

# Statistical Mechanics and Thermodynamics of Optimal Estimation

Francis J. Alexander

Los Alamos National Laboratory

LA-UR-08-06894

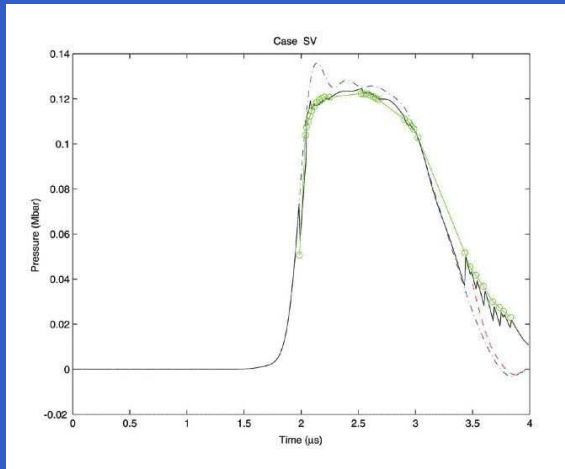
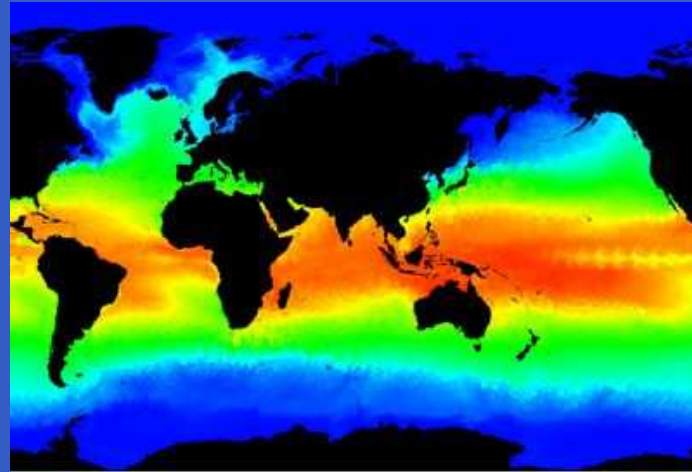
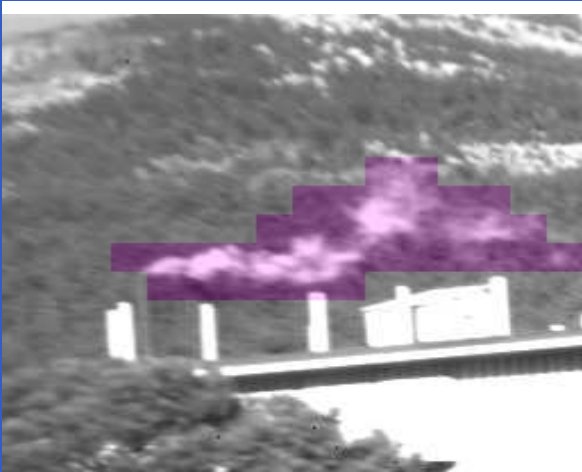
# Collaborators

- Marian Anghel (Los Alamos)
- Gregory Eyink (Johns Hopkins)
- Gregory Johnson\* (Google)
- Natali Gulbahce Johnson\* (Northeastern)
- Yannis Keverekidis (Princeton)
- Sangil Kim (Oregon State)
- Juan Restrepo (Arizona)
- Daniel Tartakovsky (UC San Diego)

# Outline

- Problem Statement
- Formal solution
- Approximate solutions
  - Phase Space Approach
  - State Space Approach
- Conclusion
- Future Directions

# Problem Statement: Optimal Estimation at LANL



# Problem Statement: Dynamics

- (Stochastic) Dynamical System:

$$d\mathbf{X}(dt) = \mathbf{f}(\mathbf{X}, t) + (2\mathbf{D}^{1/2})(\mathbf{X}, t)\mathbf{W}(t)$$

$$\mathbf{Y}(t) = \mathbf{Z}(\mathbf{X}(t), t) + \mathbf{R}^{1/2}(t)\eta(t)$$

- $\mathbf{X}(t)$ : State Vector for  $t_i \leq t \leq t_f$
- $\mathbf{f}(\mathbf{X}(t))$ : Drift Vector;  $\mathbf{D}$ : Diffusion matrix
- $\eta(t)$ : Noise (e.g., White)
- Initial conditions:  $\mathbf{X}_0$  have distribution  $\mathcal{P}_0(\mathbf{X})$ .
- $\mathbf{Y}(t)$ : observation process  $\rightarrow$  data  
 $\mathcal{Y}(t') = \{\mathbf{Y}(t) : t \leq t'\}$
- $\mathbf{R}$  is a Covariance Function
- $\mathbf{h}$  is a Measurement Function

# Problem Statement: Optimal Solution

- Determine the “*best estimate*” of the state conditioned on measurements and the dynamical model
- Conditional mean  $\mathbf{x}_S(t) = E[\mathbf{x}(t) | \mathbf{y}_1, \dots, \mathbf{y}_M]$  is the best estimate of the state
- Conditional covariance matrix  $\mathbf{C}_S(t) = E[(\mathbf{x}(t) - \mathbf{x}_S(t))(\mathbf{x}(t) - \mathbf{x}_S(t))^\top | \mathbf{y}_1, \dots, \mathbf{y}_M]$  measure of its uncertainty.
- Conditional mean  $\mathbf{x}_S(t)$  minimizes  $\text{tr} \mathbf{C}_S(t) = E[|\mathbf{x}(t) - \mathbf{x}_S(t)|^2 | \mathbf{y}_1, \dots, \mathbf{y}_M]$ , variance-minimizing estimator, or *smoother* estimate.

# Formal Solution: Filter

- *General* optimal filtering problem: solved by exactly (formally) by Stratonovich and Kushner within a Bayesian formulation.
- Conditional probability density (filter distribution)

$$\mathcal{P}_F(\mathbf{x}, t) = \mathcal{P}(\mathbf{x}, t | \mathbf{y}_1, \dots, \mathbf{y}_k),$$

given measurements  $\mathbf{y}_1, \dots, \mathbf{y}_k$ , with  $t_{k+1} > t \geq t_k$ .

- Initial condition  $\mathcal{P}_0(\mathbf{x})$  at time  $t_0 < t_1$  and between measurement times,  $\mathcal{P}_F(\mathbf{x}, t)$  solves the forward Kolmogorov equation  $\partial_t \mathcal{P}_F(\mathbf{x}, t) = \hat{L}(t) \mathcal{P}_F(\mathbf{x}, t)$
- $\hat{L}(t) = -\nabla_{\mathbf{x}} \cdot [\mathbf{f}(\mathbf{x}, t)(\cdot)] + \nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^T : [\mathbf{D}(\mathbf{x}, t)(\cdot)]$  is the Fokker-Planck operator.

# Formal Solution: Filter

- At measurement times  $t_m$ ,  $m = 1, \dots, M$ ,  $\mathcal{P}_F(\mathbf{x}, t)$  satisfies the forward “jump condition”  $\mathcal{P}_F(\mathbf{x}, t_m+) = \frac{\exp\left[\mathbf{y}_m^\top \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m) - \frac{1}{2} \mathbf{h}^\top(\mathbf{x}, t_m) \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m)\right]}{\mathcal{W}(\mathbf{y}_1, \dots, \mathbf{y}_m)} \mathcal{P}_F(\mathbf{x}, t_m-)$
- Measurements are used sequentially to obtain the filter distribution  $\mathcal{P}_F(\mathbf{x}, t)$ .
- Calculate moments: moments,  $\mathbf{x}_F(t) = \int d\mathbf{x} \mathbf{x} \mathcal{P}_F(\mathbf{x}, t)$  and  $\mathbf{C}_F(t) = \int d\mathbf{x} (\mathbf{x} - \mathbf{x}_F(t)) (\mathbf{x} - \mathbf{x}_F(t))^\top \mathcal{P}_F(\mathbf{x}, t)$  filter mean and covariance.

# Formal Solution: Smoother

- Optimal smoother distribution  $\mathcal{P}_S(\mathbf{x}, t)$  obtained (Pardoux) by an adjoint algorithm
- Starting from final condition  $\mathcal{A}_S(\mathbf{x}, t_f) = 1$   $t_f > t_M$  solve backward Kolmogorov equation
- $\partial_t \mathcal{A}_S(\mathbf{x}, t) + \hat{L}^*(t) \mathcal{A}_S(\mathbf{x}, t) = 0$
- $\hat{L}^*(t) = \mathbf{f}(\mathbf{x}, t) \cdot \nabla_{\mathbf{x}} + \mathbf{D}(\mathbf{x}, t) : \nabla_{\mathbf{x}} \nabla_{\mathbf{x}}^\top$  is the adjoint Fokker-Planck operator.

# Formal: Smoother

- Backward “jump condition”  $\mathcal{A}_S(\mathbf{x}, t_m -) = \mathcal{A}_S(\mathbf{x}, t_m +) \frac{\exp\left[\mathbf{y}_m^\top \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m) - \frac{1}{2} \mathbf{h}^\top(\mathbf{x}, t_m) \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m)\right]}{\mathcal{W}(\mathbf{y}_1, \dots, \mathbf{y}_m)}$
- $\mathcal{W}(\mathbf{y}_1, \dots, \mathbf{y}_m)$  normalization factor
- $\mathcal{P}_S(\mathbf{x}, t) = \mathcal{A}_S(\mathbf{x}, t) \mathcal{P}_F(\mathbf{x}, t)$
- $\mathcal{P}_S(\mathbf{x}, t)$  is continuous in time.
- $\mathbf{x}_S(t) = \int d\mathbf{x} \mathbf{x} \mathcal{P}_S(\mathbf{x}, t)$  and  $\mathbf{C}_S(t) = \int d\mathbf{x} (\mathbf{x} - \mathbf{x}_S(t)) (\mathbf{x} - \mathbf{x}_S(t))^\top \mathcal{P}_S(\mathbf{x}, t)$  give the smoother mean and covariance.

# Formal Solution: Special Cases

- Linear dynamics
  - Gaussian additive noise
  - Gaussian errors
  - Kalman-Bucy filter / smoother
- Weakly nonlinear dynamics
  - Extended Kalman filter / smoother
  - Problems with this approach
  - Multimodal statistics, Riccati Equation
- What if these don't apply?

# Approximate Solution: New Approaches

- Closure of KS/P equations
- Rayleigh-Ritz Mean-Field Variational method
- Path Integral Monte Carlo Methods
  - Cluster Methods\*
  - Hybrid Monte Carlo (and Generalized)
  - Adjoint Methods

# Path Integral Methods: Idea

- Cast stochastic optimal estimation of time series in path integral form
- Apply analytical and computational techniques of equilibrium statistical mechanics
- Use standard or accelerated Monte Carlo methods for smoothing, filtering and/or prediction
- Cluster Algorithms based on Fortuin-Kasteleyn Representation

# Path Integral Methods: Dynamics

- $d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), t)dt + (2\mathbf{D})^{1/2}(\mathbf{x}, t)\mathbf{W}(t), \quad t > t_0$
- For the initial state  $\mathbf{x}_0$ , either its value or probability distribution is assumed.
- Discretize stochastic dynamics using an explicit Euler-Maruyama scheme.
- $\mathbf{x}_{k+1} = \mathbf{x}_k + f(\mathbf{x}_k, t_k)\delta t + (2\delta t\mathbf{D})^{1/2}(\mathbf{x}_k, t_k)(\mathbf{W}(t_k + \delta t) - \mathbf{W}(t_k))$
- where  $k = 0, 1, 2, \dots$
- The choice of the time discretization is not unique
- Optimal (or good choice) is an open research problem

# Path Integral Methods: Noise

- The probability of the dynamics generating a given history is simply related to the probability that it experiences a specific noise history
- $\eta(t_k) = \mathbf{W}(t_k + \delta t) - \mathbf{W}(t_k)$
- at times  $t_k$ ;  $k = 0, 1, 2, \dots, M$
- For Gaussian, uncorrelated noise  
 $\text{Prob}\{\eta(t), t = t_0, t_1, \dots\} \sim \exp(-\frac{1}{2} \sum_k \eta^2(t_k)).$

# Path Integral Methods: Measurements

- Without loss of generality assume the discrete time steps to be equally spaced
- Assume measurement times are commensurate with  $\delta t$ , the time step interval. Namely, we define  $t_k = t_0 + k\delta t$  with  $k = 0, 1, \dots, T$ , and  $(t_f - t_0)/T = \delta t$ .
- Assume that the errors are Gaussian and uncorrelated with each other and uncorrelated with the state of the system (can handle more general error statistics)

# Path Integral Methods: Action (appended)

- Thus far only effects of the dynamics considered
- Include the influence of observations via Bayes' rule

- Modifies Hamiltonian, adding

$$H_{obs} = \sum_{m=1}^M [\mathbf{h}(\mathbf{x}(t_m)) - \mathbf{y}(t_m)]^\top R^{-1} [\mathbf{h}(\mathbf{x}(t_m)) - \mathbf{y}(t_m)]$$

- Corresponds to a local field or pinning term which, when the measurements are accurate, forces the state variable to be close that of the observation.
- Total Hamiltonian:

$$H = H_{dynamics} + H_{obs}$$

# Path Integral Methods: Interpretation

- PI assigns weights/probabilities to histories.
- Weights depend on both the stochastic dynamics and the measurements.
  - Histories unlikely to arise from the dynamics are given a lower weight than histories which are consistent with them
  - Histories far from the measurements are given lower weight than those closer to the measurements
- Competition between the noise in the dynamics and the errors in the measurements
- Now sample this distribution!

# Potential Applications

- Configuration of a protein evolving under Brownian dynamics
- Concentration of interacting metabolites
- Positions of atoms in a crystal undergoing a structural phase transition or nucleation, or the
- State of a queue in a stochastic fluid model
- The final state can also be a rare event on which the history is conditioned

# Path Integral Methods: Relation to Max Likelihood

- Rather than sampling the Gibbs distribution, maximum likelihood methods minimize the Hamiltonian
- The Hamiltonian is the log-likelihood
- Maximum likelihood methods determine the mode of a distribution – path-integral methods can be used to determine the mean and other moment statistics
- Close relationship between PIMC and variational/adjoint in ocean/climate community
- Spacetime Hamiltonian identical to standard cost function for weakly constrained 4D-var

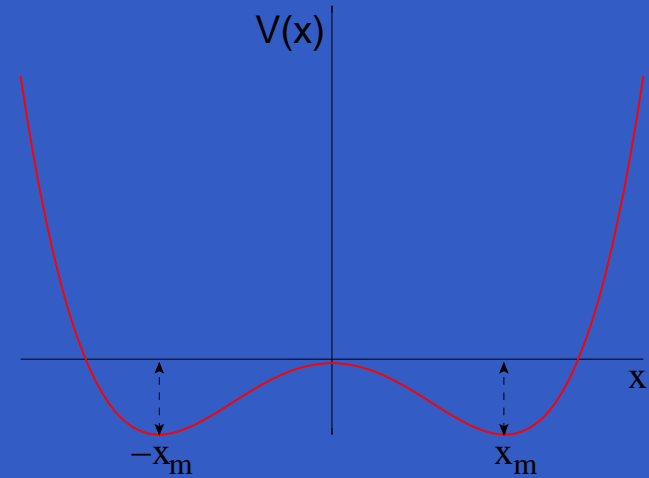
# Path Integral Methods: Sampling

- Known disadvantage of MCMC-based methods is slow convergence.
- Search for MCMC based strategy with significant increase in computational efficiency
- Unigrid, Multigrid, Fourier Acceleration, Hybrid Monte Carlo, *Cluster Methods*

# Langevin Equation

$$\phi(\mathbf{r}, t + \Delta t) = \phi(\mathbf{r}, t) + \frac{\Delta t}{\Delta x^2} \left[ \sum_i \phi(\mathbf{r}_i, t) - 4\phi(\mathbf{r}, t) \right]$$

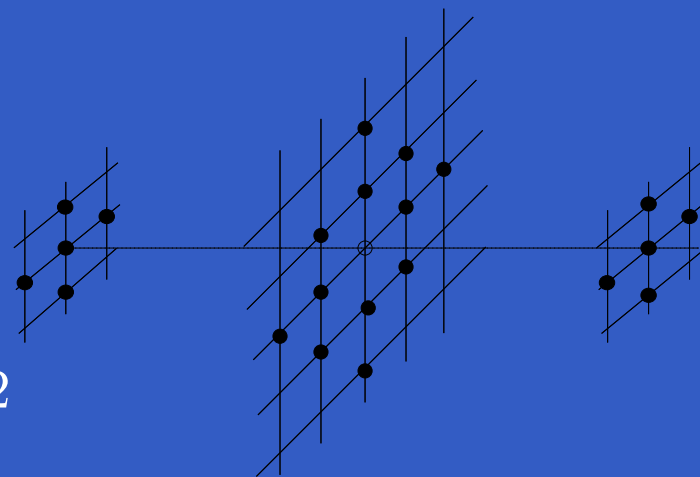
$$+ \Delta t [-a\phi(\mathbf{r}, t)^3 + b\phi(\mathbf{r}, t)] + \sqrt{\Delta t} \eta(\mathbf{r}, t)$$



# Action

$$S \equiv \frac{1}{4D\Delta t} \sum_{\mathbf{r}, t} \left( \phi(\mathbf{r}, t + \Delta t) - \phi(\mathbf{r}, t) - \right.$$

$$\Delta t \left[ -a\phi^3(\mathbf{r}, t) + b\phi(\mathbf{r}, t) \right]$$



$$-\Delta t \left[ \sum_i \phi(\mathbf{r}_i, t) - 4\phi(\mathbf{r}, t) \right]^2$$

# Cluster Monte Carlo

- Percolation-based cluster Monte Carlo - sample the statistical ‘mechanics of histories for nonlinear stochastic processes
- Application to rare event simulations and optimal estimation
- Improves statistical sampling of histories in Monte Carlo simulations significantly
- Traditional spatial cluster algorithms: the clusters represent statistically independent objects at a given time
- In the  $d + 1$  dimensions the clusters can be interpreted as statistically independent objects in space-time.

# Cluster Dynamics

- Brower and Tamayo extended Swendsen-Wang algorithm to a continuous field theory
- Embed discrete variables (spins) into field
- $V(\mathbf{r}, t) = (a/4)\phi^4(\mathbf{r}, t) - (b/2)\phi^2(\mathbf{r}, t)$
- The discrete spin variables,  $s_r$ , label the two wells in  $\phi^4$  potential such that  $\phi_r = s_r |\phi_r|$ .
- At fixed values of  $|\phi(\mathbf{r})|$  a ferromagnetic Ising model is embedded into the  $\phi^4$  field theory

# Embedded Dynamics

- Update  $\phi_r$  via a standard local Monte Carlo algorithm.
- Form percolation clusters dictated by the bond probability,  $p_{rr'} = 1 - e^{-\beta_{rr'}(1+s_r s'_r)} = 1 - e^{-(|\phi_r||\phi'_r| + \phi_r \phi'_r)}$
- Effective spin-spin coupling is  $\beta_{rr'} = |\phi_r \phi'_r|$ .
  - Note that  $p_{rr'}$  reduces to  $1 - \exp(-2\beta_{rr'})$  when the spins are the same sign.
  - Update the Ising variables:
  - Flip percolation clusters independently with probability  $1/2$ .
  - If the move is accepted, flip the sign of the fields in the cluster.

# Space-Time Hamiltonian

Local terms:(e.g.  $\phi(\mathbf{r}, t)^2$ ),  $(\mathbf{r}_j, t_k)$  as the reference site)

Nearest neighbors of  $(\mathbf{r}_j, t_{k-1})$  coupled to  $(\mathbf{r}_j, t_k)$ :

$$\beta_1 = 2\Delta t \left( \sum_i \phi(\mathbf{r}_i, t_{k-1}) \right) \phi(\mathbf{r}_j, t_k)$$

Site  $(\mathbf{r}_j, t_{k-1})$  coupled to  $(\mathbf{r}_j, t_k)$ :

$$\beta_2 = \left[ (2b-8)\Delta t - 2a\Delta t \phi^2(\mathbf{r}_j, t_{k-1}) + 2 \right] \phi(\mathbf{r}_j, t_k) \phi(\mathbf{r}_j, t_{k-1})$$

Nearest neighbors of  $(\mathbf{r}_j, t_k)$  coupled to each other:

$$\beta_3 = -\Delta t^2 \left( \sum_i \phi(\mathbf{r}_i, t_k) \right) \left( \sum_i \phi(\mathbf{r}_i, t_k) \right)$$

Nearest neighbors of  $(\mathbf{r}_j, t_k)$  coupled to  $(\mathbf{r}_j, t_k)$ :

$$\beta_4 = \left( \sum_i \phi(\mathbf{r}_i, t_k) \right) \left( [(8-2b)\Delta t^2 - 2\Delta t] \phi(\mathbf{r}_j, t_k) + 2a\Delta t^2 \phi^3(\mathbf{r}_j, t_k) \right)$$

## Now with measurements

The probability of a site having a bond with any of its neighbors is  $P_i = 1 - e^{-2\beta_i/(4D\Delta t)}$

Probability of flipping a cluster in presence of measurements.

$$p_{\text{flip}} = \frac{e^{\sum_m -2\phi(\mathbf{r},t)\phi_m(\mathbf{r},t)}}{e^{\sum_m 2|\phi(\mathbf{r},t)|\phi_m(\mathbf{r},t)} + e^{\sum_m -2|\phi(\mathbf{r},t)|\phi_m(\mathbf{r},t)}}$$

# Correlation Times

Correlation times of the magnetization  $M$  for local and cluster algorithms for several noise strengths,  $D$ . The system dimensions are  $L = 10$  and  $T = 100$ , the acceptance ratio,  $a \approx 0.5$ ,  $\Delta t = 0.05$  and  $\Delta x = 1.0$ .

$D$	$\tau_{local}$	$\tau_{cluster}$
1	947	775
5	180	134
15	25	8.8
20	19	2.9
25	12	1.4
30	9	1.1

# Correlation Times

Correlation times of  $M$  for local and cluster algorithms with measurements at different system sizes at noise strength,  $D = 25$ . The cluster algorithm consistently outperforms the local one.

$L$	$T$	$\tau_{local}$	$\tau_{cluster}$
8	32	10.8	1.7
12	72	11.3	1.6
16	128	11.9	1.5
24	288	12.3	1.7

# Conclusions

- Class of stochastic processes covered by this method
  - finite-dimensional approximations to stochastic partial differential equations, maps, coupled discrete event systems
- Can handle non-Markovian processes
- Can be applied to study pathways to rare events as well as for optimal state and parameter estimation.
- When the cluster size distribution scales, the cluster algorithm outperforms the local Monte Carlo
- Clusters are statistically independent space-time events, temporal (time-axis) extent of these clusters estimates lifetime

# Current/Future Directions

- Adjoint Methods
- Phase Transitions / Critical Phenomena
- Connection between Mean Field Estimator and Belief Propagation