

Optimal Experiment Design for State Space Models with Application to Sampled Network Data

George Michailidis

Department of Statistics and EECS
The University of Michigan
www.stat.lsa.umich.edu/~gmichail

Joint work with Harsh Singhal

QPRC 2008

Motivation: Monitoring Network Traffic

- Monitoring flow volumes is important in network management
e.g. capacity planning, root-cause analysis, identification of malicious activity, configuration of routing protocols, etc.
- Due to high volumes and resource constraints, sampling is employed (typical sampling rates range between .001-.01)
- Flows traversing a network can be observed at multiple **observation points** (routers)

Geant Network

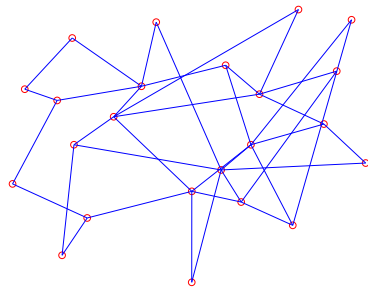
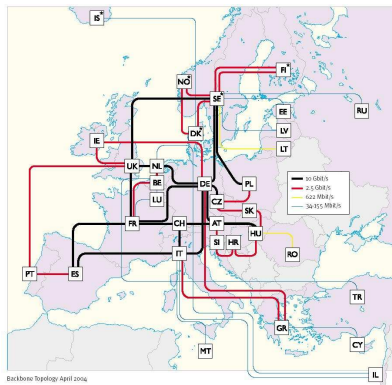


Figure: The network and its logical topology

Introduction

- Objective: Track multiple time-series given **noisy observations**.
- Causal combination of observations is a well defined **filtering** problem.
- We seek to minimize the (running) estimation error through **optimal design of measurement scheme in the filtering context**.
- Optimal experiment design approach to the above problem with network sampled data.

Flow Volumes

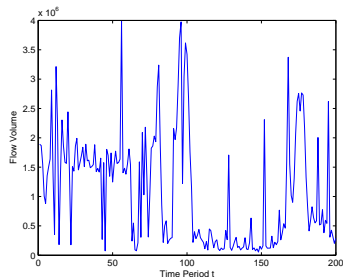
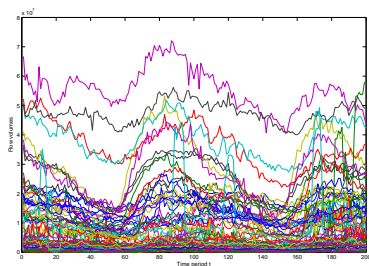


Figure: Left panel: Geant network flows; Right panel: a low volume flow

- All traffic with common origin and destination constitute a flow.
- Flows exhibit complex patterns.

Flow Volumes

- Flows observed at routers in the network.
- Sampling is employed due to high volumes and resource constraints at routers.
- Low sampling rates = **Large sampling noise**.
- Can achieve lower estimation error with same sampling rate through **filtering**.
- **Design sampling scheme**: Take into account measurement noise and process noise.

E-optimal Design for the Simplest Setting

- Assume we have independent observations $y_i \sim N(x_i, 1/m_i)$.
- Estimate $\hat{x} = y$.
- Variance of observation noise **inversely proportional** to design variable.
- Relation between information m and design variables ξ :
 $m = J\xi$.
- Budget constraint on the design variables of the form:
 $R\xi \leq b$.
- E-optimal design problem:

$$\arg \max_{R\xi \leq b} \min_i m_i$$

Minimizing the maximum MSE.

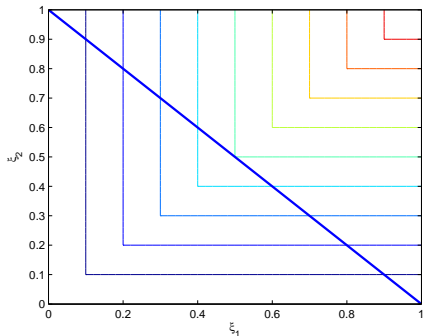


Figure: Contours of objective function for E-optimal design

$$m_1 = 50\xi_1 ; m_2 = 50\xi_2$$

$$\xi_1 + \xi_2 \leq 1$$

Steady State Optimal Design for Random Walks

- Collection of independent random walks:
 $x_i(t) = x_i(t-1) + \epsilon_i(t)$.
- Innovation variances: $\text{Var}(\epsilon_i(t)) = \sigma_i^2$.
- Noisy Observations: $y_i(t) = x_i(t) + \eta_i(t)$.
 $\text{Var}(\eta_i(t)) = 1/m_i$.
- Relation between observed information and design variables. $m = J\xi$.
- Filtering: $\hat{x}_i(t) = \mathbb{E}[x_i(t)|y_i(t), y_i(t-1), \dots]$.

Steady State Optimal Design

- Let $s_i(t) = \text{Var}(\hat{x}_i(t)|y_i(t), y_i(t-1), \dots)$.
- Let $\tilde{m}_i = \lim_{t \rightarrow \infty} 1/s_i(t)$.

Steady State Information given by:

$$\tilde{m}_i = \frac{m_i \sigma_i^2 + \sqrt{m_i^2 \sigma_i^4 + 4m_i \sigma_i^2}}{2\sigma_i^2}$$

- Steady State E-optimal Design

$$\arg \max_{R \xi \leq b} \min_i \tilde{m}_i$$

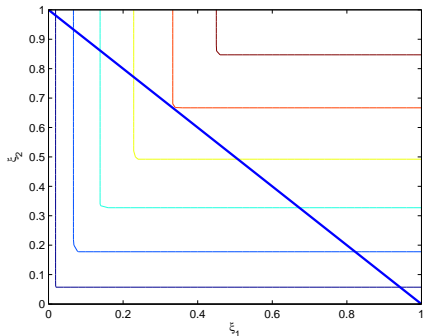


Figure: Contours of objective function for Steady State E-optimal design

$$m_1 = 50\xi_1 ; m_2 = 50\xi_2$$

$$\sigma_1 = 0.1 ; \sigma_2 = 0.2$$

$$\xi_1 + \xi_2 \leq 1$$

Optimization for Steady State E-optimal Design

maximize θ subject to

$$\frac{m_i \sigma_i^2 + \sqrt{m_i^2 \sigma_i^4 + 4m_i \sigma_i^2}}{2\sigma_i^2} \geq \theta$$

$$R\xi \leq b$$

- Constraints can be written as hyperbolic ones.
- Second order conic program: fast interior point methods exist to solve it.

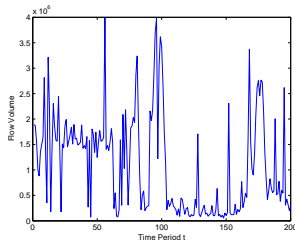
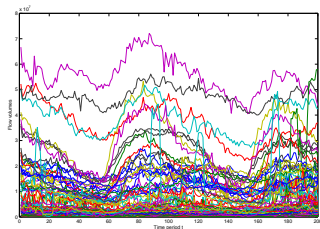
Myopic Approach: A Greedy Alternative to Steady State Optimal Design

- As before $y_i(t) = x_i(t) + \eta_i(t)$. $\text{Var}(\eta_i(t)) = 1/m_i(t)$.
- Time varying design: $m(t) = J\xi(t)$.
- As before $s_i(t) = \text{Var}(\hat{x}_i(t)|y_i(t), y_i(t-1), \dots)$.
- Information at time t : $\tilde{m}_i(t) = 1/s_i(t)$.
- $\tilde{m}_i(t)$ is a function of $\xi(t), \xi(t-1), \dots$.
- Myopic E-optimal design at time t

$$\arg \max_{R\xi(t) \leq b} \min_i \tilde{m}_i(t)$$

- Linear Program.

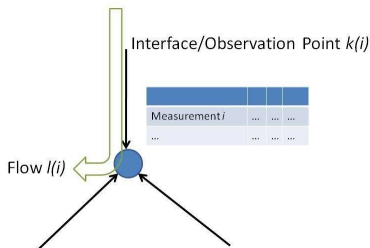
Application: Tracking Flow Volumes



- $x_i(t)$ is the volume of i th flow in time interval t .
- In the Geant network, we have ≈ 300 flows. For illustration purposes, we concentrate on largest 25%.
- **Estimate** from sampled data $y(t)$ is noisy.
- Variance of measurement noise **inversely proportional** to sampling rate ξ .

Constraints for Network Data

- All flows traversing an **observation point** (router interface) experience the same sampling rate.



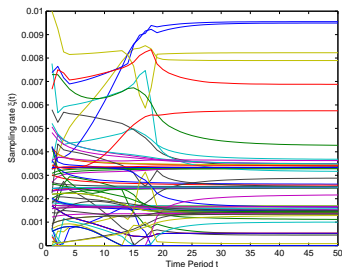
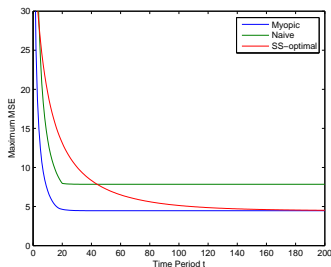
- Each router has multiple interfaces.
- Each flow may traverse multiple observation points; hence,

$$m = J\xi$$

Network Data

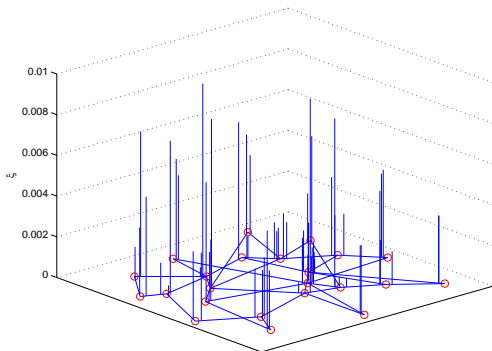
- **Common measurement budget** (CPU time etc.) for multiple observation points.
- Design problem: How to allocate this common budget.
- We assume that the sampling rates are constrained to lie in a convex polygon $R_{\xi_t} \leq b$.
- Important case: The weighted sum of sampling rates on the interfaces of a router is bounded above by the budget for that router. One linear inequality for each router.

Performance of Various Sampling Schemes



As information accumulates over time we get an improvement in performance. We achieve a 42% improvement over **naive** sampling in the steady state. The figure on the right shows that the myopic optimal sampling rates at all observation points reach a steady state.

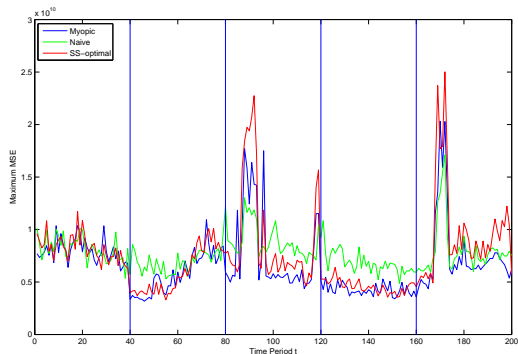
Spatial view of steady state optimal sampling rates



Departures from Linear Model

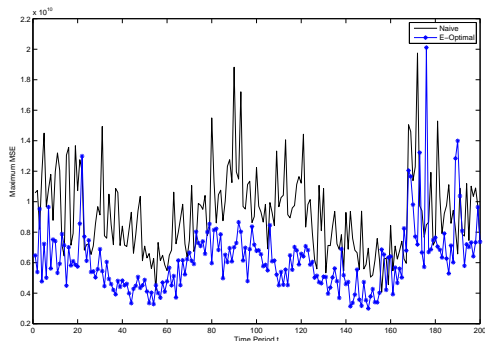
- **Process Model/State Transition Equation:** Not Random Walk. Not Independent.
- Suppose that a flow with volume X in a certain time interval is sampled at a rate ξ .
- If the number of sampled packets is N , then the usual (approximate ML) estimate of flow volume is $Y \equiv N/\xi$.
- **Measurement noise:** $\text{Var}(Y|X) \simeq X/\xi$.
- Thus J in $m = J\xi$ depends on X .
- Solution: Batch Sequential Design.
- At the beginning of each batch use current estimate of X to calculate J .
- **Measurement Model/ Observation Equation:** Local linearization of measurement error.

Performance for Geant Data



Sampling rates **adjusted only at the beginning of a 40 time period block** and constant over each block.

Geant Data



Sampling rates **adjusted at the beginning of each time period** using the myopic scheme. Measurements simulated from Geant data under the true model. Significant improvement in performance over naive allocation.

General Case: State Space Model

- State Transition Equation

$$X(t) = CX(t-1) + w(t)$$

$$\text{Cov}(w(t)) = W.$$

- Observation Equation

$$Z(t) = LX(t) + \epsilon(t)$$

Assume $\text{Cov}(\epsilon(t)) = \Psi(\xi)^{-1}$ where $\Psi(\cdot)$ is a linear function and ξ is the value of design variables. $L'\Psi(\xi)L$ is the corresponding information matrix.

Steady State Optimal Design

- Kalman Filter: Iteratively compute $\mathbb{E}[X(t)|Z(t), Z(t-1), \dots]$.
- Let the steady state estimation error covariance be $\Sigma = \tilde{M}^{-1}$.
- Algebraic Riccati Equation:

$$\tilde{M} = (C\tilde{M}^{-1}C' + W)^{-1} + L'\Psi(\xi)L$$

- No analytic solution in general.
- Steady State Optimal Design:



$$\arg \max_{R\xi \leq b} f(\tilde{M})$$

- Myopic design investigated in Singhal and Michailidis (2008).

Discussion and Future Research Directions

- We have established that steady state E-optimal design for random walks is a second order conic program.
- The A-optimality criterion leads to a tractable linear program.
- We have shown numerically that the performance of the Kalman filter can be significantly improved by optimal experiment design.
- The linear state space model is fairly analytically tractable and one would like to study more closely the steady state optimal sampling rates.
- From a practical point of view it would be useful to extend these ideas to non-linear filtering.

References

-  H. Singhal and G. Michailidis. Optimal sampling in state space models with applications to network monitoring. *Proceedings ACM SIGMETRICS 2008*.
-  H. Singhal and G. Michailidis. Optimal Design in a Filtering Context. *Annals of Applied Statistics*