INL/CON-24-78570-Revision-0

Regularized Differentiation for Bioburden Density Estimation in Planetary Protection

June 2024

Changing the World's Energy Future

Andrei Vasilyevich Gribok, Michael DiNicola, Lisa Guan

Idaho National Laboratory

INL/CON-24-78570-Revision-0

Regularized Differentiation for Bioburden Density Estimation in Planetary Protection

Andrei Vasilyevich Gribok, Michael DiNicola, Lisa Guan

June 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Jet Propulsion Laboratory
California Institute of Technology

Regularized Differentiation for Bioburden Density Estimation in Planetary Protection

Andrei Gribok[a], Mike DiNicola[b] ,and Lisa Guan[b]

[a] Idaho National Laboratory, Idaho Falls, ID, USA
[b] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

ESREL 2024

23-27 JUNE, 2024

the 34-th European Safety and Reliability Conference, Jagiellonian University, Cracow/Kraków, Poland

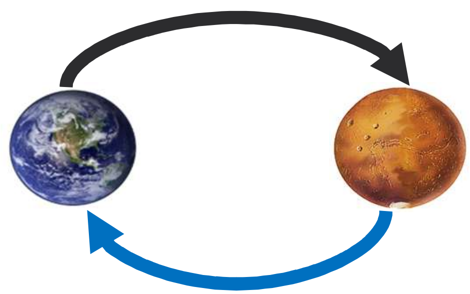
INL Idaho National Laboratory

What is Planetary Protection?

Jet Propulsion Laboratory
California Institute of Technology

Planetary protection (PP) aims to:

- Protect other worlds from microorganisms found on Earth.
- Prevent both forward and backward contamination.



Why?

- How can we look for life on other planets if we bring our own?
- Prevent non-terrestrial biocontamination of Earth.
- Meet NASA planetary protection requirements for launch approval.
- Comply with international treaty obligations.

PP requirements per NASA Procedural Requirements 8020.12D:

- For category IV missions to Mars:
  - $\leq 5\times10^5$ total spores (colony forming unit [CFU]) at launch
  - $\leq 3\times10^5$ total spores on the planned landing hardware
  - ≤ 300 spores/m^2 on exposed surfaces.
- For the outer planets:
  - Inadvertent contamination of an ocean or other liquid water body is kept at a probability of less than 1×10^{-4} per mission.
- Monitoring and/or verification of biological cleanliness is conducted using direct wipe and swab samples extracted and processed via a cultivation-based technique (NASA standard assay) that measures heat-tolerant organisms.
- The total number of spores/CFUs at launch is a parameter in PP probabilistic models.

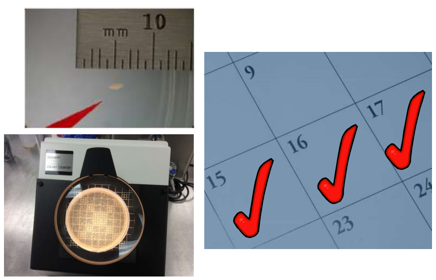
Jet Propulsion Laboratory
California Institute of Technology

Planetary Protection Biological Cleanliness Verification Sampling

Clean hardware (and table or bag) with a solvent wipe prior to sampling or installation

Sample with swab or wipe (water is used as the solvent)

Process swabs and wipes, up to 24 hours required post-assay



Count plates – 24 h, 48 h, and 72 h

- Swabs sampled a 0.0025 m^2 surface area maximum.
- Wipes sampled up to a 1.0 m^2 surface area.
- Due to the experimental procedure, the swabs assume a pour fraction of 0.8 and the wipes 0.25, representing the portion of the total sample solution plated and analyzed for CFU counts.
- Exposure is the sampled area multiplied by the pour fraction (i.e., for swabs it would be E= 0.0025 m^2 $\cdot$0.8=0.002 m^2).

Spore/CFU, a heat-tolerant reproductive cell capable of developing into a new individual without fusion with another reproductive cell.

Data Generating Model and Parameter Estimators

$X_i = x | \lambda_{true}^{i} \sim Poisson(\lambda_{true}^{i} \cdot E_i)$

$\lambda_{true}^{i} | \alpha, \beta \sim Gamma(\alpha, \beta), i = 1, \dots N$

$\widehat{\lambda_i}(x_i) = \frac{x_i}{E_i}$ – Maximum Likelihood Estimator (MLE)

$\widehat{\lambda_i}(x_i) = \frac{x_i+\alpha}{E_i+\beta}$ – Conjugate prior Bayes Estimator

$\widehat{\lambda_i} = d$ – Deterministic Estimator

X_i – random variable describing the number of CFUs on i-th sampled component
x_i – observed number of CFUs in a sample from i-th sampled component
E_i – exposure for a sample, determined by multiplying the sampled area by pour fraction, m^2
λ_{true}^{i} – true bioburden density for a sample, CFUs/m^2
$Gamma(\lambda / \alpha,\beta)$ – Gamma prior distribution of bioburden density λ_{true}^{i}
α,β – parameters of the Gamma prior distribution
$Poisson(\lambda_{true}^{i} \cdot E_i)$ – Poisson likelihood of having x CFU counts at exposure E_i given λ_{true}^{i}
N – number of components

Component 261 Sample No.	CFUs Observed	Area Sampled, m^2	Pour Fraction	Exposure, m^2
1	1	0.0025	0.8	0.0020
2	0	0.0025	0.8	0.0020
3	0	0.0025	0.8	0.0020
4	0	0.0025	0.8	0.0020
5	0	0.0025	0.8	0.0020
6	0	0.0025	0.8	0.0020
7	0	0.0025	0.8	0.0020
8	0	0.0025	0.8	0.0020
9	0	0.0025	0.8	0.0020
10	0	0.0025	0.8	0.0020
11	0	0.0025	0.8	0.0020
12	0	0.0025	0.8	0.0020
13	0	0.0025	0.8	0.0020
14	0	0.0025	0.8	0.0020
15	0	0.0025	0.8	0.0020
16	0	0.0025	0.8	0.0020
17	0	0.0025	0.8	0.0020
18	0	0.0025	0.8	0.0020
19	0	0.0025	0.8	0.0020
20	6	0.0025	0.8	0.0020
21	8	0.0025	0.8	0.0020
22	10	0.0025	0.8	0.0020
23	10	0.0025	0.8	0.0020
24	14	0.0025	0.8	0.0020
Total	52	0.0600	0.8	0.0480

Integration vs Differentiation

$$f'(x) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

$$F(x) = \int_0^x F'(t)dt, \qquad F(0) = 0$$

$$\frac{dF(x)}{dx} = \frac{d}{dx} \int_0^x F'(t)dt = F'(x)$$

$$\sin(x) = \int_0^x \cos(t)dt$$

$$\frac{d(\sin(x))}{dx} = \frac{d}{dx} \int_0^x \cos(t)dt = \cos(x)$$

The Problem of Differentiation

$$g(x) = f(x) + a \cdot \sin(\omega \cdot x)\,;\, g'(x) = f'(x) + \omega \cdot a \cdot \cos(\omega \cdot x)$$

$$g'(x) - f'(x) = \omega \cdot a \cdot \cos(\omega \cdot x)$$

$$G(x) = \int f(x)dx - \frac{a}{w} \cdot \cos(w \cdot x) + C$$

$$G(x) - \int f(x)dx = C - \frac{a}{w} \cdot \cos(w \cdot x)$$

$$F(x) + \varepsilon = \int_0^x F'(t)dt, \qquad F(0) = 0$$

$$\frac{d[F(x)+\varepsilon]}{dx} = \frac{d}{dx}\int_0^x F'(t)dt = F'(x) + \frac{d\varepsilon}{dx}$$

Jet Propulsion Laboratory
California Institute of Technology

Naïve Differentiation of the Sampled Data

$$t_1, t_2, t_3 \dots, t_{n-1}, t_n; \Delta t = t_n - t_{n-1}$$

$$y(t_1), y(t_2), y(t_3) \dots, y(t_{n-1}), y(t_n)$$

$$y'(t_1) = \frac{y(t_2) - y(t_1)}{t_2 - t_1}, y'(t_2) = \frac{y(t_3) - y(t_2)}{t_3 - t_2}, \dots, y'(t_{n-1}) = \frac{y(t_n) - y(t_{n-1})}{t_n - t_{n-1}}$$

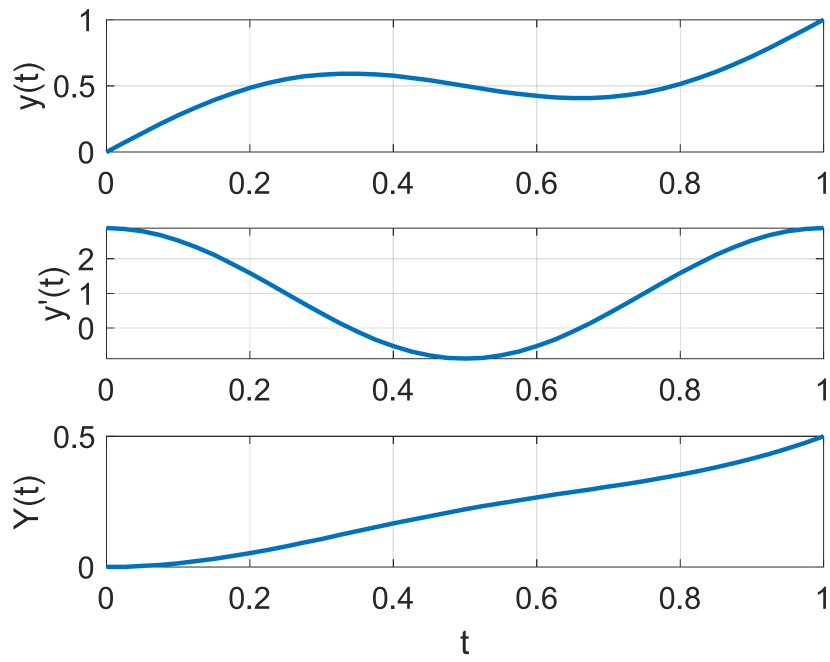
Naïve Differentiation of the Sampled Data

Jet Propulsion Laboratory
California Institute of Technology

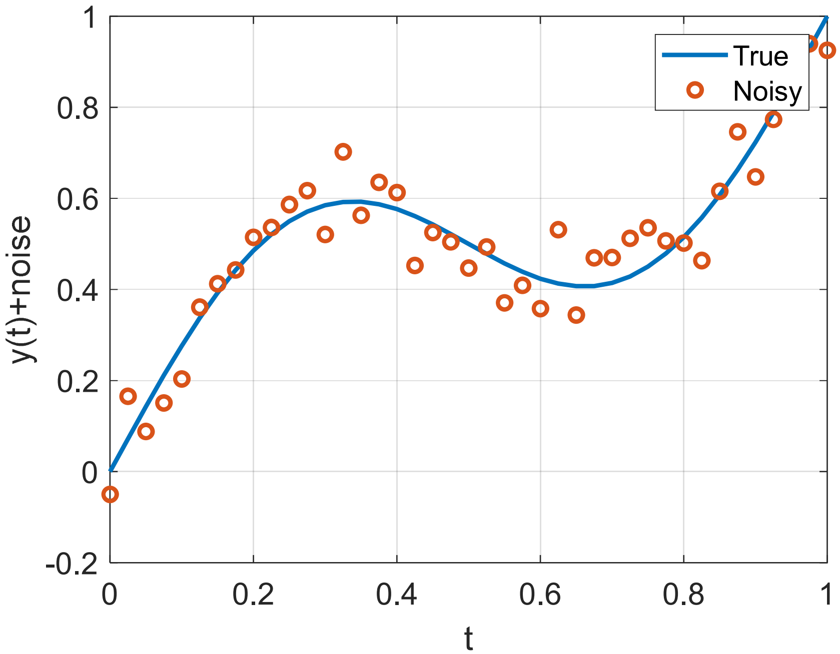
$$y(t) = t + 0.3 \cdot \sin(2\pi \cdot t)$$

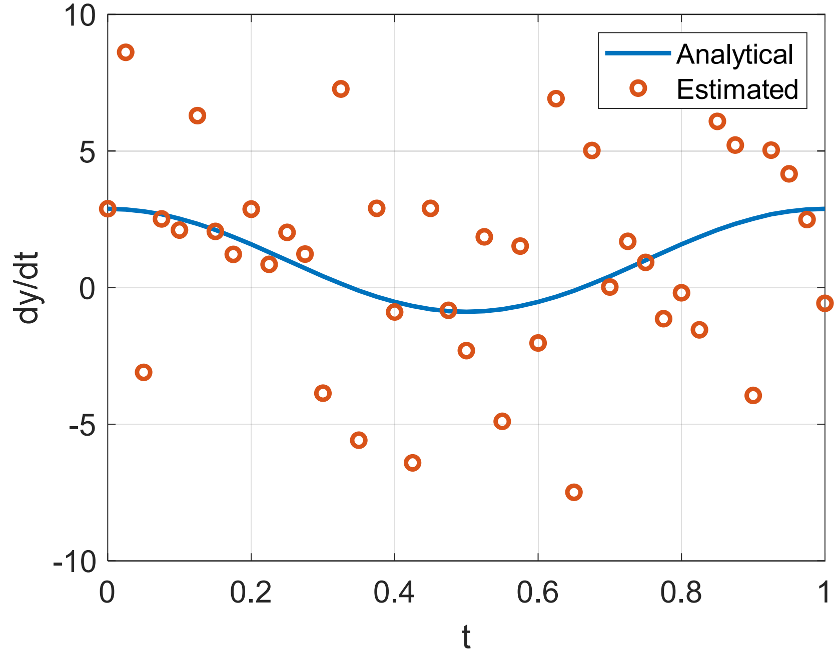
$$y'(t) = 1 + 0.3 \cdot 2\pi \cdot \cos(2\pi \cdot t)$$

$$\int_0^x t + 0.3 \cdot \sin(2\pi \cdot t) = \frac{x^2}{2} - \frac{0.3}{2\pi} \cdot \cos(2\pi \cdot x) + \frac{0.3}{2\pi}$$



Naïve Differentiation of the Sampled Data





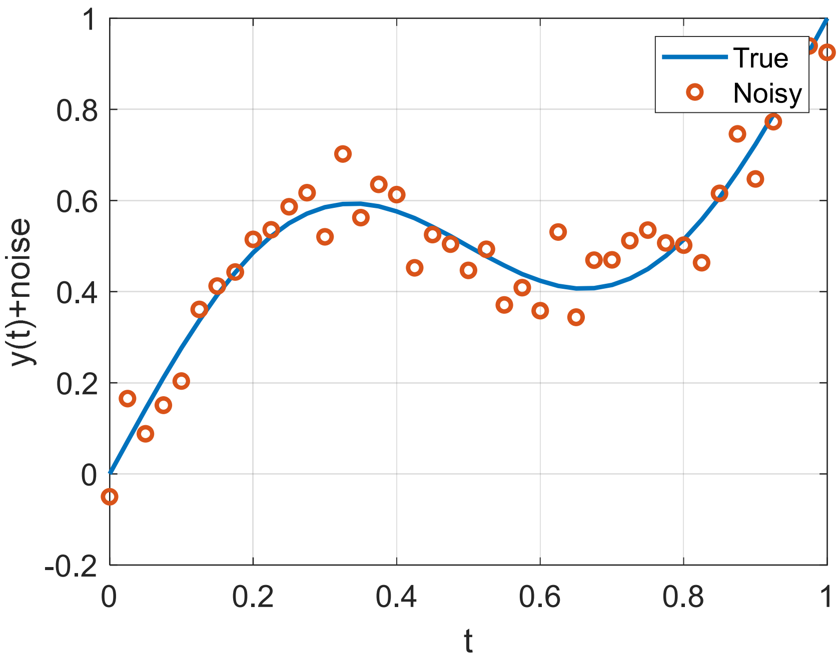
$$F(x) = \int_0^x F'(t)dt, \qquad F(0) = 0$$

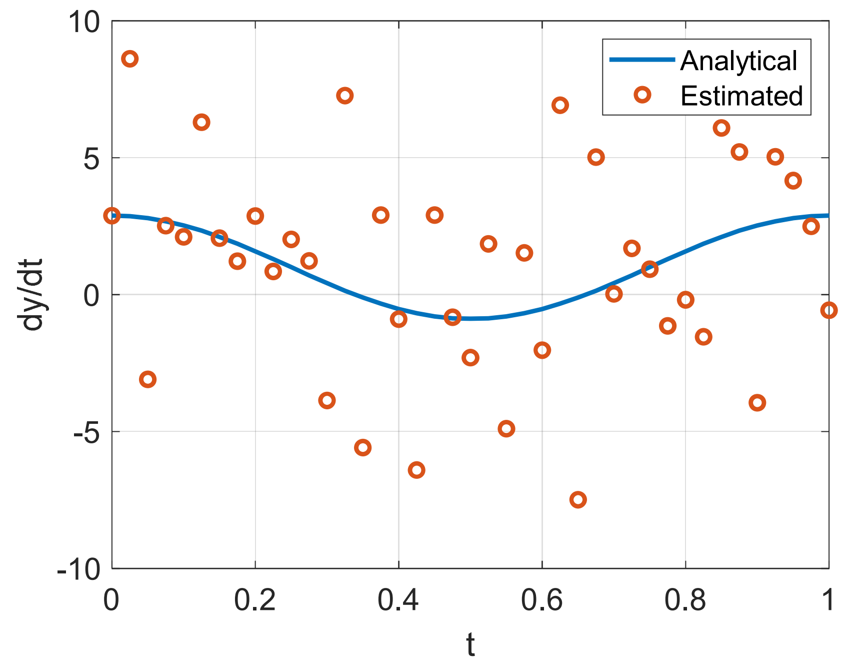
$$\frac{d[F(x) + \varepsilon]}{dx} = \frac{d}{dx}\int_0^x F'(t)dt = F'(x) + \frac{d\varepsilon}{dx}$$

$$F(x) + \varepsilon = \int_0^x F'(t)dt, \qquad F(0) = 0$$

Naïve Differentiation of the Sampled Data

$$y(t) = t + 0.3 \cdot \sin(2\pi \cdot t) + \varepsilon;\ \bar{\varepsilon} = 0, \sigma_\varepsilon = 0.05; \sigma_{Residuals} = 3.6; \frac{\sigma_{Residuals}}{\sigma_\varepsilon} = 72$$

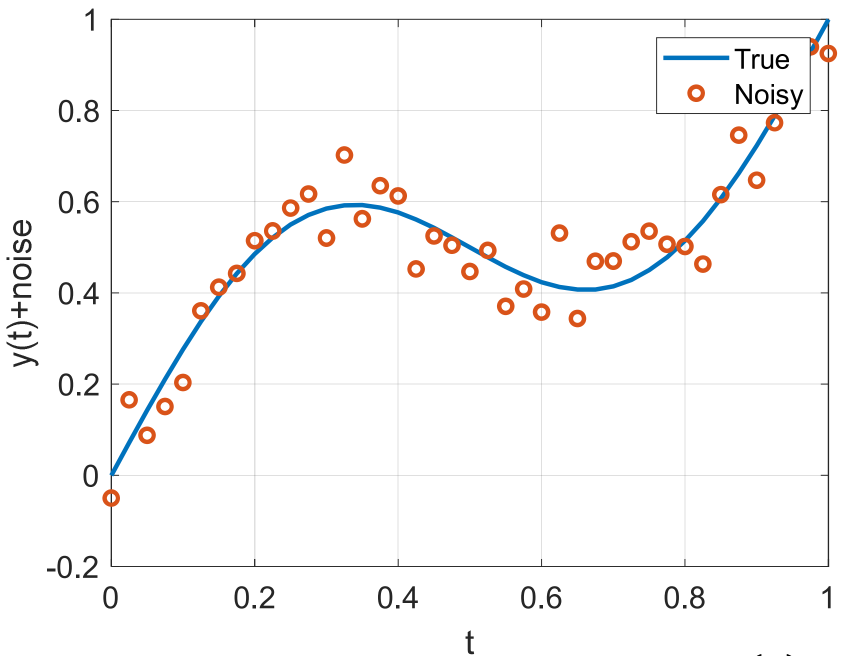


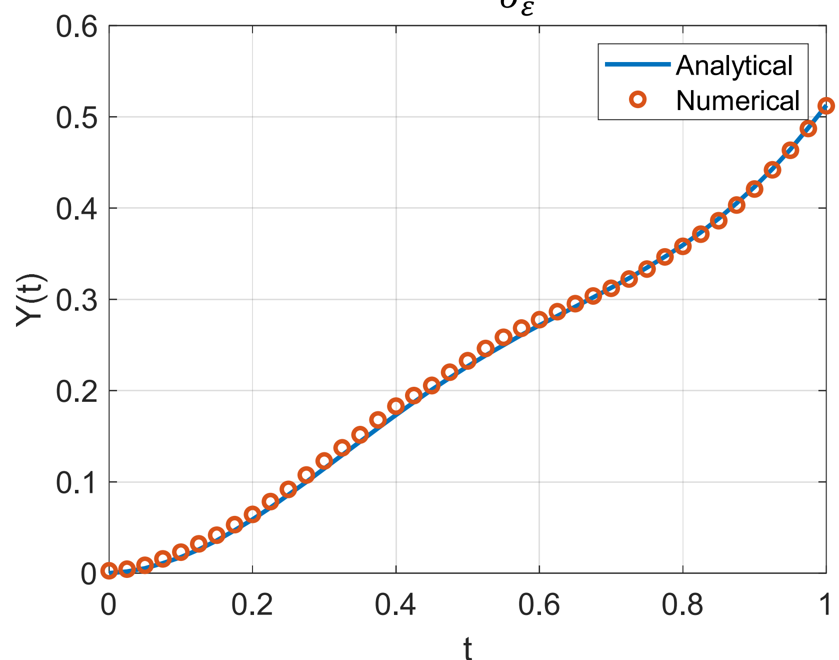


$$y'(t_1) = \frac{y(t_2) - y(t_1)}{t_2 - t_1}, y'(t_2) = \frac{y(t_3) - y(t_2)}{t_3 - t_2}, \ldots, y'(t_{n-1}) = \frac{y(t_n) - y(t_{n-1})}{t_n - t_{n-1}}$$

Integration of the Noisy Data

$$y(t) = t + 0.3 \cdot \sin(2\pi \cdot t) + \varepsilon;\ \bar{\varepsilon} = 0, \sigma_\varepsilon = 0.05; \sigma_{Residuals} = 0.004; \frac{\sigma_{Residuals}}{\sigma_\varepsilon} = 0.08$$





$$S(n) = \sum_{i=1}^{n} y(t_i) \cdot \Delta t, \Delta t = t_i - t_{i-1}$$

$$S(1) = y(t_1) \cdot \Delta t; S(2) = [y(t_1) + y(t_2)] \cdot \Delta t; S(n) = [y(t_1) + y(t_2) + \cdots + y(t_n)] \cdot \Delta t$$

Jet Propulsion Laboratory
California Institute of Technology

Integration and Differentiation in Matrix Form

$$t_1, t_2, t_3 \ldots, t_{n-1}, t_n; \Delta t = t_n - t_{n-1}$$

$$y(t_1), y(t_2), y(t_3) \ldots, y(t_{n-1}), y(t_n)$$

$$m = 1,2, \ldots n \qquad S(m) = \sum_{i=1}^{m} y(t_i) \cdot \Delta t$$

$$\underbrace{\begin{bmatrix} S(1) \\ S(2) \\ \vdots \\ S(n) \end{bmatrix}}_{nx1} = \underbrace{\begin{bmatrix} \Delta t & 0 & \ldots & 0 \\ \Delta t & \Delta t & \ldots & 0 \\ \vdots & \vdots & \vdots & 0 \\ \Delta t & \Delta t & \Delta t & \Delta t \end{bmatrix}}_{A=nxn} \cdot \underbrace{\begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_n) \end{bmatrix}}_{nx1}$$

$$[A^T \cdot A]^{-1} \cdot A^T = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots & \vdots \\ 0 & \vdots & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \cdot \frac{1}{\Delta t}$$

$$F(x) = \int_0^x F'(t)dt$$

$$\frac{dF(x)}{dx} = \frac{d}{dx}\int_0^x F'(t)dt = F'(x)$$

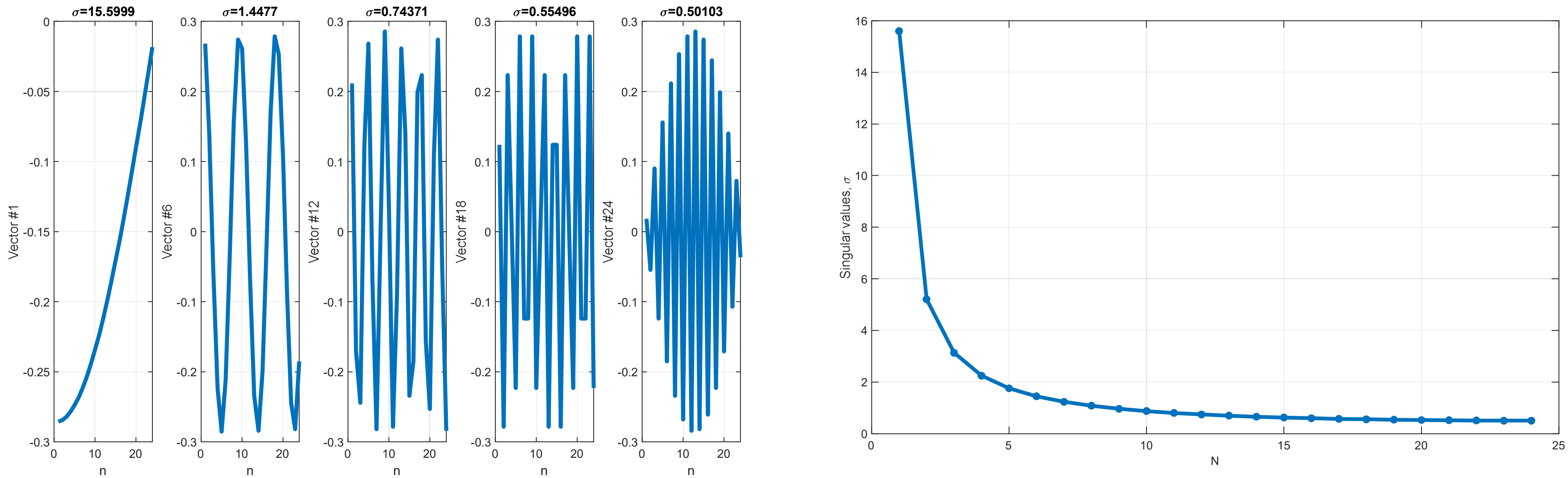
$$\{\|S - A \cdot y\|^2\} \rightarrow min$$

$$\hat{y} = [A^T \cdot A]^{-1} \cdot A^T \cdot S$$

$$\|S - A \cdot y\|^2 + \mu \cdot \|L \cdot y\|^2 \rightarrow min$$

Singular Vectors and Singular Values of the Differentiation Matrix, N=24

Jet Propulsion Laboratory
California Institute of Technology



$$\hat{y}_{OLS} = V \cdot \Sigma^{-1} \cdot U^T \cdot S = \sum_{i=1}^{N} \frac{u_i^T \cdot S}{\sigma_i} \cdot v_i$$

Frequency Response of Integration and Differentiation

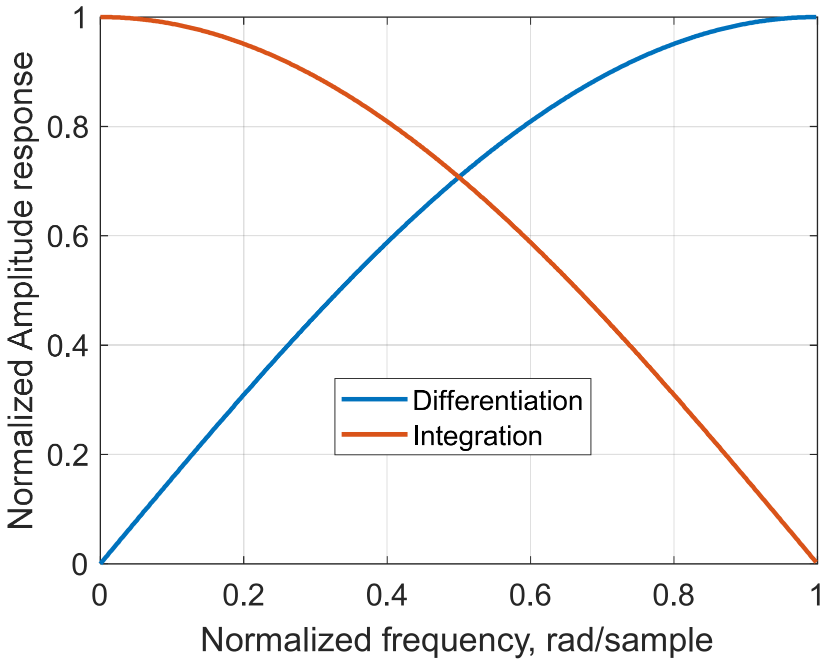
$$y'(t_1) = \frac{y(t_2) - y(t_1)}{t_2 - t_1} = [y(t_2) - y(t_1)] \cdot \frac{1}{\Delta t}; \Delta t = t_2 - t_1$$

$$h_{diff} = [1 \ \ -1], \Delta t = 1$$

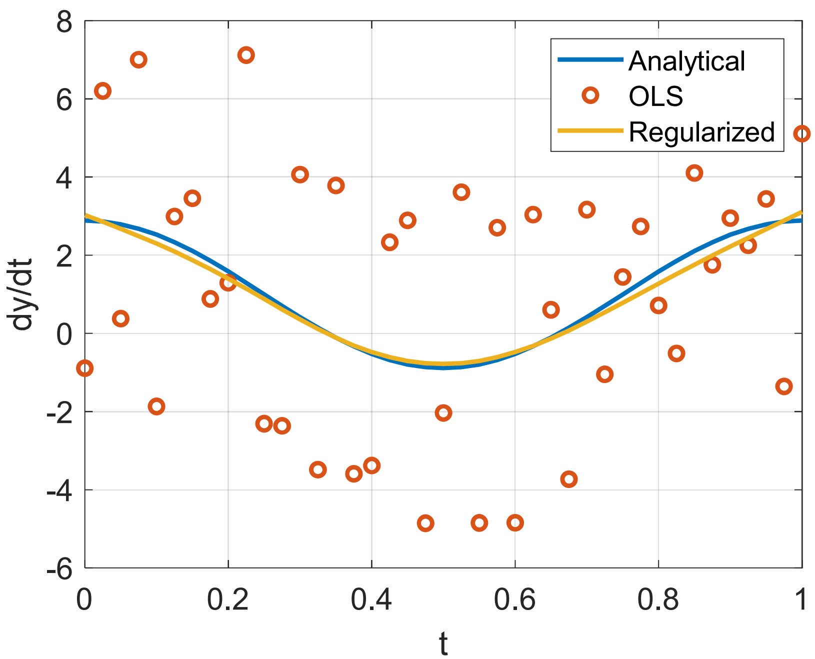
$$S = [y(t_1) + y(t_2)] \cdot \Delta t$$

$$h_{int} = [1 \ \ 1], \Delta t = 1$$

Amplitude response is the absolute value of the Fourier transform of the filter's coefficients (impulse response).



Regularized vs Naïve Derivatives



$$\underbrace{A}_{nxn} = \underbrace{U}_{nxn} \cdot \underbrace{\Sigma}_{nxn} \cdot \underbrace{V}_{nxn}{}^{T}$$

$$\|S - A \cdot y\|^2 \to min$$

$$\hat{y}_{OLS} = V \cdot \Sigma^{-1} \cdot U^T \cdot S = \sum_{i=1}^{N} \frac{u_i^T \cdot S}{\sigma_i} \cdot v_i$$

$$\|S - A \cdot y\|^2 + \mu \cdot \|L \cdot y\|^2 \to min$$

$$\hat{y}_{Reg} = [A^T \cdot A + \mu \cdot L^T \cdot L]^{-1} \cdot A^T \cdot S = V \cdot (\Sigma^2 + \mu \cdot L^T \cdot L)^{-1} \cdot \Sigma \cdot U^T \cdot S$$
$$= \sum_{i=1}^{N} f_i \frac{u_i^T \cdot S}{\sigma_i} \cdot v_i$$

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu}, i = 1, \dots, N$$

Jet Propulsion Laboratory
California Institute of Technology

Regularized Differentiation for Bioburden Density Estimation

$$\|S - A \cdot y\|^2 + \mu \cdot \|L \cdot y\|^2 \to min$$

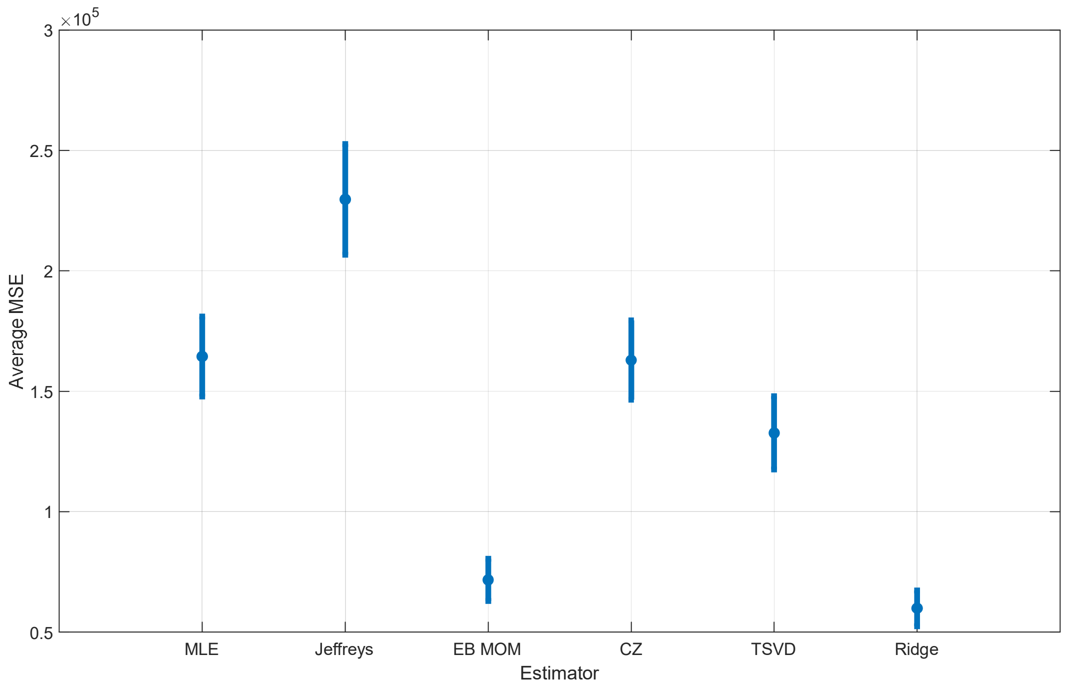
$$\hat{\lambda}_i^{MLE} = \frac{x_i}{E_i}, i = 1, \ldots \mathrm{n}$$

$$\lambda_1 = \frac{S(1) - S(0)}{E_1} = \frac{1}{0.0020}; \lambda_2 = \frac{S(2) - S(1)}{E_2} = \frac{0}{0.0020}$$

$$\lambda_{13} = \frac{S(13) - S(12)}{E_{13}} = \frac{3}{0.0020}; \ldots; \lambda_{24} = \frac{S(24) - S(23)}{E_2} = \frac{14}{0.0020}$$

Component 261 Sample No.	CFUs Observed	Area Sampled, m^2	Pour Fraction	Exposure, m^2	Cumulative CFU Count, S
0					0
1	1	0.0025	0.8	0.0020	1
2	0	0.0025	0.8	0.0020	1
3	0	0.0025	0.8	0.0020	1
4	0	0.0025	0.8	0.0020	1
5	0	0.0025	0.8	0.0020	1
6	0	0.0025	0.8	0.0020	1
7	0	0.0025	0.8	0.0020	1
8	0	0.0025	0.8	0.0020	1
9	0	0.0025	0.8	0.0020	1
10	0	0.0025	0.8	0.0020	1
11	0	0.0025	0.8	0.0020	1
12	0	0.0025	0.8	0.0020	1
13	0	0.0025	0.8	0.0020	4
14	0	0.0025	0.8	0.0020	4
15	0	0.0025	0.8	0.0020	4
16	0	0.0025	0.8	0.0020	4
17	0	0.0025	0.8	0.0020	4
18	0	0.0025	0.8	0.0020	4
19	0	0.0025	0.8	0.0020	4
20	6	0.0025	0.8	0.0020	10
21	8	0.0025	0.8	0.0020	18
22	10	0.0025	0.8	0.0020	28
23	10	0.0025	0.8	0.0020	38
24	14	0.0025	0.8	0.0020	52
Total	52	0.0600	0.8	0.0480	-

Simulated Data



$$\hat{\lambda}_i(x_i) = \frac{x_i}{E_i} \qquad \hat{\lambda}_i(x_i) = \frac{x_i+0.5}{E_i}$$

$$\hat{\lambda}_i(x_i) = \frac{x_i+\alpha_{MOM}}{E_i+\beta_{MOM}} \qquad \hat{\lambda}_{TSVD} = \sum_{i=1}^{k} \frac{u_i^T \cdot S}{\sigma_i} \cdot v_i$$

$$\hat{\lambda}_i^{CZ} = \frac{x_i}{E_i} - \frac{\gamma + N - 1}{\sum \frac{x_i}{E_i} + \gamma + N - 1} \cdot \left(\frac{x_i}{E_i}\right), i = 1, \ldots N$$

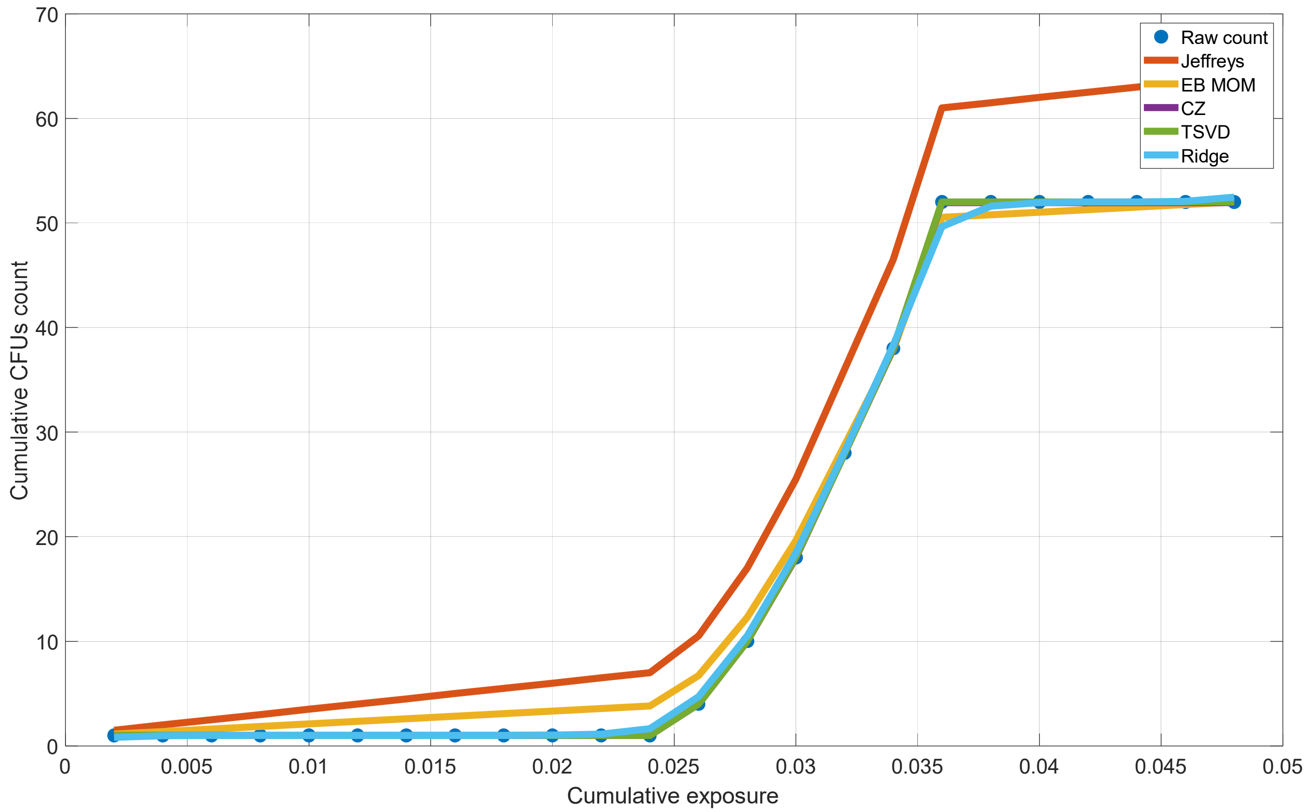
$$\hat{\lambda}_{Ridge} = \sum_{i=1}^{N} f_i \frac{u_i^T \cdot S}{\sigma_i} \cdot v_i$$

$$P(X = x|\lambda_{true}) = \frac{(\lambda_{true} \cdot E)^x}{x!} e^{-\lambda_{true} \cdot E}, x = 0,1,2, \ldots, \lambda_{true} \in (0, \infty), \mathrm{E} \in (0, \infty)$$

$$Gamma(\lambda_{true}|\alpha = 1.8, \beta = 0.001), \hat{\lambda} = 179.5$$

Jet Propulsion Laboratory
California Institute of Technology

Component 261



Conclusions

- MLE of bioburden density is a naïve derivative of CFUs cumulative count.
- There are two sources of noise: sampling noise and differentiation noise
- Regularized differentiation prevents differentiation noise amplification
- Regularized differentiation does not make any assumptions about sampling noise
- The variance of the sampling noise is a very important variable for selection of the regularization parameter and can be obtained from sampling efficiency studies
- The regularized differentiation approach integrates estimation and noise filtering into a single framework which may produce a better bioburden density estimate
- Regularized least squares does not work if ALL CFU data are zeros.

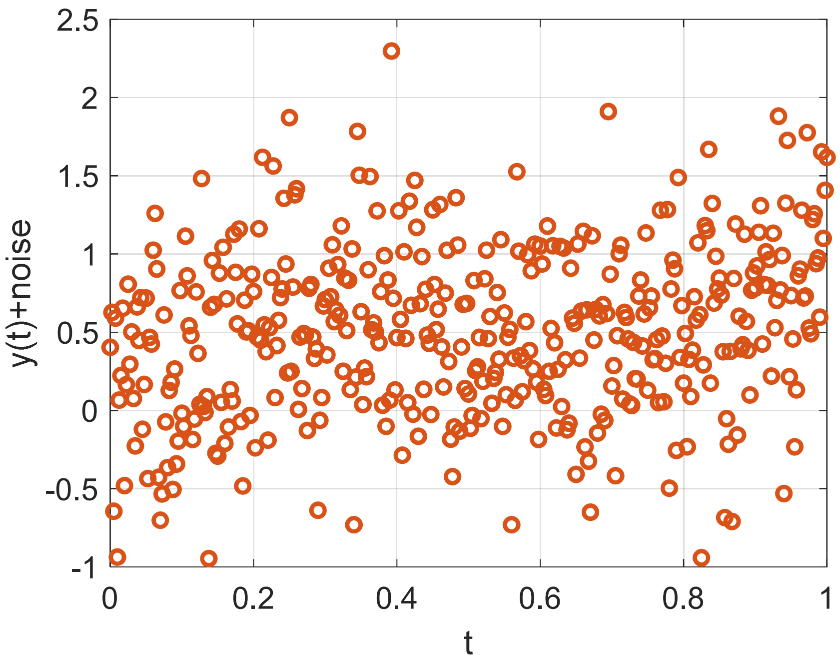
Acknowledgements

This work was supported by U.S. DOE-NASA Strategic Partnership Project (SPP) #19701.

The authors are grateful to Dr. J. Nick Benardini and Dr. Elaine E. Seasly for their contribution to this project.

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy. INL is the nation's center for nuclear energy research and development, and also performs research in each of DOE's strategic goal areas: energy, national security, science and the environment.

Learning from Noisy Data

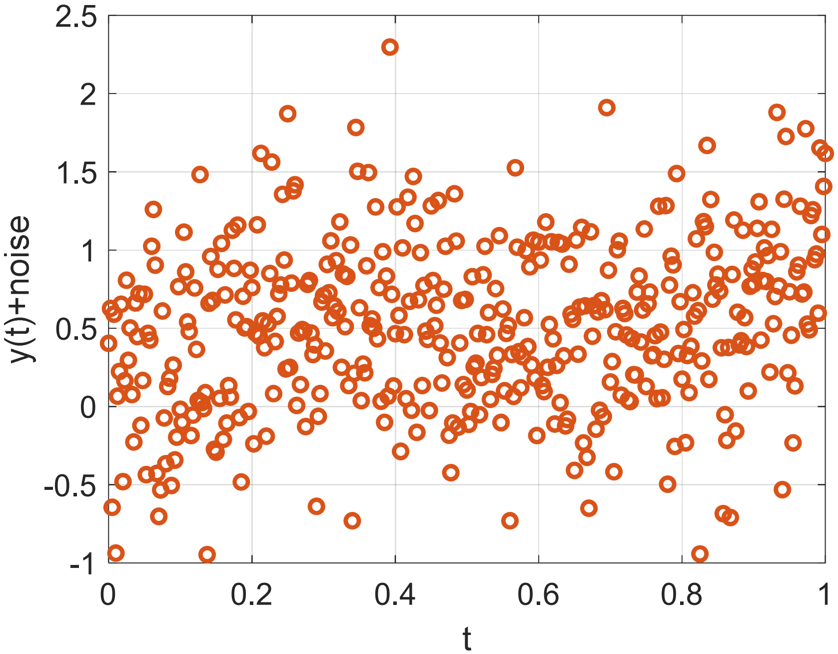


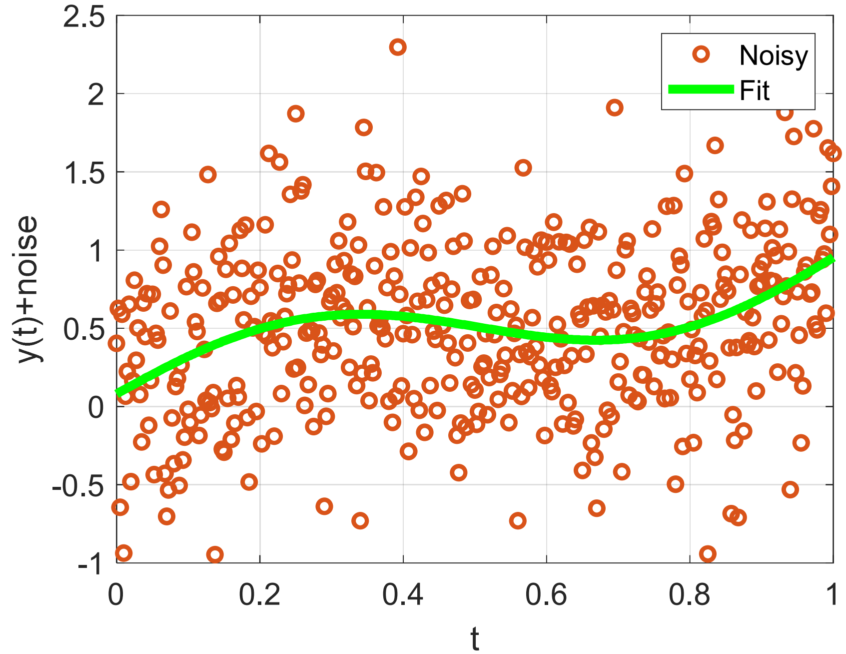
In 1923, the French mathematician Hadamard introduced the notion of well-posed problems.

According to Hadamard, a problem is called well-posed iff:

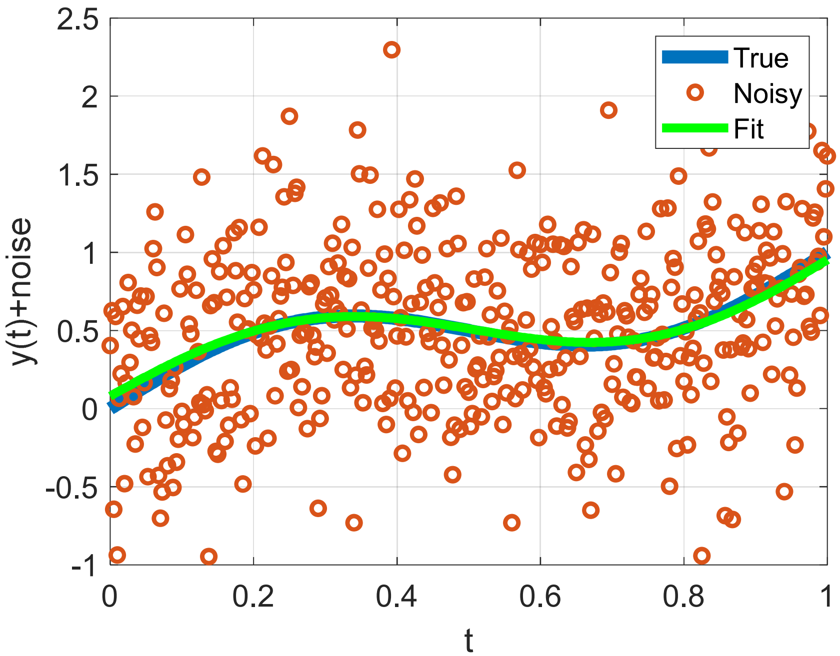
1. A solution to the problem exists (existence).
2. This solution is unique (uniqueness).
3. This unique solution is stable under small perturbations in the data, in other words small perturbations in the data should cause small perturbations in the solution (stability).

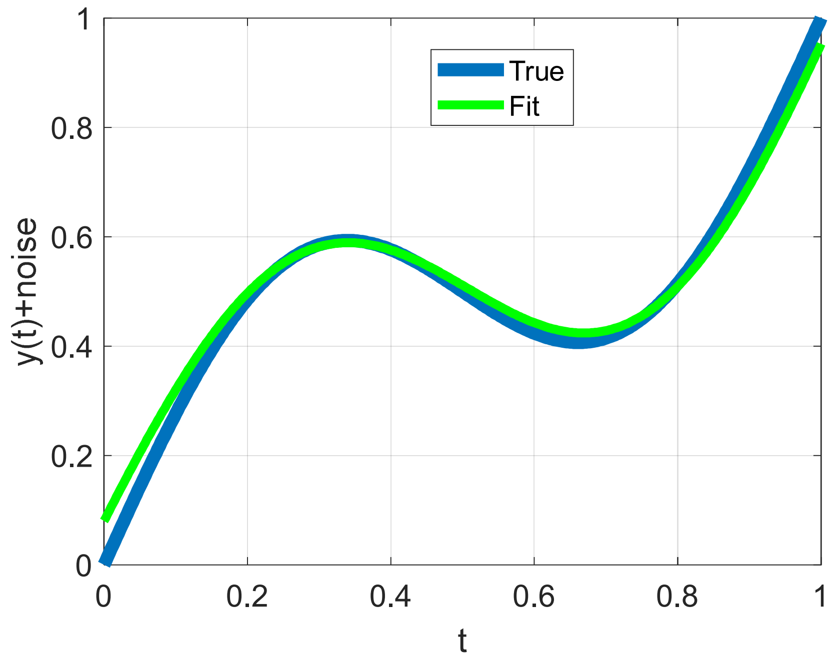
Learning from Noisy Data





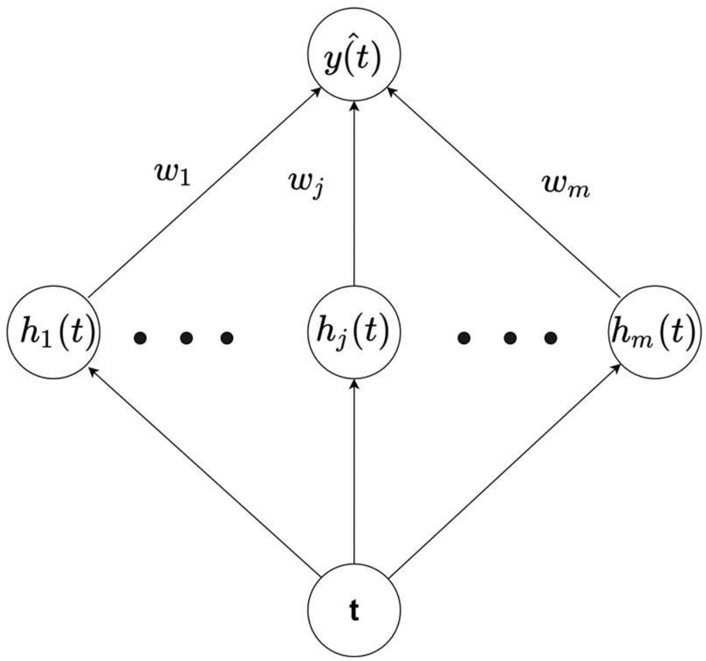
Learning from Noisy Data





Jet Propulsion Laboratory
California Institute of Technology

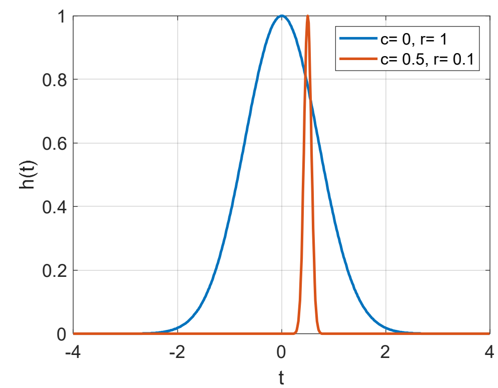
Radial Basis Function Network and Regularization



$$\hat{y}(t_1) = w_1 \cdot h_1(t_1) + w_2 \cdot h_2(t_1) + \cdots w_m \cdot h_m(t_1)$$
$$\hat{y}(t_2) = w_1 \cdot h_1(t_2) + w_2 \cdot h_2(t_2) + \cdots w_m \cdot h_m(t_2)$$
$$\vdots$$
$$\hat{y}(t_p) = w_1 \cdot h_1(t_p) + w_2 \cdot h_2(t_p) + \cdots w_m \cdot h_m(t_p)$$

$$\hat{y}(t) = \sum_{j=1}^{m} w_j \cdot h_j(t)$$

$$\underbrace{\hat{y}}_{px1} = \underbrace{H}_{pxm} \cdot \underbrace{w}_{mx1}$$



$$h(t) = e^{\left(-\frac{(t-c)^2}{r^2}\right)}$$

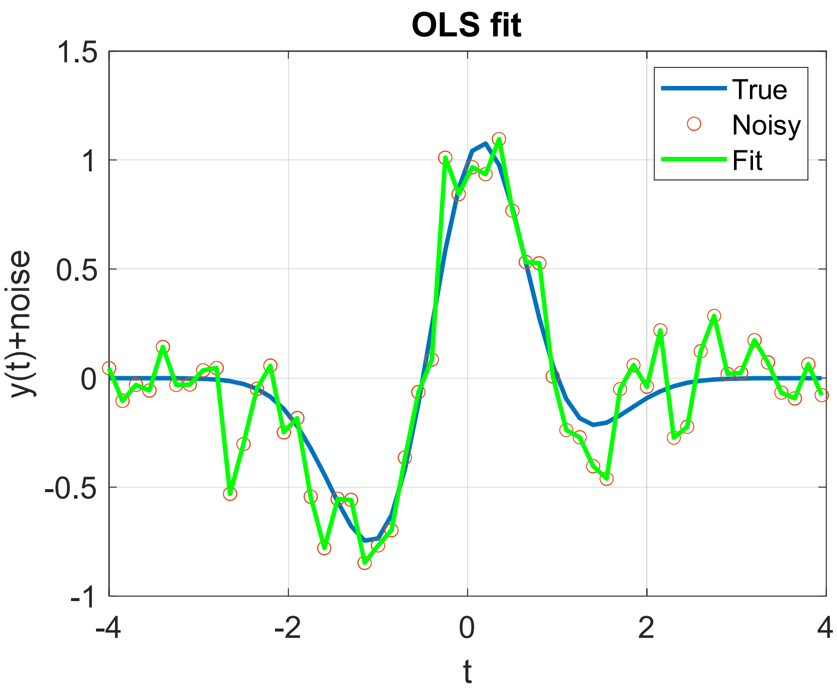
Ordinary Least Squares

$$E = \sum_{i=1}^{p} (\hat{y}(t_i) - \tilde{y}_i)^2 = (\tilde{y} - H \cdot w)^T \cdot (\tilde{y} - H \cdot w)$$

$$w = (H^T \cdot H)^{-1} H^T \cdot \tilde{y}$$

$$\hat{y}(t) = \sum_{j=1}^{m} w_j \cdot h_j(t)$$

$$\{(t_i, \tilde{y}_i)\}_{i=1}^{p}$$



Regularized Least Squares and Generalized Cross-Validation (GCV)

$$E = \sum_{i=1}^{p} (\hat{y}(t_i) - \tilde{y}_i)^2 + \mu \cdot \sum_{j=1}^{m} w_j^2 = (\tilde{y} - H \cdot w)^T \cdot (\tilde{y} - H \cdot w) + \mu \cdot w^T \cdot w$$

$$w_\mu = (H^T \cdot H + \mu \cdot I_m)^{-1} H^T \cdot \tilde{y}$$

$$\underbrace{\hat{y}}_{px1} = \underbrace{H}_{pxm} \cdot \underbrace{(H^T \cdot H + \mu \cdot I)^{-1} H^T \cdot \tilde{y}}_{mx1} = A_\mu \cdot \tilde{y}$$

$$A_\mu = H \cdot (H^T \cdot H + \mu \cdot I)^{-1} H^T$$

$$GCV(\mu) = \frac{\left\| H \cdot w_\mu - \tilde{y} \right\|^2}{\left(m - trace(A_\mu)\right)^2}$$