval estimation

Pointwise estimate of parameter $\boldsymbol{\theta} = \mathbf{R}$ from observations $\mathbf{Z}$

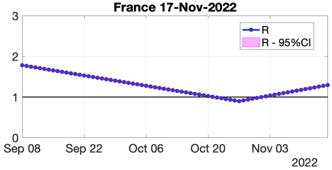
$$\underset{\mathbf{R} \in \mathbb{R}^T}{\text{minimize}} \quad f(\boldsymbol{\theta}|\mathbf{Z}) + h(\mathbf{A}\boldsymbol{\theta}) \quad \text{(Pascal et al., 2022, } \textit{Trans. Sig. Process.}\text{)}$$

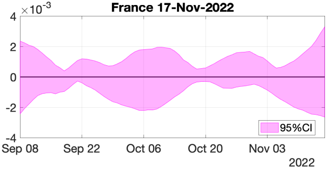
Bayesian reformulation: interpret $\widehat{\mathbf{R}}^{\mathrm{PL}}$ as the Maximum A Posteriori of

$$\pi(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta}|\mathbf{Z}) - h(\mathbf{A}\boldsymbol{\theta}))$$

- $\exp(-f(\boldsymbol{\theta}|\mathbf{Z})) \sim$ likelihood of the observation
- $\exp(-h(\mathbf{A}\boldsymbol{\theta})) \sim$ prior on the parameter of interest

$\Longrightarrow$ instead of focusing on $\widehat{\mathrm{R}}_t$, the **pointwise** MAP, probe π to get $\mathrm{R}_t \in [\underline{\mathrm{R}}_t, \overline{\mathrm{R}}_t]$ with 95% probability, i.e., **credibility interval** estimates





$$\widehat{\mathrm{R}}_T \in [1.2987, 1.3047]$$

Purpose: sampling the random variable $\boldsymbol{\theta} = \mathbf{R} \in \mathbb{R}^T$ according to the posterior[†]

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

[†] π is defined up to a normalizing constant

Purpose: sampling the random variable $\boldsymbol{\theta} = \mathbf{R} \in \mathbb{R}^T$ according to the posterior[†]

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

Principle: 1) generate a random sequence $\{\boldsymbol{\theta}^n,\, n \in \mathbb{N}\}$ such that

- $\boldsymbol{\theta}^{n+1}$ only depends on $\boldsymbol{\theta}^n$,
- at convergence, i.e., as $n \to \infty$, $\boldsymbol{\theta}^n \sim \pi$,

2) compute Bayesian estimators, e.g., credibility intervals, on samples $\{\boldsymbol{\theta}^n,\, n \geq N\}$

[†] π is defined up to a normalizing constant

Purpose: sampling the random variable $\boldsymbol{\theta} = \mathbf{R} \in \mathbb{R}^T$ according to the posterior[†]

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

Principle: 1) generate a random sequence $\{\boldsymbol{\theta}^n, n \in \mathbb{N}\}$ such that

- $\boldsymbol{\theta}^{n+1}$ only depends on $\boldsymbol{\theta}^n$,
- at convergence, i.e., as $n \to \infty$, $\boldsymbol{\theta}^n \sim \pi$,

2) compute Bayesian estimators, e.g., credibility intervals, on samples $\{\boldsymbol{\theta}^n, n \geq N\}$

State-of-the-art: *Hastings-Metropolis random walk*

(i) propose a random move according to

$$\boldsymbol{\theta}^{n+\frac{1}{2}} = \boldsymbol{\theta}^n + \sqrt{2\gamma}\Gamma\xi^{n+1}, \quad \xi^{n+1} \sim \mathcal{N}_T(0, \mathrm{I})$$

with γ positive step size, $\Gamma \in \mathbb{R}^{T \times T}$

[†] π is defined up to a normalizing constant

Purpose: sampling the random variable $\boldsymbol{\theta} = \mathbf{R} \in \mathbb{R}^T$ according to the posterior[†]

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

Principle: 1) generate a random sequence $\{\boldsymbol{\theta}^n, n \in \mathbb{N}\}$ such that

- $\boldsymbol{\theta}^{n+1}$ only depends on $\boldsymbol{\theta}^n$,
- at convergence, i.e., as $n \to \infty$, $\boldsymbol{\theta}^n \sim \pi$,

2) compute Bayesian estimators, e.g., credibility intervals, on samples $\{\boldsymbol{\theta}^n, n \geq N\}$

State-of-the-art: *Hastings-Metropolis random walk*

(i) propose a random move according to

$$\boldsymbol{\theta}^{n+\frac{1}{2}} = \boldsymbol{\theta}^n + \sqrt{2\gamma}\Gamma\xi^{n+1}, \quad \xi^{n+1} \sim \mathcal{N}_T(0, \mathrm{I})$$

with γ positive step size, $\Gamma \in \mathbb{R}^{T \times T}$

(ii) accept: $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+\frac{1}{2}}$, with probability $1 \wedge \dfrac{\pi(\boldsymbol{\theta}^{n+\frac{1}{2}})}{\pi(\boldsymbol{\theta}^n)}$, or reject: $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$

† π is defined up to a normalizing constant

Metropolis Adjusted Langevin Algorithm (MALA)

Langevin dynamics: $\boldsymbol{\theta}^{n+\frac{1}{2}} = \mu(\boldsymbol{\theta}^n) + \sqrt{2\gamma}\xi^{n+1}$, (Kent, 1978, *Adv Appl Probab*)

$$\mu(\boldsymbol{\theta}) \text{ adapted to } \pi(\boldsymbol{\theta}) = \exp(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta}))\mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

Langevin dynamics: $\boldsymbol{\theta}^{n+\frac{1}{2}} = \mu(\boldsymbol{\theta}^n) + \sqrt{2\gamma}\xi^{n+1}$, (Kent, 1978, *Adv Appl Probab*)

$$\mu(\boldsymbol{\theta}) \text{ adapted to } \pi(\boldsymbol{\theta}) = \exp(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta}))\mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

Case 1: $g = 0$ and $-\ln \pi = f$ is smooth (Roberts & Tweedie, 1996, *Bernoulli*)

$$\mu(\boldsymbol{\theta}) = \boldsymbol{\theta} - \gamma\Gamma\Gamma^\top \nabla f(\boldsymbol{\theta}) = \boldsymbol{\theta} + \gamma\Gamma\Gamma^\top \nabla \ln \pi(\boldsymbol{\theta})$$

$\Longrightarrow$ move towards areas of higher probability

Metropolis Adjusted Langevin Algorithm (MALA)

Langevin dynamics: $\boldsymbol{\theta}^{n+\frac{1}{2}} = \mu(\boldsymbol{\theta}^n) + \sqrt{2\gamma}\xi^{n+1}$, (Kent, 1978, *Adv Appl Probab*)

$$\mu(\boldsymbol{\theta}) \text{ adapted to } \pi(\boldsymbol{\theta}) = \exp(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta}))\mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$$

Case 1: $g = 0$ and $-\ln \pi = f$ is smooth (Roberts & Tweedie, 1996, *Bernoulli*)

$$\mu(\boldsymbol{\theta}) = \boldsymbol{\theta} - \gamma\Gamma\Gamma^\top \nabla f(\boldsymbol{\theta}) = \boldsymbol{\theta} + \gamma\Gamma\Gamma^\top \nabla \ln \pi(\boldsymbol{\theta})$$

$\Longrightarrow$ move towards areas of higher probability

Case 2: $-\ln \pi = f + g$ is nonsmooth

$$\mu(\boldsymbol{\theta}) = \operatorname{prox}_{\gamma g}^{\Gamma\Gamma^\top}(\boldsymbol{\theta} - \gamma\Gamma\Gamma^\top \nabla f(\boldsymbol{\theta}))$$

combining *Langevin* and *proximal*† approaches

† $\operatorname{prox}_{\gamma g}^{\Gamma\Gamma^\top}(y) = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left(\frac{1}{2}\|x-y\|^2_{\Gamma\Gamma^\top} + \gamma g(x)\right)$: preconditioned proximity operator of g

Posterior density of $\boldsymbol{\theta} = \mathbf{R}$: $\pi(\boldsymbol{\theta}) \propto \exp\left(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$

- smooth negative log-likelihood

$$\text{if } \boldsymbol{\theta} \in \mathcal{D}, \quad f(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \left(\mathrm{Z}_t \ln \mathrm{p}_t(\boldsymbol{\theta}) - \mathrm{p}_t(\boldsymbol{\theta})\right), \quad \mathrm{p}_t(\boldsymbol{\theta}) = \mathrm{R}_t(\Phi \mathrm{Z})_t$$

- nonsmooth convex lower-semicontinuous negative a priori log-distribution

$$g(\boldsymbol{\theta}) = \lambda_{\mathrm{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 = h(\mathbf{A}\boldsymbol{\theta})$$

$\mathbf{A} : \boldsymbol{\theta} \mapsto \mathbf{D}_2\mathbf{R}$ linear operator, $h(\cdot) = \lambda_{\mathrm{R}} \|\cdot\|_1$

Proximal-Gradient dual sampler PGdual

Posterior density of $\boldsymbol{\theta} = \mathbf{R}$: $\pi(\boldsymbol{\theta}) \propto \exp\left(-f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{\theta})$

- smooth negative log-likelihood

$$\text{if } \boldsymbol{\theta} \in \mathcal{D}, \quad f(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \left(\mathsf{Z}_t \ln \mathsf{p}_t(\boldsymbol{\theta}) - \mathsf{p}_t(\boldsymbol{\theta})\right), \quad \mathsf{p}_t(\boldsymbol{\theta}) = \mathsf{R}_t(\Phi \mathsf{Z})_t$$

- nonsmooth convex lower-semicontinuous negative a priori log-distribution

$$g(\boldsymbol{\theta}) = \lambda_{\mathsf{R}} \|\mathbf{D}_2 \mathbf{R}\|_1 = h(\mathbf{A}\boldsymbol{\theta})$$

$\mathbf{A} : \boldsymbol{\theta} \mapsto \mathbf{D}_2\mathbf{R}$ linear operator, $h(\cdot) = \lambda_{\mathsf{R}} \|\cdot\|_1$

Case 3: $-\ln \pi = f + h(\mathbf{A}\cdot)$ (Fort et al., 2022, *preprint*)

closed-form expression of $\text{prox}_{\gamma h}$ but not of $\text{prox}_{\gamma h(\mathbf{A}\cdot)}$

1) extend $\mathbf{A}$ into **invertible** $\overline{\mathbf{A}}$, and h in $\bar{h}$ such that $\bar{h}(\overline{\mathbf{A}}\boldsymbol{\theta}) = h(\mathbf{A}\boldsymbol{\theta})$

2) reason on the **dual** variable $\tilde{\boldsymbol{\theta}} = \overline{\mathbf{A}}\boldsymbol{\theta}$

Data: $\overline{\mathbf{D}} = \overline{\mathbf{D}}_2$ (`Invert`) or $\overline{\mathbf{D}} = \overline{\mathbf{D}}_o$ (`Ortho`)

$\gamma_{\mathrm{R}}, \gamma_{\mathrm{O}} > 0$, $N_{\max} \in \mathbb{N}_\star$, $\boldsymbol{\theta}^0 = (\mathbf{R}^0, \mathbf{O}^0) \in \mathcal{D}$

Result: A $\mathcal{D}$-valued sequence $\{\boldsymbol{\theta}^n = (\mathbf{R}^n, \mathbf{O}^n), n \in 0, \ldots, N_{\max}\}$

for $n = 0, \ldots, N_{\max} - 1$ **do**

Sample $\xi_{\mathrm{R}}^{n+1} \sim \mathcal{N}_T(0, \mathrm{I})$ and $\xi_{\mathrm{O}}^{n+1} \sim \mathcal{N}_T(0, \mathrm{I})$;

Set $\mathbf{R}^{n+\frac{1}{2}} = \mu_{\mathrm{R}}(\boldsymbol{\theta}^n) + \sqrt{2\gamma_{\mathrm{R}}}\,\overline{\mathbf{D}}^{-1}\overline{\mathbf{D}}^{-\top}\,\xi_{\mathrm{R}}^{n+1}$;

$\mathbf{O}^{n+\frac{1}{2}} = \mu_{\mathrm{O}}(\boldsymbol{\theta}^n) + \sqrt{2\gamma_{\mathrm{O}}}\,\xi_{\mathrm{O}}^{n+1}$;

$\boldsymbol{\theta}^{n+\frac{1}{2}} = (\mathbf{R}^{n+\frac{1}{2}}, \mathbf{O}^{n+\frac{1}{2}})$;

Set $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^{n+\frac{1}{2}}$ with probability

$$1 \wedge \frac{\pi(\boldsymbol{\theta}^{n+\frac{1}{2}})}{\pi(\boldsymbol{\theta}^n)} \frac{q_{\mathrm{R}}(\boldsymbol{\theta}^{n+\frac{1}{2}}, \boldsymbol{\theta}_{\mathrm{R}}^n)}{q_{\mathrm{R}}(\boldsymbol{\theta}^n, \boldsymbol{\theta}_{\mathrm{R}}^{n+\frac{1}{2}})} \frac{q_{\mathrm{O}}(\boldsymbol{\theta}^{n+\frac{1}{2}}, \boldsymbol{\theta}_{\mathrm{O}}^n)}{q_{\mathrm{O}}(\boldsymbol{\theta}^n, \boldsymbol{\theta}_{\mathrm{O}}^{n+\frac{1}{2}})},$$

$q_{\mathrm{R/O}}$: Gaussian kernel stemming from nonsymmetric proposal

and $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$ otherwise.

Algorithm 1: Proximal-Gradient dual: `PGdual Invert` and `PGdual Ortho`