Submodular Optimization for Discovery of Key Entities in Complex Systems

ARUN V. SATHANUR

Pacific Northwest National Laboratory
CCSI Workshop, ACC, 2017, Seattle, WA
Outline

- Project Objectives

- Background: Submodular optimization

- Spatial system: Sensor placement
  - Scalable algorithms for sensor placement in large spatial domains

- Networked system: Influence Maximization
  - Influence Maximization with Intrinsic Nodal Activation

- Looking forward
Objectives and Motivation

- Development of scalable submodular optimization approaches to address observability and controllability in large complex systems via discovery of key entities

Motivation scenarios:

- Placement of sensors across a given large geographical area
- Placement of PMUs in a power network for maximizing observability
- Joint observability-controllability optimization to localize the spread of power outage via defensive islanding
- Observing nodes on social media to monitor opinions on a topic and identifying influential nodes to incentivize towards adoption of a specific behavior.
Submodular Functions

Consider a set of entities $V$. Let $S \subseteq V$. Let $f : 2^V \rightarrow \mathcal{R}$, be a mapping that associates every such $S$ with a real number. $f$ is submodular provided it satisfies the following.

For every $S \subseteq T \subseteq V$ and any $v \in V, v \notin T$ we have

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

A function $f : 2^V \rightarrow \mathcal{R}$ is monotone provided, for any $S \subseteq T$, $f(S) \leq f(T)$

We want to select the optimal subset called the seed set $S$ such that $f(S)$ is maximized under a constraint on the cardinality of the seed set $S$ i.e $|S| \leq k$.

For this problem the simple Greedy algorithm works very well

Let $S^*$ denote the optimal solution to the problem. If $f$ is monotone submodular and $f(\phi) = 0$, then due to a powerful result derived by Nemhauser et al.

$$f(S) \geq \left(1 - \frac{1}{e}\right) f(S^*)$$
Scalability: Two Pieces to Address

Scale the greedy algorithm. Ordinary Greedy Algorithm: \( O(kn) \) oracle calls

Scale the computation oracle.
Oracle Complexity: Varies between applications.
Example: Sensor Placement: \( O(n^3) \)
General Sensor Placement

- Assume that the quantity to be sensed can be modeled as a Gaussian Process with a given covariance function

- Given a set of ‘N’ possible sites distributed either spatially or over a network and given that the NxN co-variance matrix can be computed

- Given a budget of ‘k’ sensors, find the sites where the sensors need to be placed in order to maximize the observability

- Metric for the optimization: The metric that we consider is the mutual information between the sensed and the un-sensed locations

- Given a set of locations where sensors are placed, the marginal gain in MI by adding a new element y to the set A is given by

\[
\frac{\sigma_{yA}^2 - \Sigma_y \Sigma_A^{-1} \Sigma_{Ay}}{\sigma_{yA}^2 - \Sigma_y \Sigma_A^{-1} \Sigma_{A\bar{A}} \Sigma_{\bar{A}y}}
\]
Initial Results

Scaling
- Ordinary Greedy: 3.6
- Stochastic Greedy: 1.2
- Covariance Build: 2.0

Quality
Network Scenario: Influence Maximization Problem

- A large class of natural and man-made systems with rich dynamics can be studied through the abstraction of graphs.

- Signals arrive at each node along the incoming edges, undergo (non-linear) processing at the node and the processed signal is transmitted along the out-going edges.

- Given the models for node behavior and edge interactions and the objectives we are interested in, how can we find those influential nodes which have maximal impact on the system?

- Applications in diverse fields
  - Viral marketing for product adoption
  - Spread of content on social media
  - Spread of diseases in contact-networks
  - Keystone species in microbial communities
  - Controllability and Observability in complex systems
Our Contribution: Inclusion of Intrinsic Activation

Activation at each node is split into two mechanisms to allow correspondences to real-world situations

Intrinsic activation: Activation at each node attributable to its own intrinsic mechanisms

Influenced activation: Activation originating at each node attributable to influence of the neighboring nodes

Parameterize by the tendency towards intrinsic activation denoted by $\alpha$ and that towards influenced activation by $\beta$

Experiments on the Twitter network

Independent Cascade model over-estimates the number of activations

Ran the algorithm with three different alpha ranges and finally the scenario where the alpha values were set to be proportional to the node out-degree. The last one maximizes engagement.
Looking forward

- Distributed submodular optimization algorithms for even larger problems

- Online submodular optimization for real-time applications

- Working with domain experts to apply the algorithms to power grid applications at scale:
  - PMU placement
  - Controlled Islanding
  - Synchronization

- Scalable algorithms for influence maximization
Thank You