Empowering Nonlinear and Stochastic Optimization for Large-scale Data Analysis

Guanghui (George) Lan

Assistant Professor, Ph.D.
Dept. of Industrial and Systems Engineering
University of Florida
June 19th, 2013

Joined UF since August 2009.

Research Areas

- Methodology: stochastic and nonlinear optimization.
- Applications in large-scale data analysis: machine learning, image processing and simulation input/output analysis etc.
- Goal: transform raw data into useful knowledge to support decision-making, e.g., in healthcare, national security, energy and transportation etc.

The Role of Optimization

Since its beginning, nonlinear and stochastic optimization has been recognized as an important modeling and solution technique in data analysis. Application examples include

- Linear regression: $\min \mathbb{E}_{u,v}\left[(u^T x - v)^2\right]$.
- Maximum likelihood estimation: $\max \mathbb{E}_u\left[\log f(p, u)\right]$.
- Support vector machine: $\min \mathbb{E}_{u,v}\left[\max\{0, v\langle x, u\rangle\right] + \rho\|x\|_2^2$.
- Compressed sensing: $\min_x \|Ax - b\|^2 + \lambda\|x\|_1$.
- Total variation minimization: $\min_x \|Ax - b\|^2 + \lambda \mathrm{TV}(x)$.
- Matrix completion: $\min_x \|Ax - b\|^2 + \lambda \sum_i \sigma_i(x)$.

If the dataset is relatively small, routinely solved by the off-the-shelf solvers, e.g., those based on second-order interior point methods.

Big-data Challenges in Optimization

Examples:

- Netflix problem: Rows - ratings from Customer; Columns - movies. Dataset: 100 million ratings from over 480 thousand customers on nearly 18 thousand movie titles.
- Machine learning: the largest dataset in UCI (University of California, Irvine) repository prior to 1990 had about $8,000$ samples, while the largest dataset currently in the repository has 8 million samples.

Challenges:

- High dimensionality (the number of unknowns, 10^4 to 10^{12}).
- Uncertainty (dataset: samples from unknown distribution).
- Structural ambiguity: smoothness, regularity and convexity.
- Increasing need to solve the problem in real time.

Recent Advances

Scalable, robust and efficient optimization algorithms, along with strong sampling and iteration complexity results.

- Stochastic Optimization: Robust stochastic approximation (SA) by Nemirovski, Juditsky, Lan and Shapiro, 09; Accelerated SA by Lan (10), stochastic first- and zeroth-order methods by Ghadimi and Lan (12).
- Deterministic Optimization: Nesterov's optimal method and smoothing technique (Nesterov, 83, 05), uniformly optimal prox-level methods (Lan 11, 13) and universal gradient method (Nesterov 13).
- Block decomposition and parallel computing (Shalev-Shwartz and Tewari 11, Nesterov 12, Dang and Lan 13).

Impact: ≈ 240 citations to the 2009 paper. Recognized by various prizes/awards from INFORMS, MOS and NSF.

Research Grants in Data Analysis

NSF Operations Research and Computational Mathematics:

- National Science Foundation (CMMI-1000347), Theory and Applications of Stochastic First-order Methods for Large-Scale Stochastic Convex Optimization, May 2010 - April 2014.
- National Science Foundation (CMMI-1254446), CAREER: Reduced-order Methods for **Big-Data Challenges** in Nonlinear and Stochastic Optimization, Jan 2013 - Dec. 2017.
- National Science Foundation (DMS-1319050), Accelerated Algorithms for a Class of Saddle Point problems and Variational Inequalities, Sep. 2013 - Aug. 2016, recommended for funding (with Yunmei Chen).

Limits and Opportunities at UF

- Gaps between between theoretical and applied research.
  - Workshops, new courses, stronger student recruitment and training programs, and joint faculty appointments?
- Facilitating the formation of **big big-data** research groups
  - DOE and NIH big-data opportunities.
- Educational programs in data analytics.
  - e.g., Columbia and Northwestern.
- Support from the state government and local industry.

Thanks!

- Email: glan@ise.ufl.edu.
- Phone: 352-392-1464 ext. 2005.