Entropy, Targeted, and Redundant Observations For Filtering Turbulent Signals

Marcus Grote

Marcus.Grote@unibas.ch

University of Basel, Switzerland

joint work with:

Andrew Majda, Courant Institute and CAOS, New York University

UNI BASEL

Outline

- Data assimilation
- Bayesian statistics
- Information theory
- Kalman filtering
- Prototype stochastic model
- Sparse irregularly spaced observations
- Redundant observations
- Targeted observations
- Concluding remarks

UNI BASEL

Data assimilation

- available data (+/- 3-h): $10^4 - 10^5$ (nonuniform, 4D) observations
- numerical model (PEM, 1° horiz., 20 vert.): $\sim 10^7$ dof's

Ghil & Malanotte-Rizzoli (1991):

> *...noisy, inaccurate data should not be fitted by exact interpolation, but rather by a procedure designed to achieve two goals simultaneously: (1) to extract the valuable information contained in the data, and (2) to filter out the spurious information, i.e. the noise.*

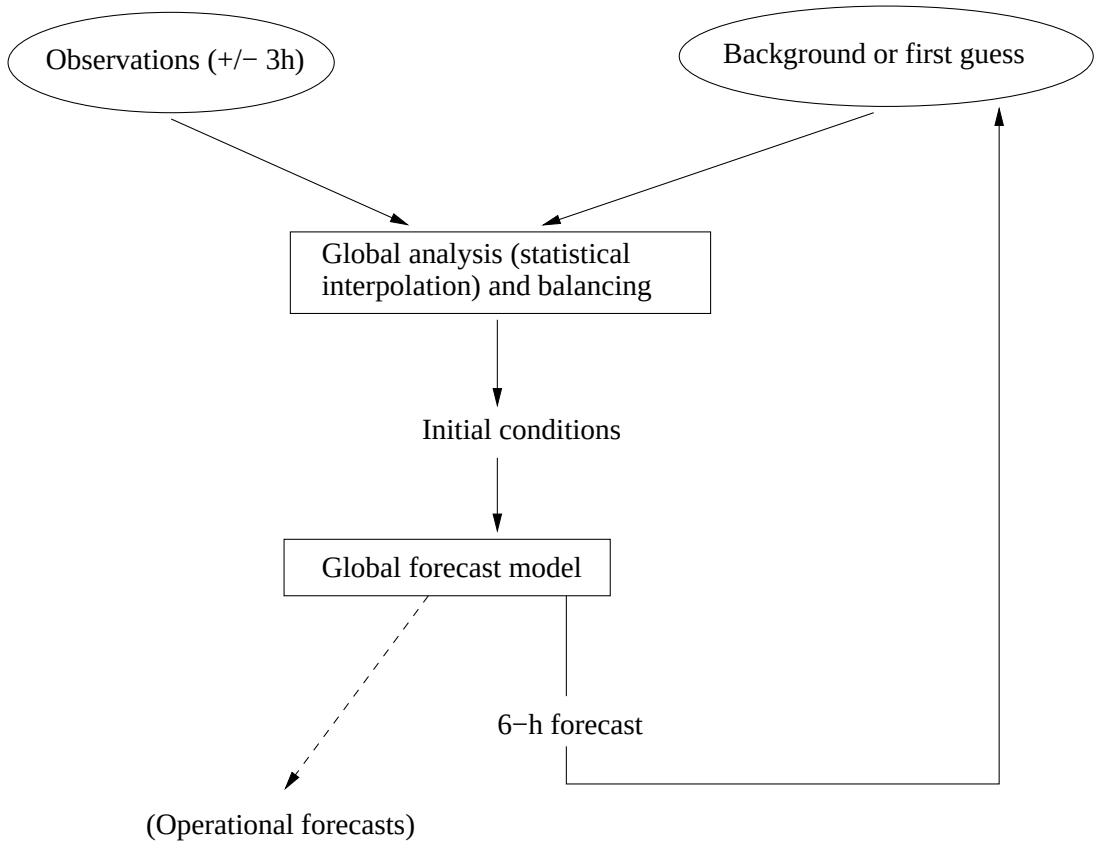
Talagrand (1997):

> *Assimilation of meteorological or oceanographical observations can be described as the process through which all the available information is used in order to estimate as accurately as possible the state of the atmospheric or oceanic flow. The available information essentially consists of the observations proper, and of physical laws that govern the evolution of the flow. The latter are available in practice under the form of a numerical model.*
>
> *...should produce not only an estimate of the state of the flow, but also an estimate of the associated uncertainty.*

UNI BASEL

Data assimilation cycle in NWP

Typical (6-h) data assimilation cycle:



Analysis: add innovations to model forecast with weights $\mathbf{W}$

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}\left[\mathbf{y}^{obs} - \mathbf{G}(\mathbf{x}^b)\right]$$

Statistical least squares estimation

Goal: estimate temperature T
Data: independent measurements T_1, T_2

Error statistics:

$$T_1 = T + \varepsilon_1, \quad E[\varepsilon_1] = 0, \quad E[\varepsilon_1^2] = \sigma_1^2$$
$$T_2 = T + \varepsilon_2, \quad E[\varepsilon_2] = 0, \quad E[\varepsilon_2^2] = \sigma_2^2, \quad E[\varepsilon_1 \varepsilon_1] = 0.$$

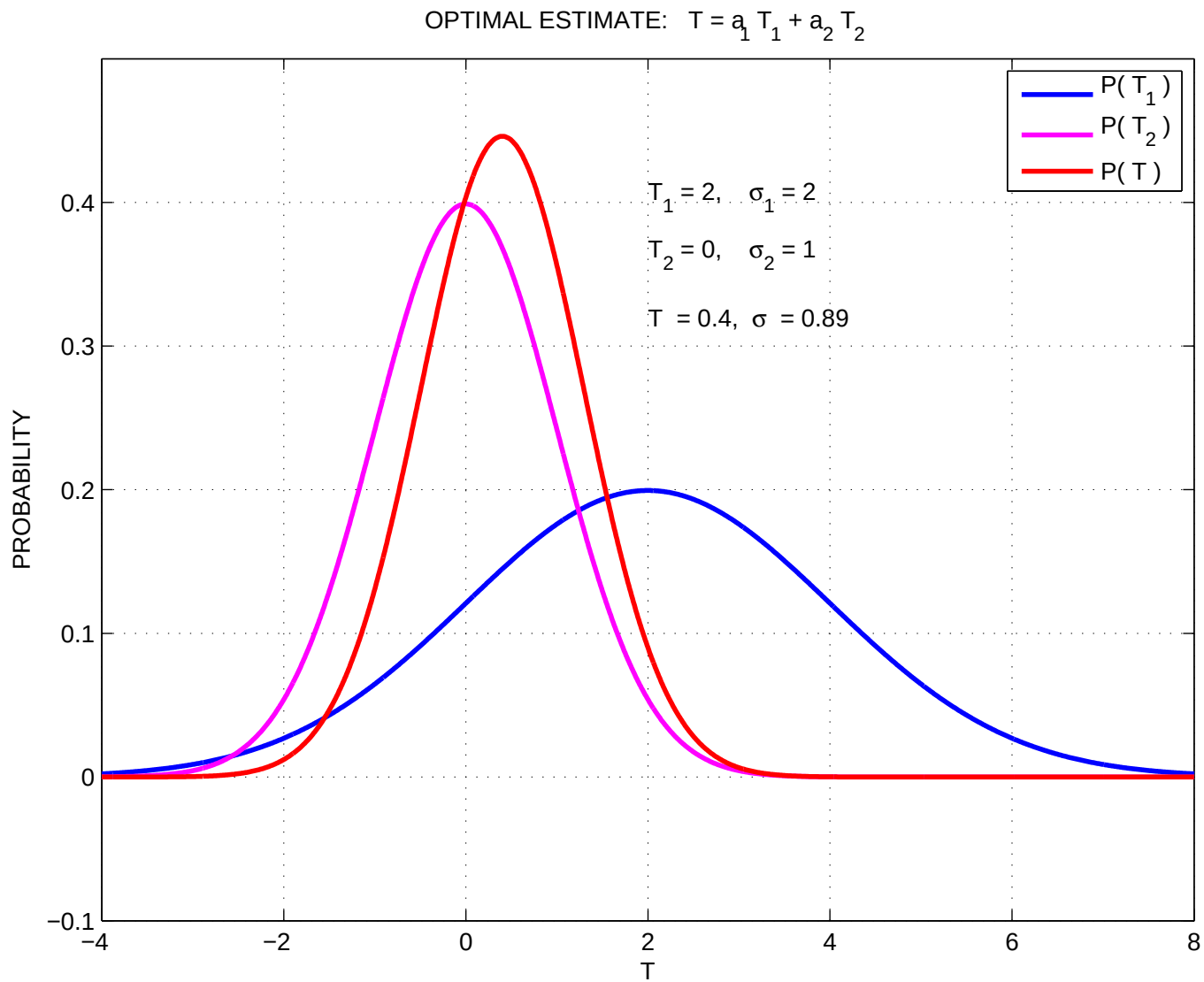
Linear estimator: $T^a = a_1 T_1 + a_2 T_2$

unbiased: $E[T^a] = T \;\Rightarrow\; a_1 + a_2 = 1$
minimal variance: $\sigma^2 = E[(T^a - T)^2]$ must be minimal

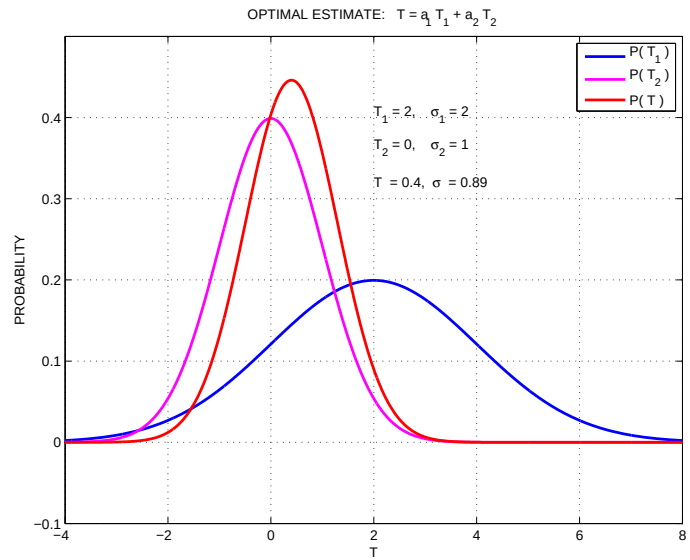
$$\Rightarrow \qquad a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \qquad a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$
$$\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \qquad \frac{1}{\sigma^2} = \text{“precision”}$$

Statistical least squares estimation (contd.)

Bayesian interpretation: $$P_{post}(T) = P(T|T_2) = \frac{P(T_2|T)\, P_{prior}(T)}{P(T_2)}$$



Statistical least squares estimation (contd.)



The PDF $p_{post}(T)$

- is “narrower”,
- has a smaller variance,
- is less “uncertain”,
- is more “informative”...

But how do we quantify the information content associated with $p_{post}(T)$?

Information Theory

C. Shannon (1916–2001)

founder of information theory

boolean algebra, digital circuit design

signal processing, cryptography

Shannon's mouse, chess program

radio controlled cars, juggling, unicycling...

ultimate machine

Information theory (Shannon, 1948)

Let $A = \{\text{binary words of length } n\}$, $\#A = N = 2^n$

Amount of information needed for $x \in A$, $x = \underbrace{01011001\ldots001}_{n}$:

$$n = \log_2 N$$

Next, let $A = A_1 \cup A_2 \cup \cdots \cup A_k$, $\#A_i = N_i$, $A_i \cap A_j = \emptyset, i \neq j$:

$$\mathcal{P}(x \in A_i) = N_i/N = p_i, \qquad \text{event } A_i.$$

$$\underbrace{\sum p_i \log_2 N_i}_{\text{info if } x\in A_i} \underbrace{- \sum p_i \log_2 p_i}_{\text{lack of info}} = \underbrace{\log_2 N}_{\text{total info } x\in A}$$

Shannon entropy: $\mathcal{S}(p) = -\sum_{i=1}^{k} p_i \ln p_i > 0$

see Majda & Wang 2006

The Shannon entropy

The Shannon entropy $\mathcal{S}(p) = -\sum_{i=1}^{k} p_i \ln p_i$ has the following properties:

1. $\mathcal{S}(p_1, \ldots, p_k)$ continuous,
2. $S(1/k, \ldots, 1/k)$ is monotonic increasing as $k \to \infty$,
3. Composition law: $\mathcal{S}(p)$ invariant under splitting $A = A_1 \cup A_2$

The Shannon entropy is unique up to scaling (Jaynes, 1957),

and measures the lack of information or the uncertainty associated with p.

For a continuous probability density function $p(\lambda)$:

$$\mathcal{S}(p) = -\int_{-\infty}^{\infty} p \ln(p)\, d\lambda > 0 \ .$$

UNI BASEL

The Shannon entropy: Gaussian case

For a normal distribution $X \sim \mathcal{N}(\bar{X}, R)$, $X \in \mathbb{R}^n$, $R \in \mathbb{R}^{n \times n}$,

$$\mathcal{S}(p) = \frac{1}{2}[\ln \det(R) + n + n \ln(2\pi)] \ .$$

In the special case $n = 1$, $X \sim \mathcal{N}(\mu, \sigma)$, $\mu \in \mathbb{R}$ and $\sigma > 0$ we have:

$$\mathcal{S}(p) = \ln \sigma + (1 + \ln(2\pi))/2 \ .$$

Thus in the Gaussian case, minimizing the Shannon entropy corresponds to minimizing the variance $\Rightarrow$ intuitively correct.

However, Shannon entropy immediately applies to the non-Gaussian case, too!

Statistical Gaussian least squares

Random variable : $X \in \mathbb{R}^n, \quad X \sim \mathcal{N}(\bar{X}_{pr}, R_{pr})$

Observation : $Y = GX + \sigma, \quad G : \mathbb{R}^n \to \mathbb{R}^m, \quad \sigma \sim \mathcal{N}(0, R^o)$

Then the posterior (Gaussian) distribution is:

$$p(X|Y) = \mathcal{N}(\bar{X}_{post}, R_{post})$$

$$\begin{aligned} \bar{X}_{post} &= (I - KG)\bar{X}_{pr} + KY \qquad \text{(unbiased minimal variance estimator)} \\ R_{post} &= (I - KG)R_{pr} \\ K &= R_{pr}G^\top (GR_{pr}G^\top + R^o)^{-1} \qquad \text{(Kalman gain matrix)} \end{aligned}$$

REMARK:

- If $KG \longrightarrow I$, trust the observation
- If $KG \longrightarrow 0$, trust the prediction

Statistical Gaussian least squares (contd.)

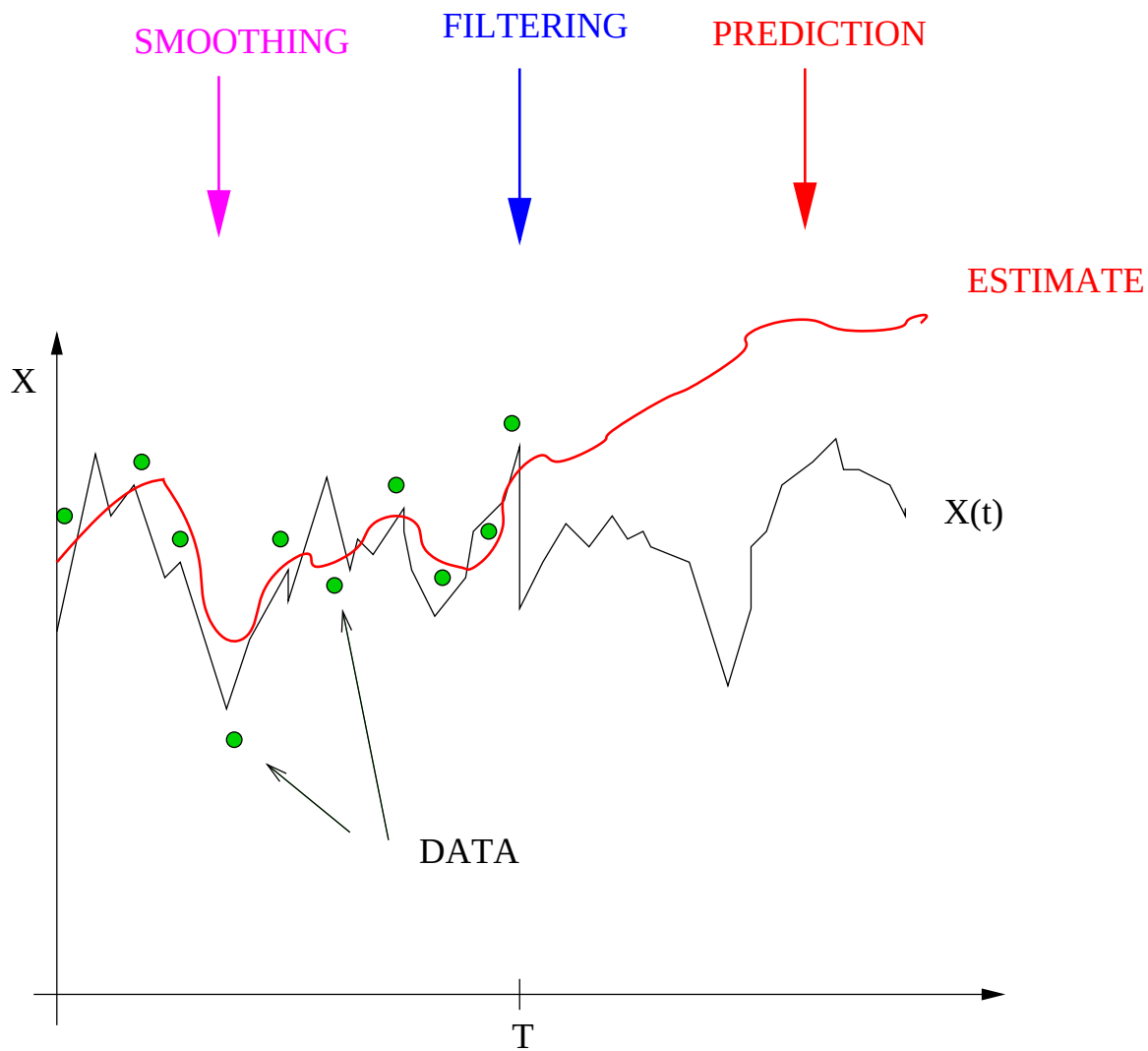
The gain in information or reduction in uncertainty due to the observation Y is given by the Shannon entropy difference:

$$\begin{aligned}\mathcal{S}(p_{pr}) - \mathcal{S}(p_{post}) &= \frac{1}{2}[\ln\det(R_{pr}) - \ln\det(R_{post})] \\ &= \frac{1}{2}[\ln\det(R_{pr}) - \ln\det((I-KG)R_{pr})] \\ &= -\frac{1}{2}\ln\det(I-KG).\end{aligned}$$

Let λ_i be the singular values of the $m \times n$ scaled observation operator $M = (R^o)^{-1/2} G R_{pr}^{1/2}$. Then one can show (Xu, Tellus 2006):

$$\Delta\mathcal{S} = \mathcal{S}(p_{pr}) - \mathcal{S}(p_{post}) = \frac{1}{2}\sum_i \ln(1+\lambda_i^2) \geq 0.$$

Smoothing, filtering and prediction



Bayesian filtering: nonstationary case

state vector $\{X_k\}_{k=0}^{\infty}, \quad X_k \in \mathbb{R}^N,$ Markov process

$$\pi(x_{k+1}|x_0, x_1, \ldots, x_k) = \pi(x_{k+1}|x_k)$$

observation $\{Y_k\}_{k=0}^{\infty}, \quad Y_k \in \mathbb{R}^M,$ Markov process w.r.t. $\{X_k\}$

$$\pi(y_k|x_0, x_1, \ldots, x_k) = \pi(y_k|x_k)$$

and also assume

$$\pi(x_{k+1}|x_k, y_1, \ldots, y_k) = \pi(x_{k+1}|x_k).$$

Evolution-observation model:

$$\begin{array}{ccccccccc} X_0 & \to & X_1 & \to & X_2 & \to \cdots \to & X_n & \to \ldots \\ & & \downarrow & & \downarrow & & \downarrow & \\ & & Y_1 & & Y_2 & & Y_n & \end{array}$$

Bayesian filtering: nonstationary case (contd.)

Denote all measurements until t_k as: $D_k = \{y_1, \ldots, y_k\}$

FILTER UPDATING STEPS:

Time evolution: Given $\pi(x_k|D_k)$, find $\pi(x_{k+1}|D_k)$
based on Markov transition kernel $\pi(x_{k+1}, x_k)$

Observation: Given $\pi(x_{k+1}|D_k)$, find $\pi(x_{k+1}|D_{k+1})$
based on the new observation y_{k+1} and the likelihood $\pi(y_{k+1}|x_{k+1})$

GOAL: obtain $\pi(x_{k+1}|D_{k+1})$, $\Rightarrow$ estimate X_{k+1} given all available data D_{k+1}

The Kalman filter (1960)

Linear state equations with additive noise

$$X_{k+1} = FX_k + W_{k+1}, \qquad k = 0, 1, \ldots, \quad X_k \in \mathbb{R}^N$$
$$Y_{k+1} = GX_{k+1} + V_{k+1}, \qquad Y_k \in \mathbb{R}^M$$

where $X_0, \{V_k\}, \{W_k\}$ are mutually independent and Gaussian, with

$$E[X_0] = x_0, \quad E[X_0 X_0^\top] = R_0,$$
$$E[V_k] = 0, \quad E[V_k V_j^\top] = \delta_{kj} R_v,$$
$$E[W_k] = 0, \quad E[W_k W_j^\top] = \delta_{kj} R_w,$$

with F, G, R_0, R_v, R_w constant matrices, for simplicity

The Kalman filter (contd.)

Notation: $x_{m|\ell} = E[X_m|D_\ell]$, $R_{m|\ell} = \mathsf{cov}(X_m|D_\ell)$

Time evolution updating

$$x_{m+1|m} = F\,x_{m|m}, \tag{1}$$
$$R_{m+1|m} = F\,R_{m|m}F^\top + R_w, \tag{2}$$

Observation updating

$$x_{m+1|m+1} = x_{m+1|m} + K_{m+1}(y_{m+1} - Gx_{m+1|m}), \tag{3}$$
$$R_{m+1|m+1} = (I - K_{m+1}G)\,R_{m+1|m}, \tag{4}$$

where K_{m+1} is the Kalman gain matrix:

$$K_{m+1} = R_{m+1|m}\,G^\top(G\,R_{m+1|m}\,G^\top + R_v)^{-1}$$

Remark: Typically $R_{m|m} \to R_{\infty,\infty}$ and $K_m \to K_\infty$, $m \to 1$.
Both can be computed off-line.

Prototype stochastic model

Consider the damped and driven stochastic PDE in $[0, 2\pi]$:

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = -du + \mu\,\frac{\partial^2 u}{\partial x^2} + F(x,t) + \sigma(x)\dot{W}(t), \qquad c,\ \mu > 0\ .$$

In Fourier space:
add spatially correlated white noise in time.

$$du_k(t) = -(d + ick + \mu k^2)u_k(t)\,dt + F_k(t)\,dt + \sigma_k\,dW_k(t)\ ,$$

where W_k are independent (complex) Wiener processes, with

$$E_k = E_0|k|^{-\beta} = \frac{\sigma_k^2}{2(d+\mu k^2)}$$

Remark: Linear SDE: solved analytically
$\Rightarrow$ exact update formulas available for mean and variance.

(see Harlim, Majda, JCP, 2008)

Entropy and redundant observations

GOAL: Given asymptotic Kalman gain K_∞ and covariance $R_{\infty,\infty}$, measure degree of information redundancy in observations at $x_1, \ldots, x_M$.

IDEA:

1. Apply one Kalman evolution step: $R_{pr} = F R_{\infty,\infty} F^\top + R$
2. Compute $M = (R^o)^{-1/2} G R_{pr}^{1/2}$ with singular values λ_i
3. Monitor individual terms in information gain $\Delta\mathcal{S} = \frac{1}{2}\sum_i \ln(1+\lambda_i^2)$

Remark:
The dominant left singular vectors of $M = U\Lambda V^\top$ can be used for determining super-observations.

Regular sparse observations (resonant)

Negligible information loss: 21 vs. 3 locations, $\Delta t = 10$, $\mu = 0.01$

$\Delta \mathcal{S} = 3.75$

21 observations:

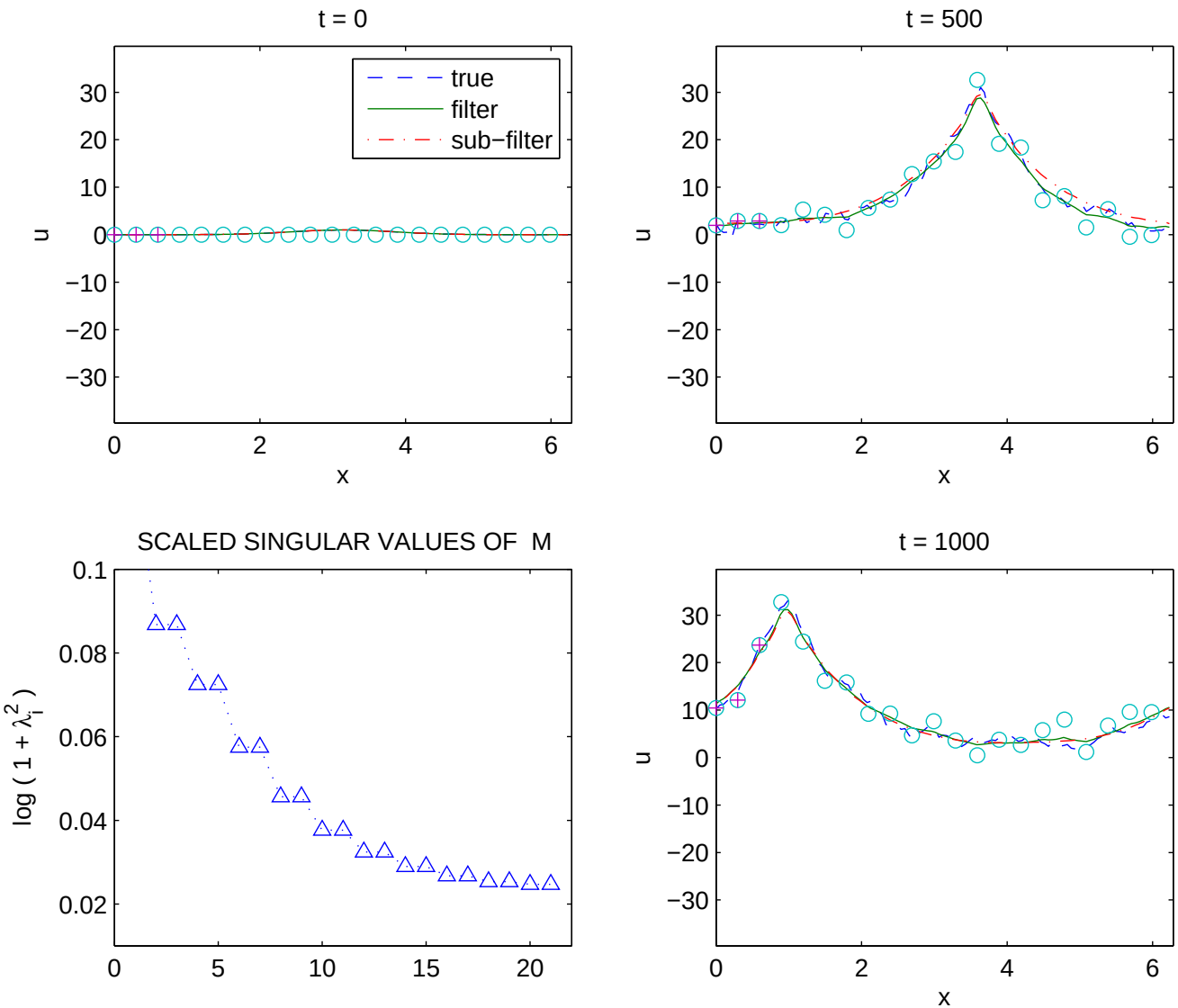
$< \text{err}_{\text{RMS}} >= 1.20$

$< \text{correl} >= 0.994$

3 observations:

$< \text{err}_{\text{RMS}} >= 1.40$

$< \text{correl} >= 0.991$



Regular sparse observations (non-resonant)

Significant information loss: 21 vs. 3 locations, $\Delta t = 0.1$, $\mu = 0.1$

$\Delta \mathcal{S} = 16.83$

21 observations:

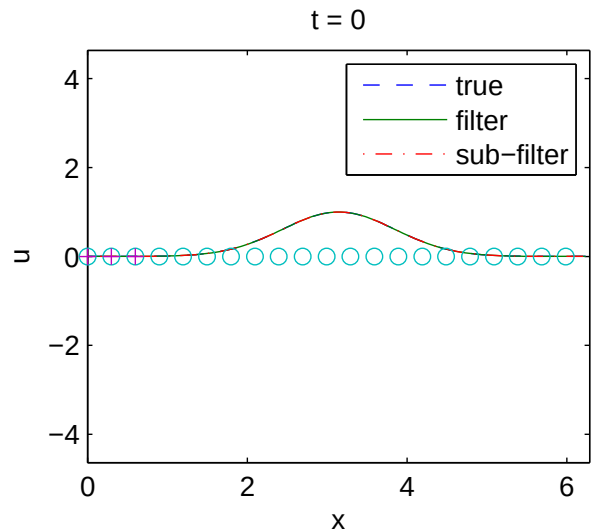
$< \text{err}_{\text{RMS}} >= 0.82$

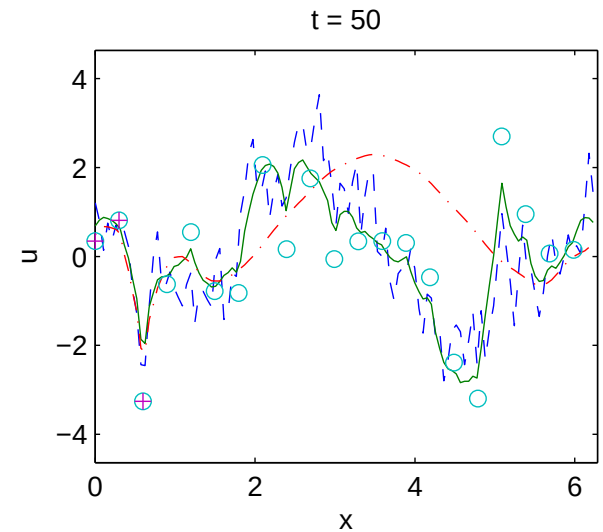
$< \text{correl} >= 0.82$

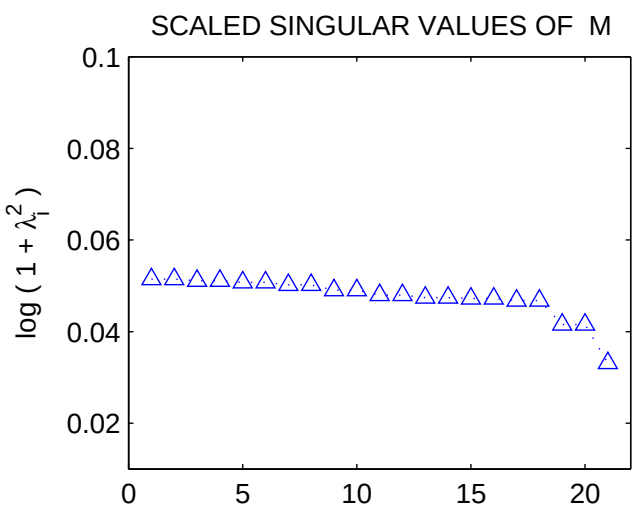
3 observations:

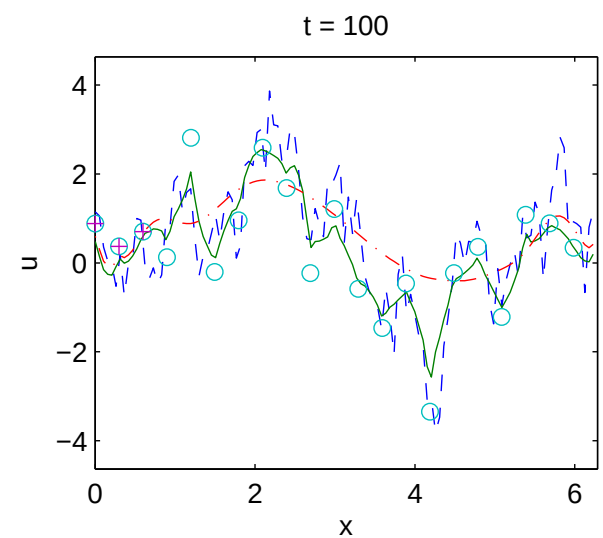
$< \text{err}_{\text{RMS}} >= 1.23$

$< \text{correl} >= 0.54$









Irregular sparse (packed) observations

Negligible information loss: 21 vs. 3 locations, $\Delta t = 0.1$, $\mu = 0.1$

$\Delta \mathcal{S} = 8.05$

21 observations:

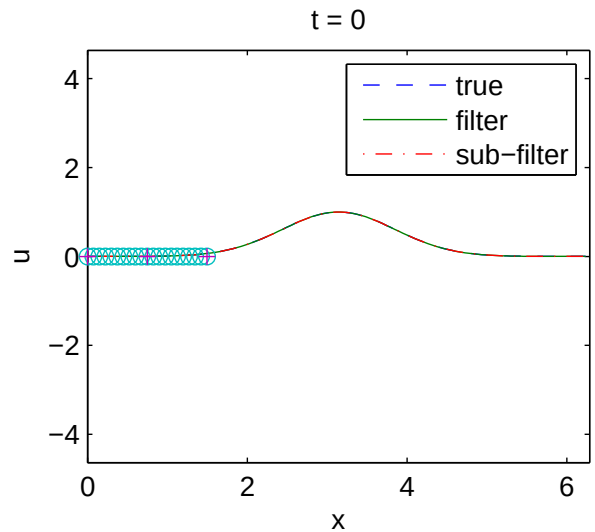
$< \text{err}_{\text{RMS}} >= 1.10$

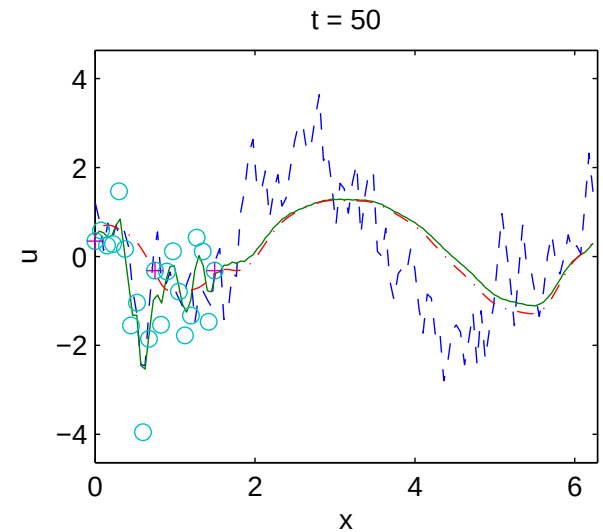
$< \text{correl} >= 0.65$

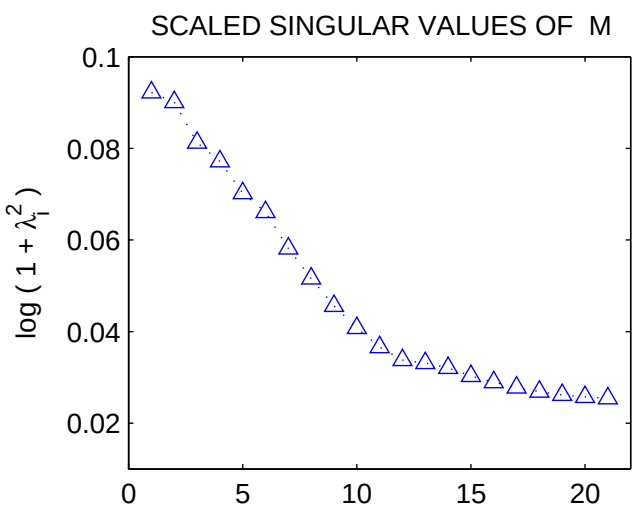
3 observations:

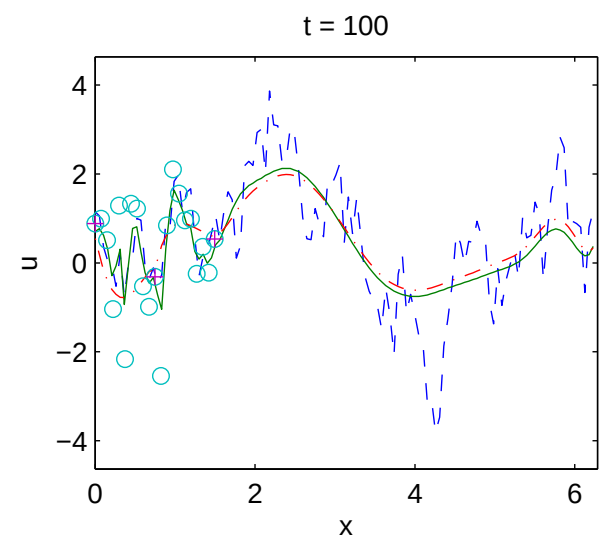
$< \text{err}_{\text{RMS}} >= 1.17$

$< \text{correl} >= 0.60$









Targeted observations

GOAL: Given observations at $x_1, \ldots, x_{M_1}$, select additional locations $x_{M_1+1}, \ldots, x_M$ that maximize information gain.

IDEA:

Given R_{∞,∞,M_1}

Maximize $\mathcal{S}(p_{M_1}; x_1, \ldots, x_{M_1}) - \mathcal{S}(p_M; x_1, \ldots, x_M)$

DIFFICULTIES:

- multi-dimensional non-convex optimization problem
- Every evaluation of $\mathcal{S}(p_M; x_1, \ldots, x_M)$ requires computing the asymptotic Kalman filter
- too expensive !

UNI BASEL

Targeted observations

Determine new observation site x^* by solving one-dimensional optimization problem for statistical least-squares.
By monotonicity of the Kalman filter we have:

$$\mathcal{S}(p_M; x_1, \ldots, x_{M_1}, x) \leq \mathcal{S}(p_{\text{post}}; x_1, \ldots, x_{M_1}, x) \leq \mathcal{S}(p_{M_1}; x_1, \ldots, x_{M_1})$$

ONE-STEP ALGORITHM:

Given R_{∞,∞,M_1}

Maximize $\mathcal{S}(p_{M_1}; x_1, \ldots, x_{M_1}) - \mathcal{S}(p_{\text{post}}; x_1, \ldots, x_{M_1}, x)$

Determine new site x^*

Compute true R_{∞,∞,M_1+1} with $x_{M+1} = x^*$

Targeted observations: example 1

Given 3 observations, determine new optimal site $x^* \in [\pi/4, 2\pi]$

3 observations:

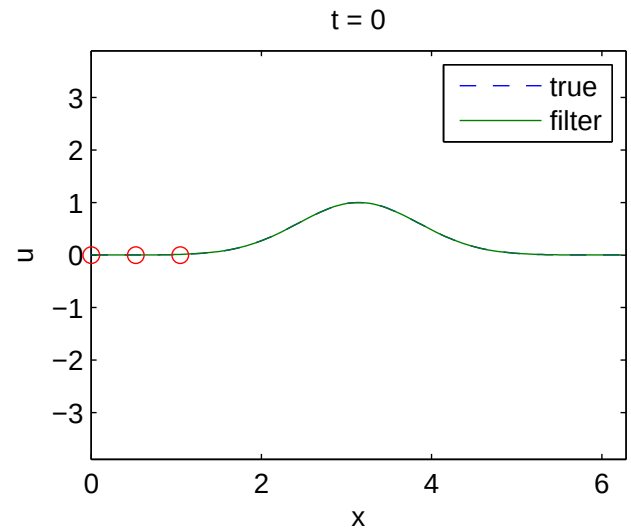
$\Delta\mathcal{S} = 3.15$

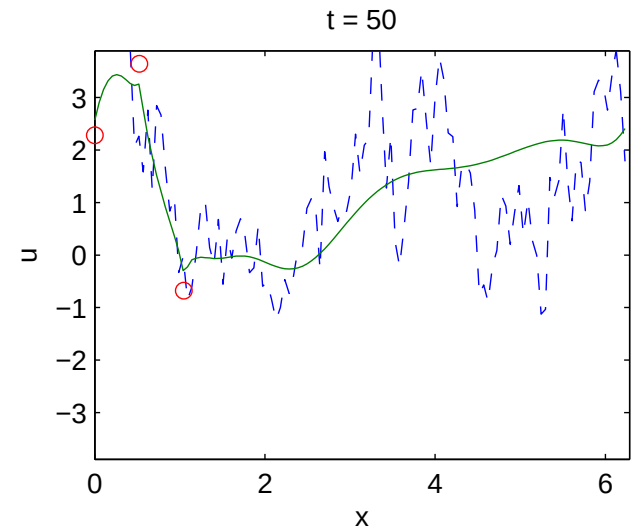
$< \text{err}_{\text{RMS}} >= 1.21$

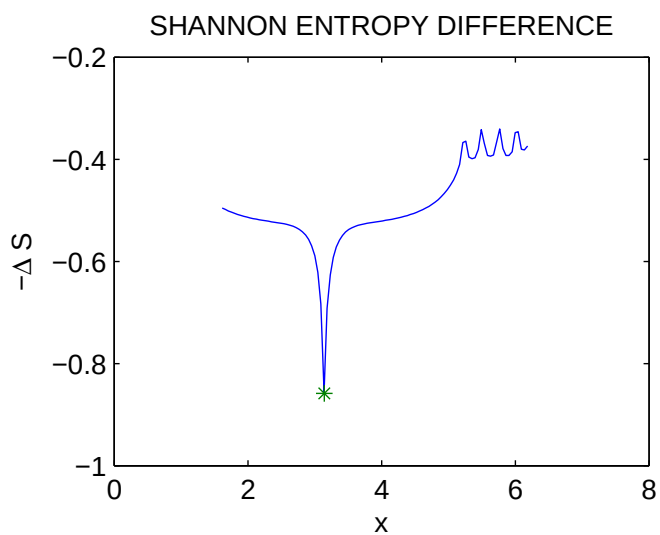
$< \text{correl} >= 0.63$

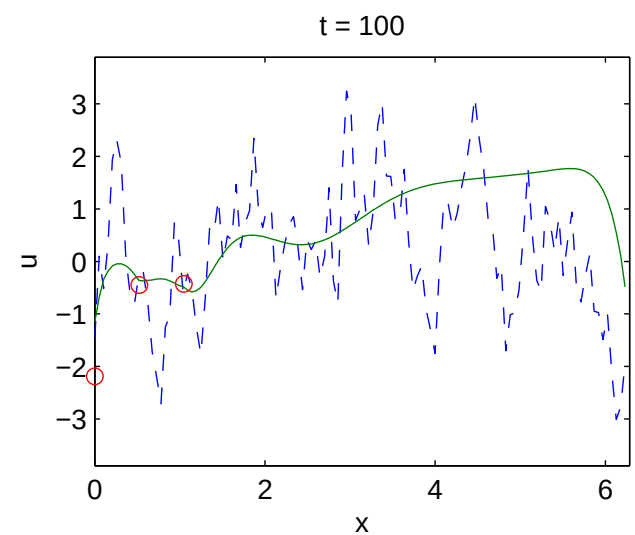
New site:

$x^* = 3.14$









Targeted observations: example 1

Given 4 observations, determine new optimal site $x^* \in [\pi/4, 2\pi]$

4 observations:

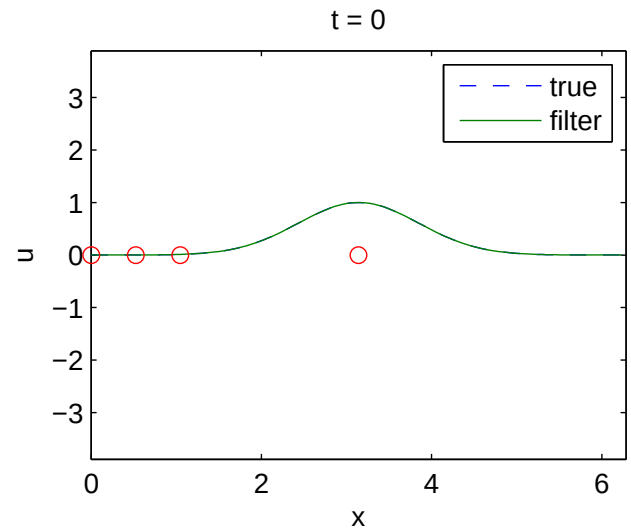
$\Delta\mathcal{S} = 1.56$

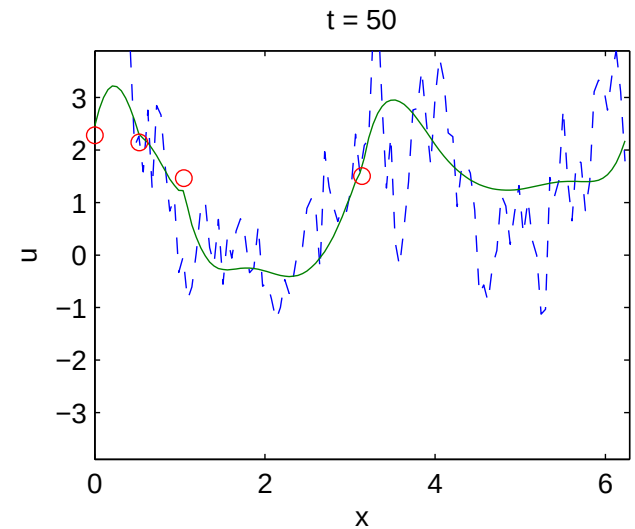
$< \text{err}_{\text{RMS}} >= 1.14$

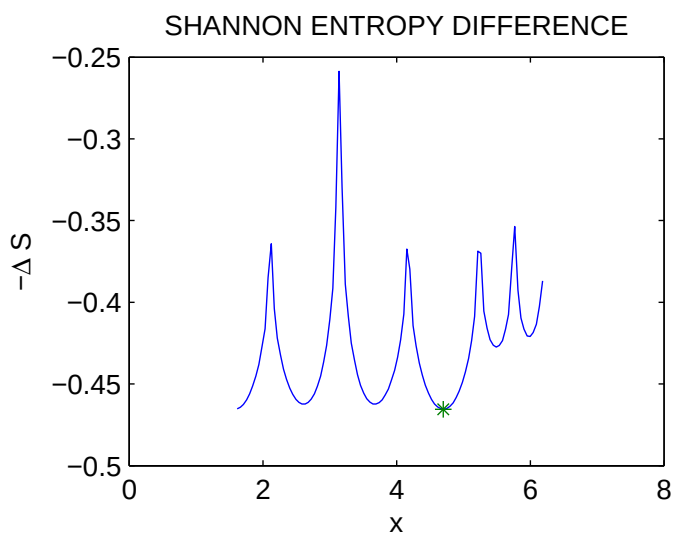
$< \text{correl} >= 0.68$

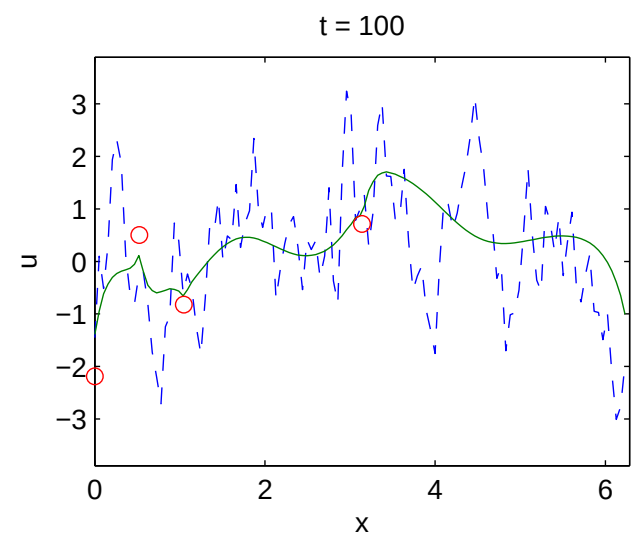
New site:

$x^* = 4.17$









Targeted observations: example 1

Given 5 observations, determine new optimal site $x^* \in [\pi/4, 2\pi]$

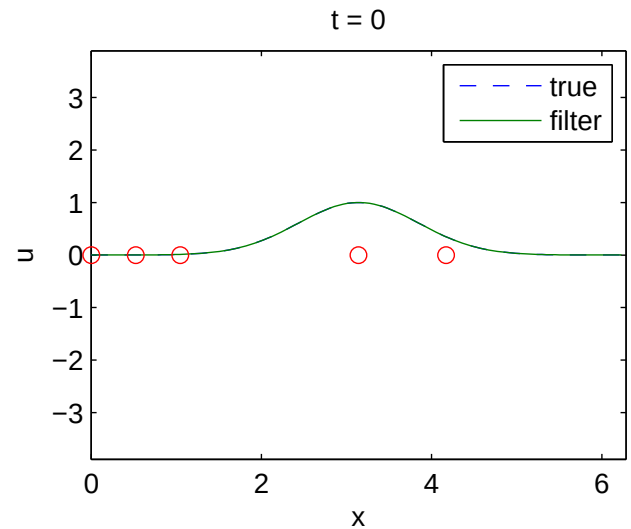
5 observations:

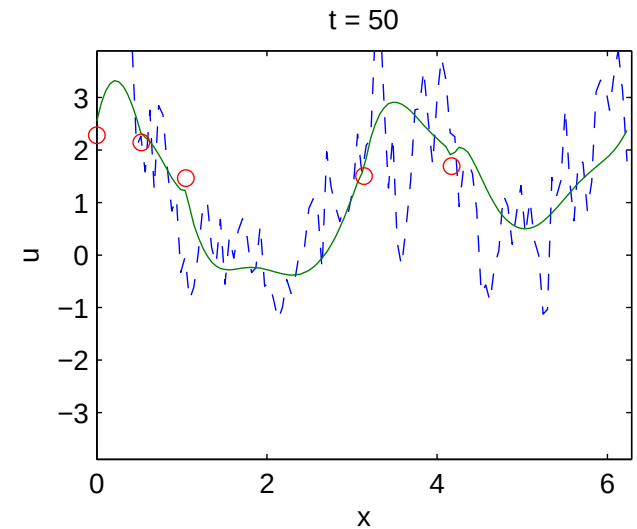
$< \text{err}_{\text{RMS}} >= 1.09$

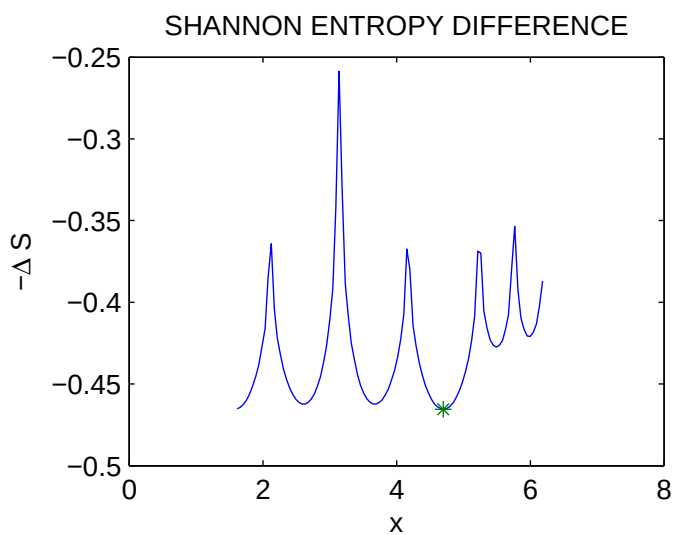
$< \text{correl} >= 0.71$

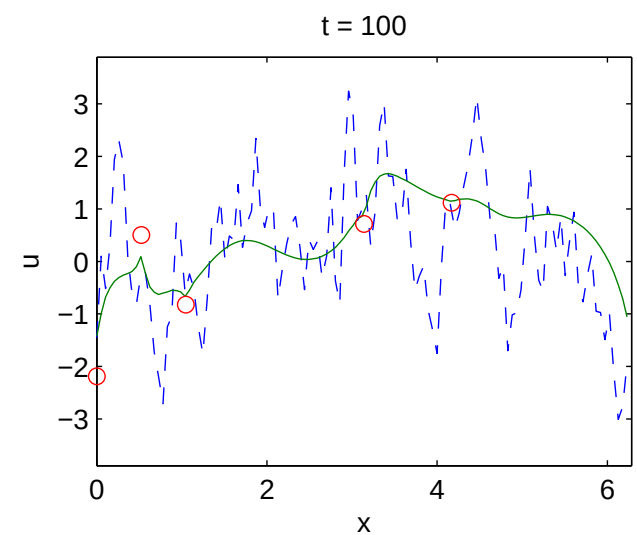
New site:

$x^* = 4.70$









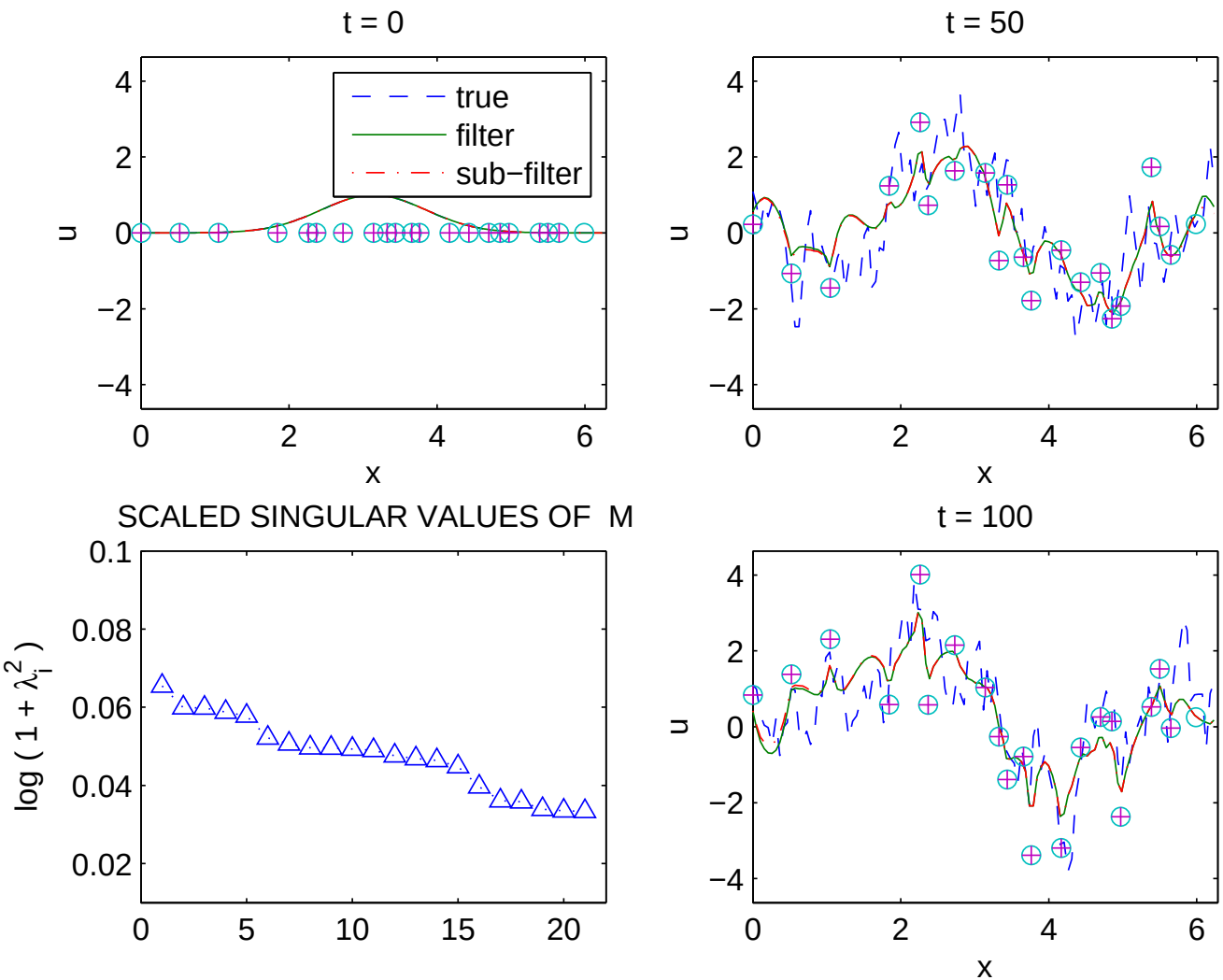
Targeted observations: example 1

Keep iterating...18 new sites:

21 observations:

$< \text{err}_{\text{RMS}} >= 0.84$

$< \text{correl} >= 0.83$



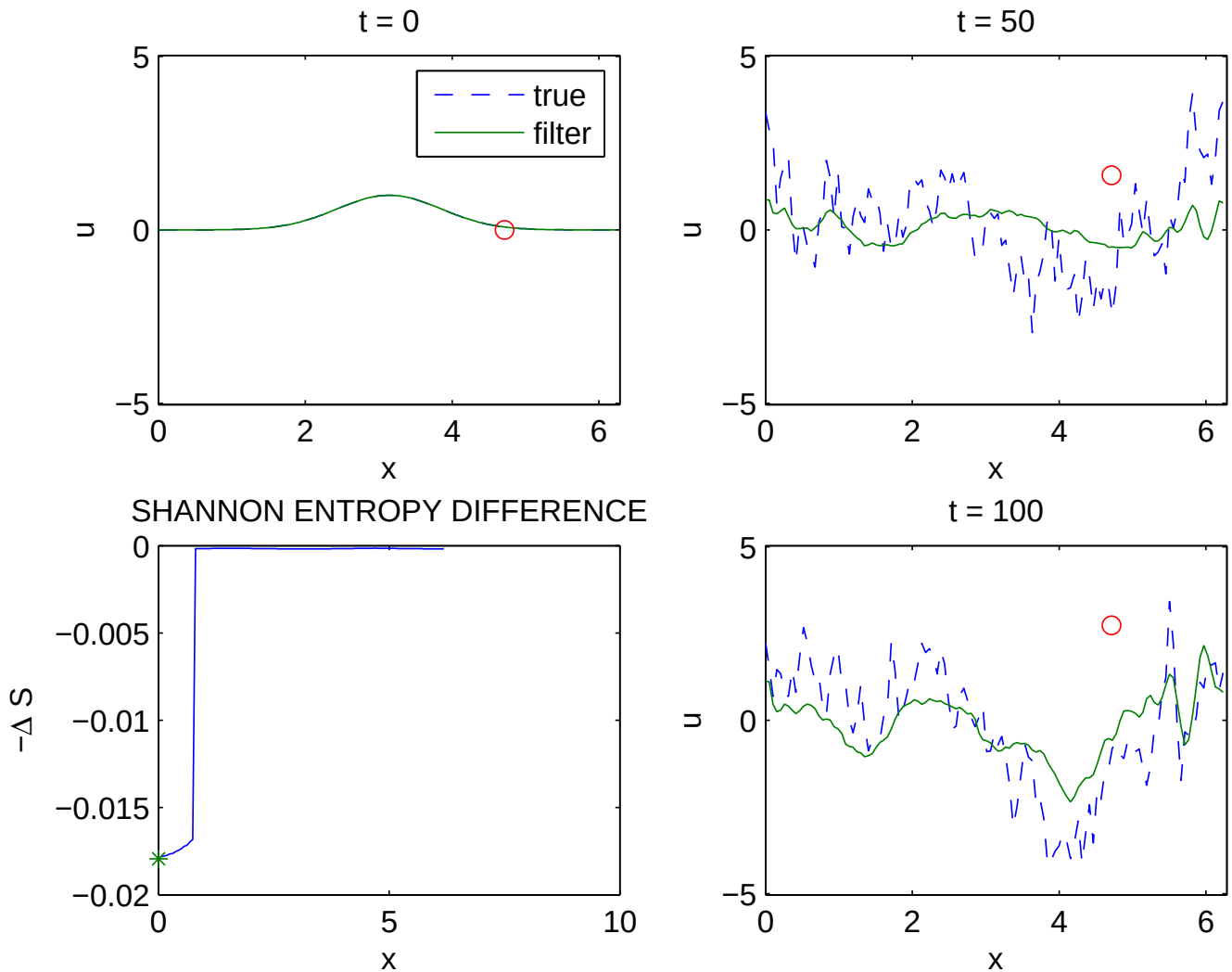
Targeted observations: example 2

Targeted vs. equidistant observations, $\Delta t = 0.1$, $\mu = 0.00001$
Given 1 observation at $x = 3\pi/2$, determine 3 new observation sites
Observation accuracy ten times higher for $0 \leq x \leq \pi/4$

1 observation:

$< \text{err}_{\text{RMS}} >= 1.05$

$< \text{correl} >= 0.72$



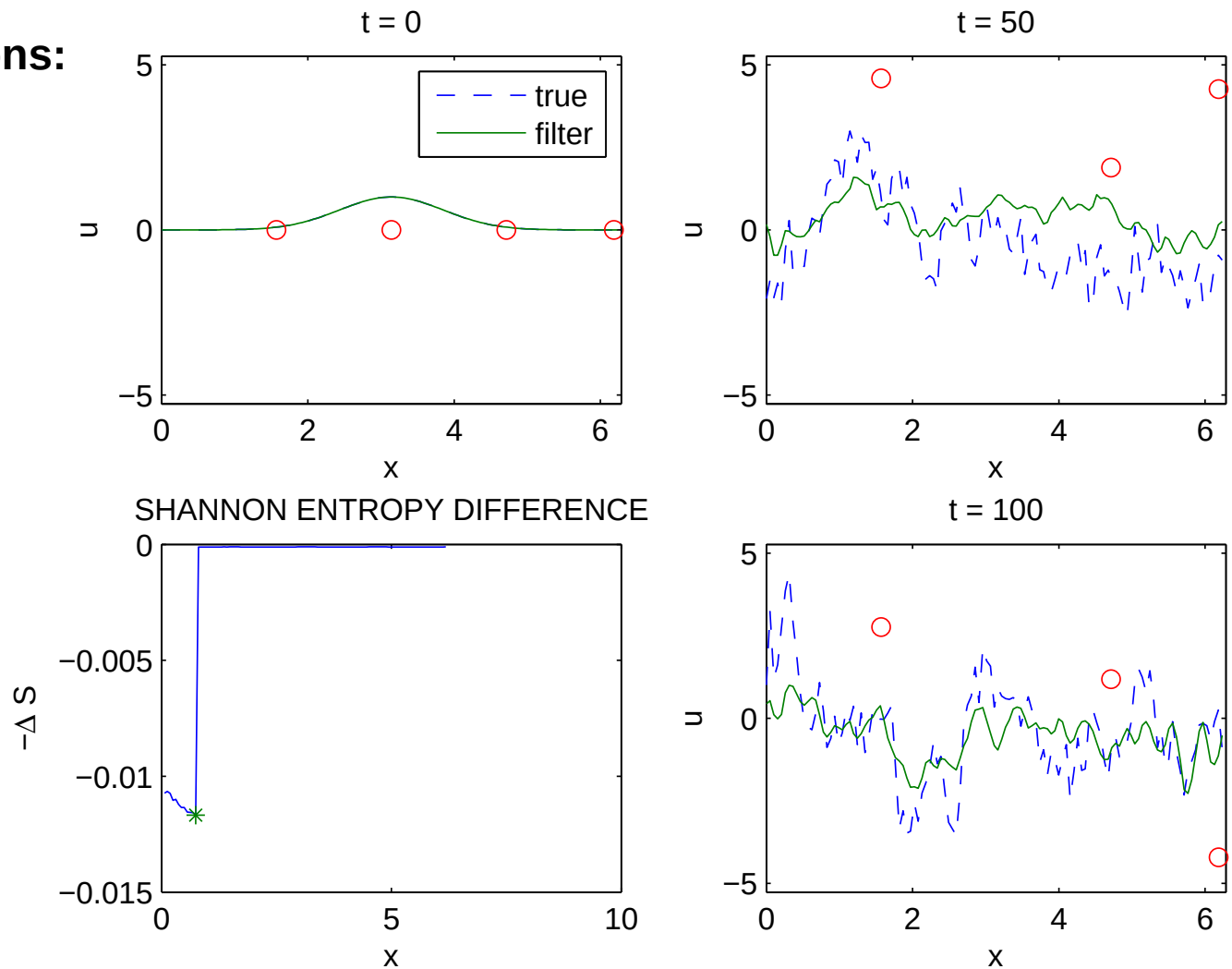
Targeted observations: example 2

Targeted vs. equidistant observations, $\Delta t = 0.1$, $\mu = 0.00001$
Given 1 observation at $x = 3\pi/2$, determine 3 new observation sites
Observation accuracy ten times higher for $0 \leq x \leq \pi/4$

4 equidistant observations:

$< \text{err}_{\text{RMS}} >= 0.95$

$< \text{correl} >= 0.73$



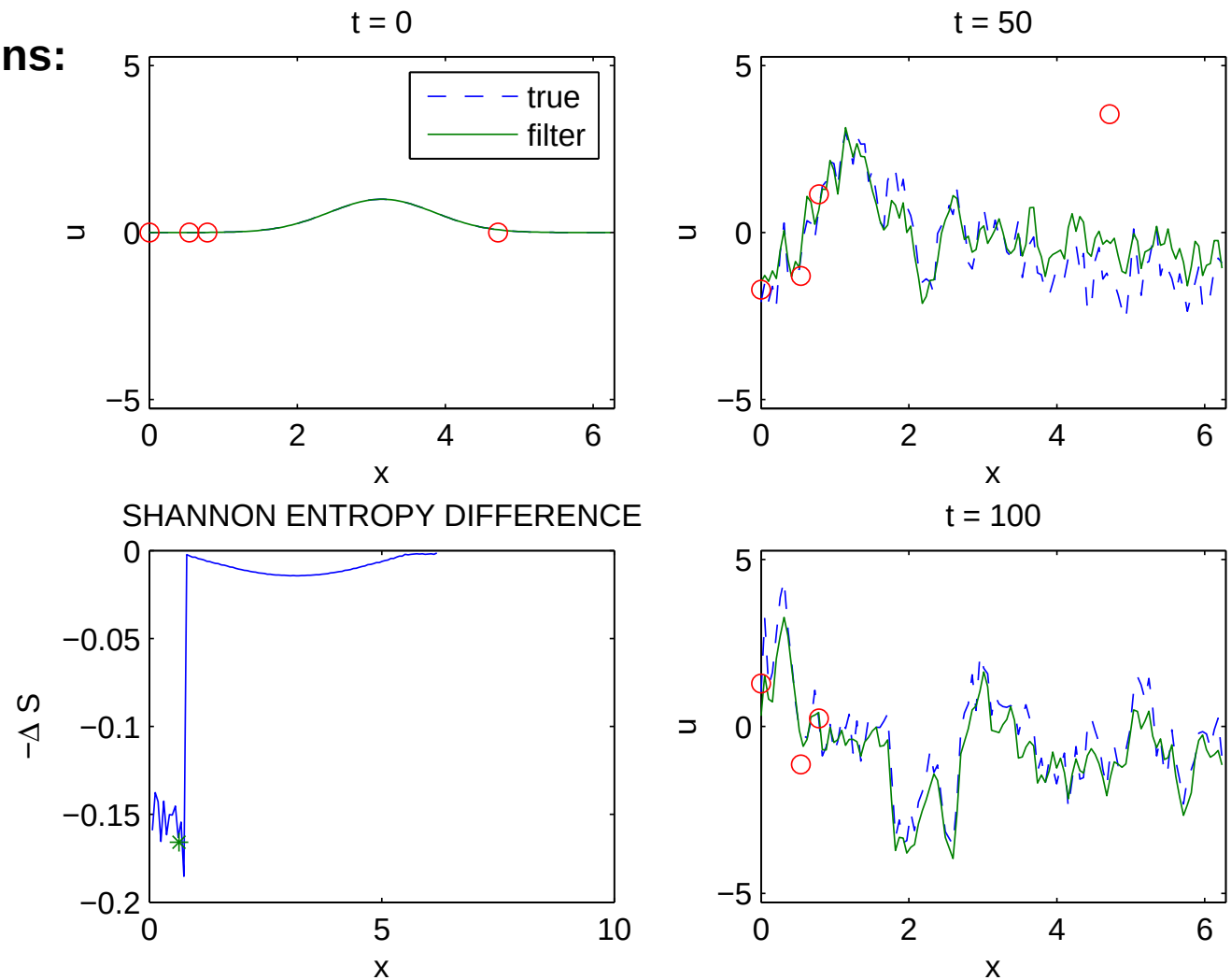
Targeted observations: example 2

Targeted vs. equidistant observations, $\Delta t = 0.1$, $\mu = 0.00001$
Given 1 observation at $x = 3\pi/2$, determine 3 new observation sites
Observation accuracy ten times higher for $0 \leq x \leq \pi/4$

1 + 3 targeted observations:

$< \text{err}_{\text{RMS}} >= 0.52$

$< \text{correl} >= 0.93$



Concluding remarks

- Shannon entropy measures information content from observations and identifies redundancy in data
- Maximizing information gain leads to effective strategies for targeted observations
- Applies to non-Gaussian case, too.
- Current work:
  - Study effect of simple (cheap) least-squares and full (expensive) asymptotic Kalman filter in Shannon entropy difference $\Delta\mathcal{S}(x)$
  - Use relative entropy to include effects of the (time varying) mean.
  - Devise strategies for real-time adaptive targeted observations
  - Include the effect of model error
  - Drive mesh adaptivity not through numerical error control (of the mean) but maximization of information

UNI BASEL