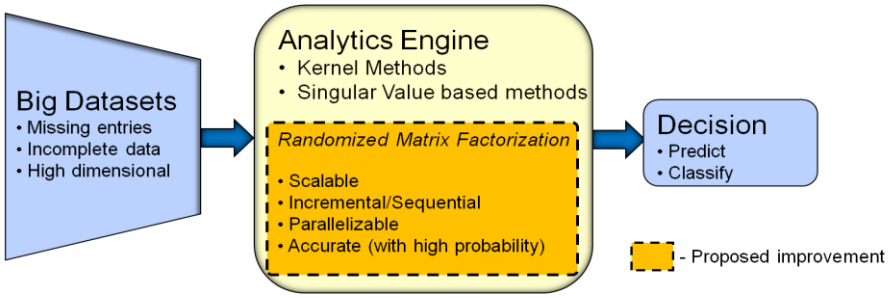
Randomized Algorithms for Data Analytics

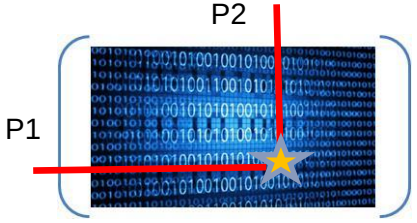
Dr. Shaunak D. Bopardikar
Staff Research Scientist
Controls Group
United Technologies Research Center
East Hartford, CT USA

UTRC Collaborators: George Ekladious, Kunal Srivastava
Academic Collaborators: Joel Tropp (CalTech), C. Langbort (UIUC)

Randomized Algorithms for Scalable Analytics

Funded by: Office of Naval Research (PI: Shaunak D. Bopardikar) 2016-19, UTRC Innovation funds





Min-max problems (Game)

- Randomized algorithms to explore scalability versus accuracy tradeoffs
  - Projection-based (matrix factorization) methods for analytics

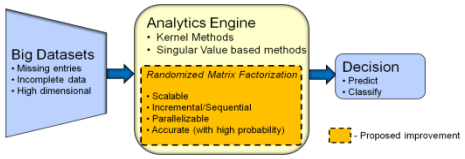
    [Ref: Bopardikar et al ACC 2013, IEEE BigData, ACC 2017 (to appear)]
  - Sampling-based (scenario optimization) for min-max problems / games

    [Ref: Bopardikar et al Automatica 2013, CDC 2014, Automatica 2016]

- Analytic guarantees on accuracy and estimates on computational complexity
- Results hold with “high probability” (which is also a tuning/design parameter)

Projection-based (Matrix Factorization) Approaches

- Applications
  - Streaming Gaussian Processes (ONR funded effort)



Given training data: $\boldsymbol{X}_{n\times d}, \boldsymbol{Y}_{n\times 1}$, kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$

Estimate output y at a new (test) point x_*

Closed form:

$$f_*(\mathbf{x}_*) = \mathbf{k}_{\mathbf{x}_*}^\top (k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I})^{-1}\mathbf{Y},$$
$$k_*(\mathbf{x}_*,\mathbf{x}_*) = k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_{\mathbf{x}_*}^\top (k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I})^{-1}\mathbf{k}_{\mathbf{x}_*}$$

$O(n^2)$: batch setting

IEEE Big Data 2016

  - Kalman filtering in high dimensions

$$x_{t+1} = A_t x_t + w_t,$$
$$y_t = C_t x_t + v_t,$$

$$\hat{x}_{t+1} = A_t\hat{x}_t + K_t(y_t - C_t\hat{x}_t),$$
$$P_{t+1} = (I - K_t C_t)(A_t P_t A_t^T + Q_t),$$
$$K_t = (A_t P_t A_t^T + Q_t)C_t^T(C_t(A_t P_t A_t^T + Q_t)C_t^T + R_t)^{-1}$$

$O(n^3)$

ACC 2017 (to appear)

- Main idea: store data matrix (A) in factored form
  - Efficiently update/re-compute with new data

$$\underbrace{B_{t-1}, C_{t-1}}_{\text{Previous factors}} \longrightarrow \begin{bmatrix} B_{t-1}\, C_{t-1} & A_{\text{col}} \\ A_{\text{row}} & A_{\text{row,col}} \end{bmatrix} \xrightarrow{\text{Randomized Range Approx}} B_t, C_t$$

Ref: Bopardikar et al 2013

  - *New algorithms with complexity $O(n)$ with high probability error bounds*