Fast Feasibility Pursuit for Nonconvex QCQP using First-Order Methods

Aritra Konar

Dept. of ECE, UMN

Advisor: Prof. Nikos D. Sidiropoulos

March 8, 2017

Nonconvex QCQPs

- General Form:

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathcal{X}}. \quad & \mathbf{x}^T\mathbf{Q}\mathbf{x} \\
\text{s.t.} \quad & \mathbf{x}^T\mathbf{A}_m\mathbf{x} \leq b_m, \ \forall\, m \in \mathcal{M}_{\mathcal{I}} \\
& \mathbf{x}^T\mathbf{C}_m\mathbf{x} = d_m, \ \forall\, m \in \mathcal{M}_{\mathcal{E}}
\end{aligned}
$$

- NP-Hard (in general)
- Ubiquitous in wireless communications, signal processing, power systems etc.
  - Multicast beamforming [Sidiropoulos *et al*. 2006]
  - Phase Retrieval [Fienup 1978]
  - Optimal Power Flow [Carpentier 1962]
  - Power System State Estimation [Schweppe *et al.* 1970]

Nonconvex QCQPs

- Existing approaches
  - Semidefinite Relaxation [Wolkowicz 2000, Luo *et al.* 2010]
    - Solve rank relaxed SDP and use post-processing step (deterministic or randomized) to generate feasible solution; fails in most instances
  - Successive Convex Approximation [Beck *et al.* 2010, Scutari *et al.* 2014]
    - Approximate problem via sequence of convex problems; guaranteed convergence to stationary points
    - Requires feasible point for initialization; non-trivial to determine
  - Feasible Point Pursuit [Mehanna *et al.* 2015, Kanatsoulis *et al.* 2015]
    - Use SCA + slack variables to approximate feasibility problem
    - Works with any choice of initialization; empirically performs very well
  - Consensus ADMM [Huang *et al.* 2016]
    - Decompose problem into multiple parallel QCQP-1 subproblems at every iteration; QCQP-1 is optimally solvable
    - Enforce consensus among solutions to determine global variable

Nonconvex QCQPs

- Drawbacks
  - FPP-SCA and C-ADMM require computing eigendecompositions; additionally FPP-SCA requires storing the positive and negative definite parts in memory
  - FPP-SCA requires solving a conic programming problem at every iteration incurring complexity $\mathcal{O}(M+N)^{3.5}$
  - C-ADMM is very memory intensive, one local variable created for every constraint
- Computationally demanding/memory intensive
  - Cannot be applied to large-scale problems

- We propose a FOM based approach for feasibility pursuit with low computational and memory requirements
  - Works well in practice

Problem Statement

- Exact Penalty Formulation

$$\min_{\mathbf{x}\in\mathcal{X}}\left\{F^{(ns)}(\mathbf{x}) := \sum_{m=1}^{M_I} \max\{\mathbf{x}^T\mathbf{A}_m\mathbf{x} - b_m, 0\} + \sum_{m=1}^{M_E} |\mathbf{x}^T\mathbf{C}_m\mathbf{x} - d_m|\right\}$$

- Equivalently, in smooth form

$$\begin{aligned}
\min_{\substack{\mathbf{x}\in\mathcal{X},\, \mathbf{s}_{\mathcal{I}}\in\mathbb{R}^{M_I},\\ \mathbf{s}_{\mathcal{E}}\in\mathbb{R}^{M_E}}} \quad & \sum_{m=1}^{M_I} s_{\mathcal{I}}(m) + \sum_{m=1}^{M_E} s_{\mathcal{E}}(m) \\
\text{s.t.} \quad & \mathbf{x}^T\mathbf{A}_m\mathbf{x} - b_m \le s_{\mathcal{I}}(m),\ s_{\mathcal{I}}(m) \ge 0,\ \forall\, m \in \mathcal{M}_{\mathcal{I}} \\
& s_{\mathcal{E}}(m) \le \mathbf{x}^T\mathbf{C}_m\mathbf{x} - d_m \le s_{\mathcal{E}}(m),\ \forall\, m \in \mathcal{M}_{\mathcal{E}}
\end{aligned}$$

- FPP-SCA corresponds to performing SCA on above problem

- Use FOMs on original formulation instead?
  - Non-differentiable!

Problem Formulation

- Inequality constraints:
  - Define $f_m(\mathbf{x}) = \max\{\mathbf{x}^T \mathbf{A}_m \mathbf{x} - b_m, 0\} = \max_{0 \leq y \leq 1} \{y(\mathbf{x}^T \mathbf{A}_m \mathbf{x} - b_m)\}, \forall\, m \in \mathcal{M}_{\mathcal{I}}$
  - Smooth surrogate: [Nesterov 2004]

$$
\begin{aligned}
f_m^{(\mu)}(\mathbf{x}) &= \max_{0 \leq y \leq 1} \{y(\mathbf{x}^T \mathbf{A}_m \mathbf{x} - b_m) - \mu \frac{y^2}{2}\}, \forall\, m \in \mathcal{M}_{\mathcal{I}} \\
&= \begin{cases} 0, & \text{if } \mathbf{x}^T \mathbf{A}_m \mathbf{x} \leq b_m \\ \frac{(\mathbf{x}^T \mathbf{A}_m \mathbf{x} - b_m)^2}{2\mu}, & \text{if } b_m < \mathbf{x}^T \mathbf{A}_m \mathbf{x} \leq b_m + \mu \\ \mathbf{x}^T \mathbf{A}_m \mathbf{x} - b_m - \frac{\mu}{2}, & \text{if } \mathbf{x}^T \mathbf{A}_m \mathbf{x} > b_m + \mu \end{cases}
\end{aligned}
$$

  - Quality of approximation: [Nesterov 2004]

$$
f_m^{(\mu)}(\mathbf{x}) \leq f_m(\mathbf{x}) \leq f_m^{(\mu)}(\mathbf{x}) + \frac{\mu}{2}, \forall\, \mathbf{x} \in \mathbb{R}^N, \forall\, m \in \mathcal{M}_{\mathcal{I}}
$$

- Equality constraints:
  - Define $g_m^{(q)}(\mathbf{x}) := (\mathbf{x}^T \mathbf{C}_m \mathbf{x} - d_m)^2, \forall\, m \in \mathcal{M}_{\mathcal{E}}$

- Overall formulation: $\min_{\mathbf{x} \in \mathcal{X}} \left\{ F^{(s)}(\mathbf{x}) := \frac{1}{M} \left(\sum_{m=1}^{M_I} f_m^{(\mu)}(\mathbf{x}) + \sum_{m=1}^{M_E} g_m^{(q)}(\mathbf{x}) \right) \right\}$ $\quad (M := M_I + M_E)$

Overview of FOMs

- Minimizing average of finite sums via FOMs:

$$\min_{\mathbf{x}\in\mathcal{X}}\left\{F(\mathbf{x}):=\frac{1}{M}\sum_{m=1}^{M}f_m(\mathbf{x})\right\}$$

- Gradient Descent (GD): [Cauchy 1847]

$$\mathbf{x}^{(k)}=\Pi_{\mathcal{X}}\left(\mathbf{x}^{(k-1)}-\frac{\alpha_k}{M}\sum_{m=1}^{M}\nabla f_m(\mathbf{x}^{(k-1)})\right),\forall\, k\in\mathbb{N}$$

- Stochastic Gradient Descent (SGD): [Robbins and Munro 1953]
  - Sample $m_k\in[M]$ uniformly at random (with replacement)

$$\mathbf{x}^{(k)}=\Pi_{\mathcal{X}}\left(\mathbf{x}^{(k-1)}-\alpha_k\nabla f_{m_k}(\mathbf{x}^{(k-1)})\right),\forall\, k\in\mathbb{N}$$

- Stochastic Variance Reduced Gradient (SVRG): [Johnson *et al.* 2014]
  - Define stage s and inner stochastic iterations

$$\mathbf{x}_s^{(k)}=\Pi_{\mathcal{X}}\left(\mathbf{x}_s^{(k-1)}-\alpha_s^{(k)}(\nabla f_{m_k}(\mathbf{x}_s^{(k-1)})-\nabla f_{m_k}(\mathbf{y}_s)+\nabla F(\mathbf{y}_s))\right),\forall\, k\in[K],\forall\, s\in\mathbb{N}$$

Convergence results for FOMs

- Convergence to stationary points
  - Assumption: Lipschitz continuity of $F(\mathbf{x})$ and $\nabla F(\mathbf{x})$
    - GD [Nesterov 2004, Ghadimi *et al.* 2016]
    - SGD [Ghadimi and Lan 2013]
    - SVRG [Reddi *et al.* 2016]
- Convergence to local minima
  - Assumption: $F(\mathbf{x})$ satisfies the strict-saddle property [Ge *et al.* 2015]
    - GD (w/ random initialization) [Lee *et al.* 2016]
    - SGD [Ge *et al.* 2015]
- Convergence to global minima (at linear rate!)
  - Assumption: $F(\mathbf{x})$ satisfies the Polyak-Lojasiewicz (PL) inequality
    - GD and SGD [Karimi *et al.* 2016]

For our problem……

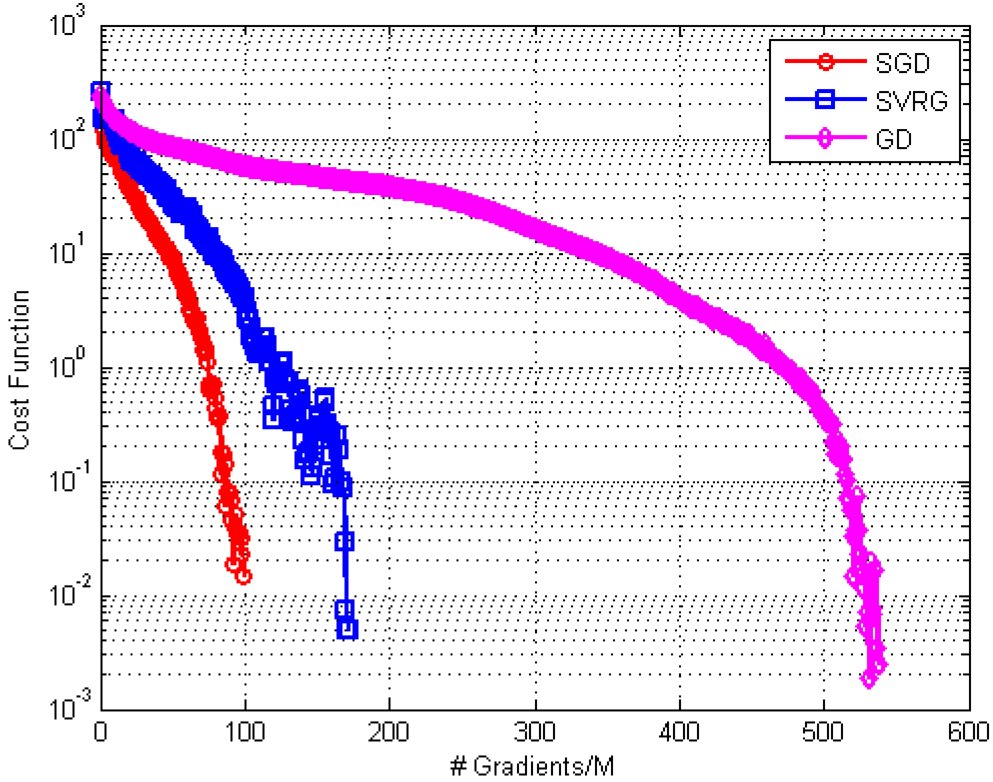
- **Unconstrained Case**
  - Not applicable in general; $F^{(s)}(\mathbf{x})$ is a quartic polynomial
- **Constrained Case**
  - Requires step-size $\mathcal{O}(\mu)$
    - Too small to work well in practice
    - Stationary point not guaranteed to be feasible
- **Heuristic Choices**
  - Diminishing: $\mathcal{O}(1/k^{\gamma}), \gamma \in [0.5, 1]$
  - Polynomial: $\mathcal{O}(1/(1+\alpha k/M))^{\gamma}, \alpha > 0, \gamma \in [0.5, 1]$
    - Generalization of inverse-t step schedule for SGD
  - N-LMS: $\mathcal{O}(1/\|\mathbf{x}^{(k)}\|_2^2)$
    - Simple counter-example where this works for minimizing a quartic function and all other reasonable step-sizes fail [Re *et al.* 2015]

Synthetic Experiments

- Feasibility for random systems of quadratic inequalities
  - Generate nonconvex quadratic feasibility problem such that there exists a feasible solution $\mathbf{p}$ with unit norm
  - Generate $\{\mathbf{A}_m\}_{m=1}^{M}$ from i.i.d. standard normal distribution
  - Generate $b_m \sim \mathcal{N}(\mathbf{p}^T\mathbf{A}_m\mathbf{p}, 1), \forall\, m \in \mathcal{M}$
- Algorithmic Setup:
  - Set $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^N | \|\mathbf{x}\|_2 \leq 1\}$, $\mu = 10^{-4}, K = 4M$
  - Initialize GD, SVRG and SGD from the same randomly generated unit-vector (no restarts)
  - GD, SVRG and SGD have a total gradient budget of $1000M$ gradients
  - Polynomial step-size rule for GD and SVRG; diminishing step-size rule for SGD
  - Feasibility declared if $F^{(ns)}(\mathbf{x}) < 10^{-6}$

Illustrative Example

N = 200, M = 1000, single instance

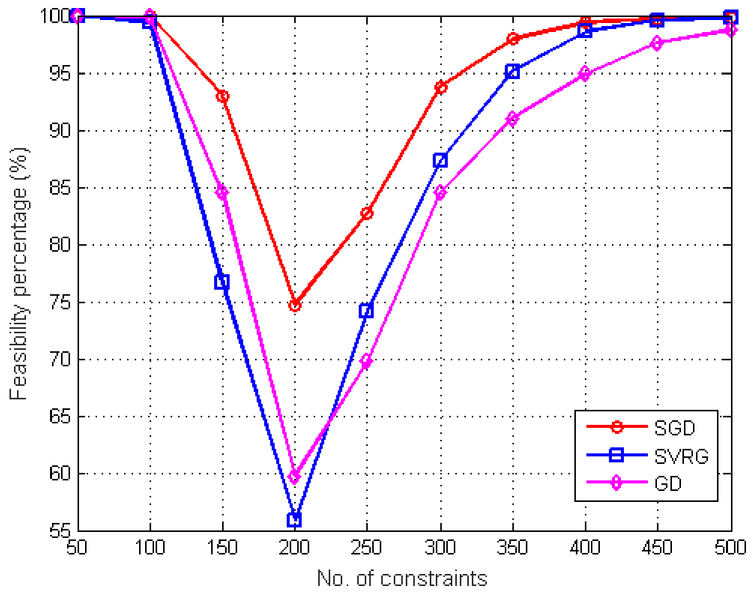


Timing: SGD – 17 secs, SVRG – 27 secs, GD – 83 secs

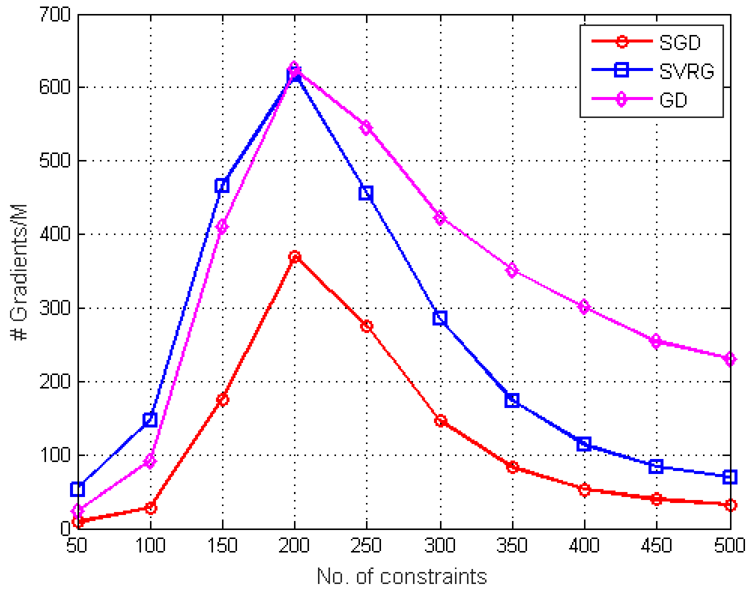
Detailed Experiments

N = 50 variables, varying M, 1000 instances for each value of M

Feasibility Percentage vs. M



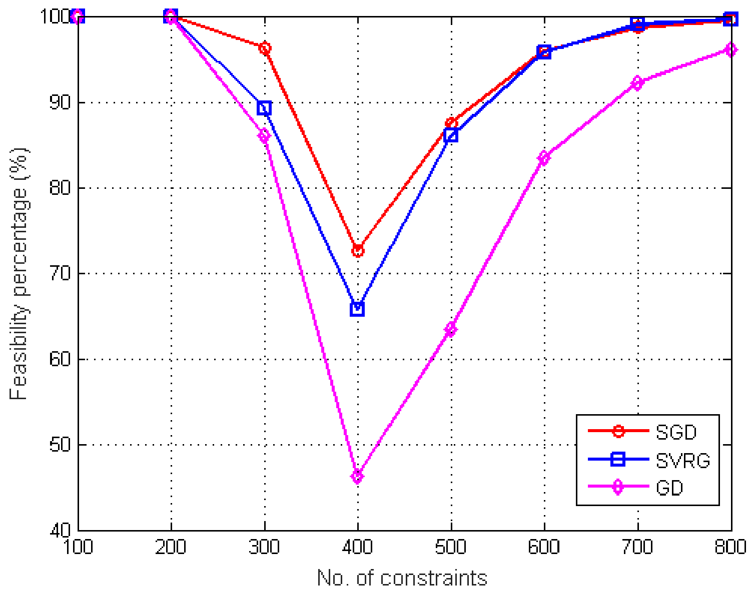
Gradients/M vs. M (feasible cases)



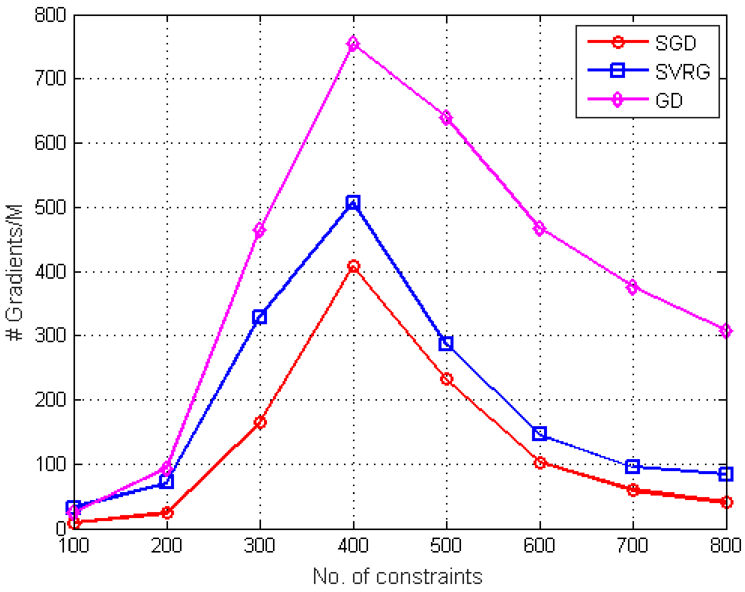
Detailed Experiments

N = 100 variables, varying M, 1000 instances for each value of M

Feasibility Percentage vs. M



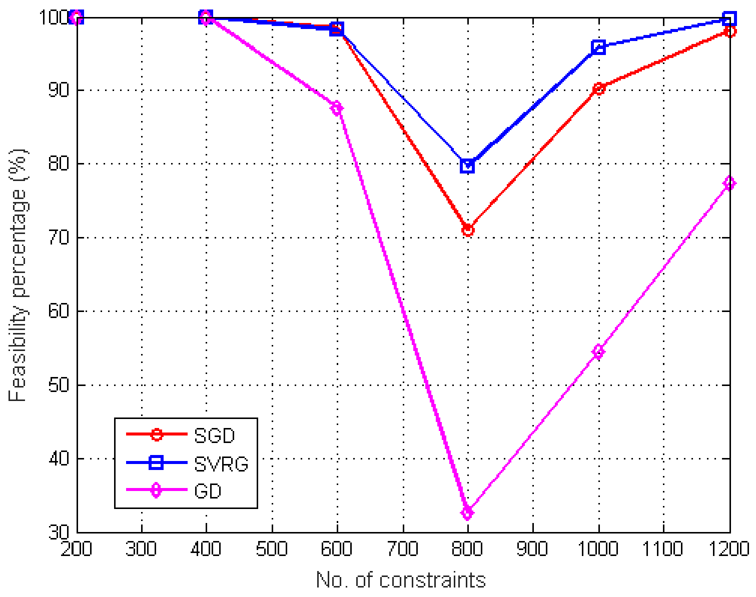
Gradients/M vs. M (feasible cases)



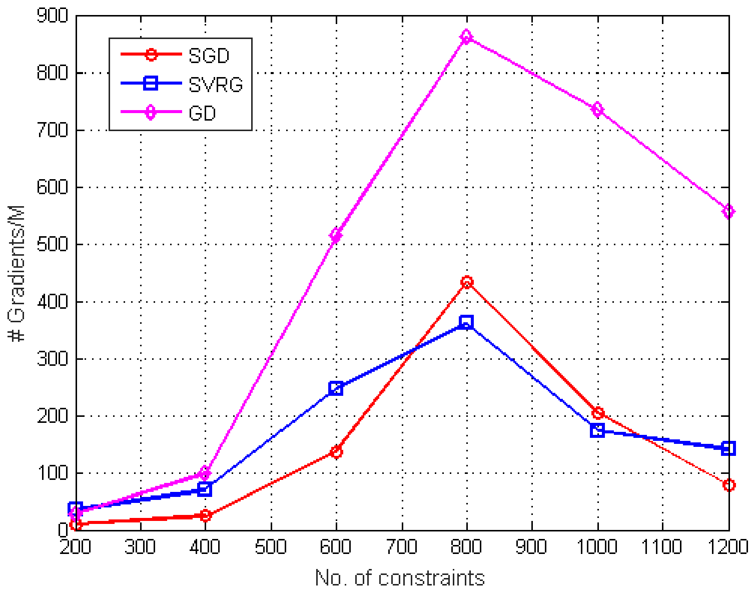
Detailed Experiments

N = 200 variables, varying M, 1000 instances for each value of M

Feasibility Percentage vs. M



Gradients/M vs. M (feasible cases)

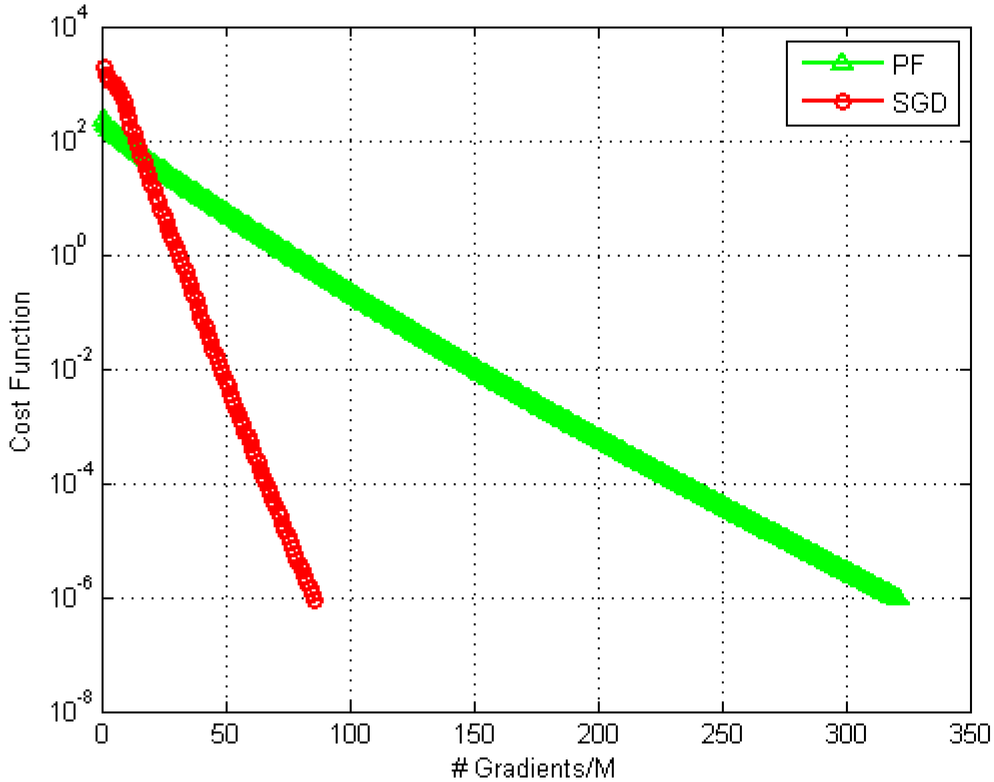


Synthetic Experiments (contd…)

- Solving random systems of quadratic equalities
  - Generate $\{\mathbf{C}_m\}_{m=1}^M$ from spiked Gaussian ensemble
  - A special case of the Matrix Sensing problem [Bhojanapalli *et al.* 2015]
  - If $M = \Omega(N)$, then RIP satisfied with high probability
  - Strict-saddle property satisfied; plus no spurious local minima exist (i.e., all local minima are also global minima)
  - GD and SGD converge to global minima!
- Algorithmic Setup:
  - Set $\mathcal{X} = \mathbb{R}^N$
  - Initialize GD with spectral initialization plus constant step-size; guaranteed (local) linear convergence rate [Tu *et al.* 2015]
  - Initialize SGD with random initialization plus normalized step-size rule; guaranteed convergence in polynomial-time [Ge *et al.* 2015]
  - Gradient budget and termination criterion same as before

Illustrative Example

N = 50, M = 200, single instance

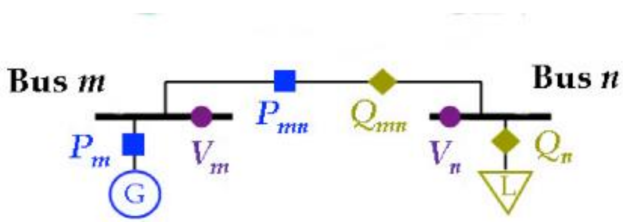


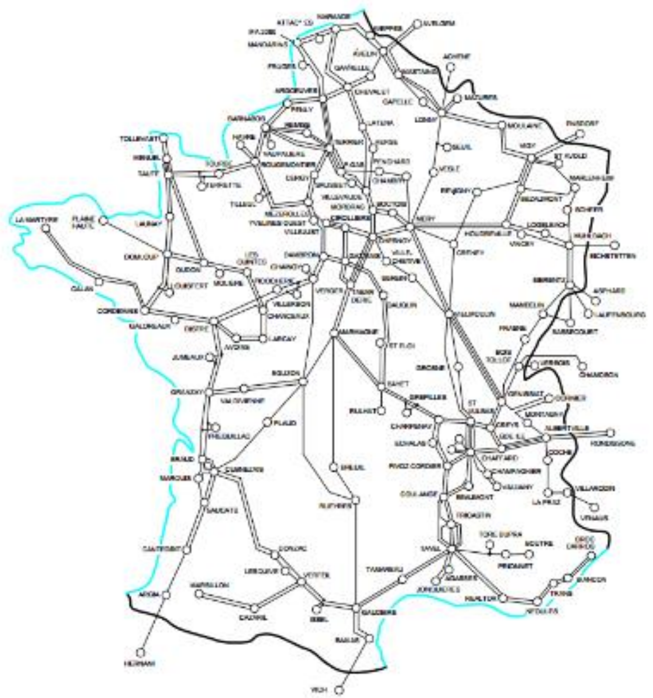
SGD works better in practice

Power System State Estimation

- Problem:
  - Estimate complex voltages at all buses from noisy (Gaussian) power measurements
  - Noisy Case
    - Weighted Least Squares formulation

$$\min_{\mathbf{v}\in\mathbb{R}^{2N}} \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\mathbf{v}^T\mathbf{Y}_m\mathbf{v}-z_m}{\sigma_m}\right)^2$$





Power Transmission Network

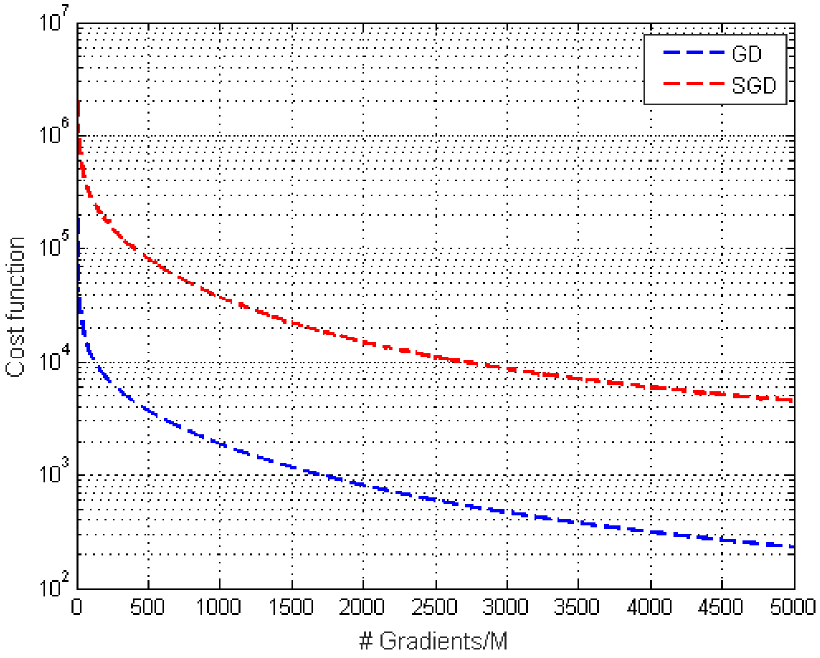
Experiments

- Test Networks obtained from the NESTA archive
  - Voltage profile with magnitude $\sim \mathcal{U}[0.9, 1.1]$ and phase $\sim \mathcal{U}[-0.1\pi, 0.1\pi]$
  - Generate SCADA measurements using MATPOWER
  - Gaussian noise with variances 10 dBm and 13 dBm added to voltage and power measurements respectively
  - Phase of reference bus set to zero
- Algorithmic Setup:
  - Add Gauss-Newton (GN) method (with backtracking line-search) for comparison
  - Initialize GN, GD and SGD from flat start
  - GD and SGD have a total gradient budget of $5000M$ gradients
  - GD with backtracking line-search (provable convergence!); minibatch SGD with normalized step-size rule
  - Output of SGD refined with 1-2 iterations of FPP-SCA [Wang *et al.*, 2016]

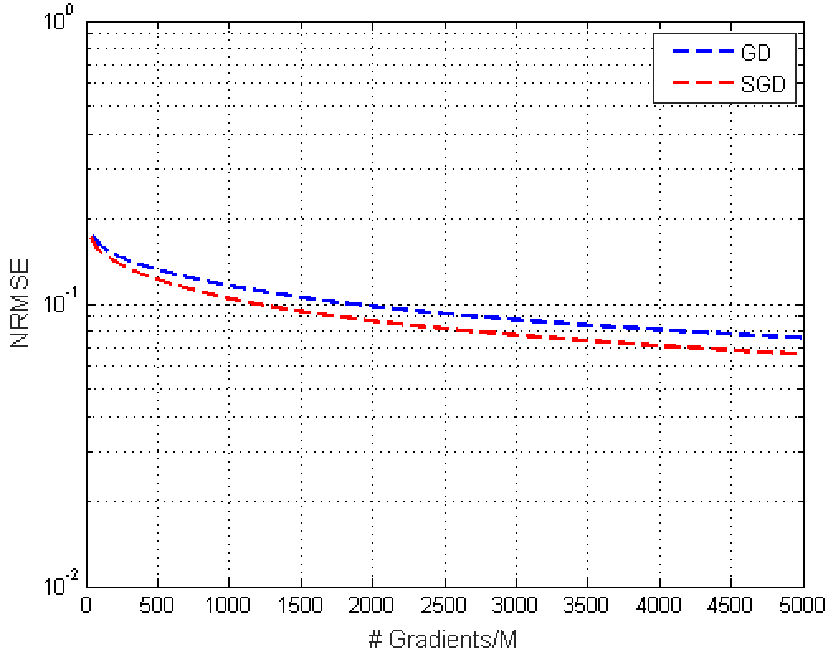
Illustrative Example

IEEE-162 bus network, N = 324 variables, M = 1054 measurements

WLS Cost Function vs. # Gradients/M



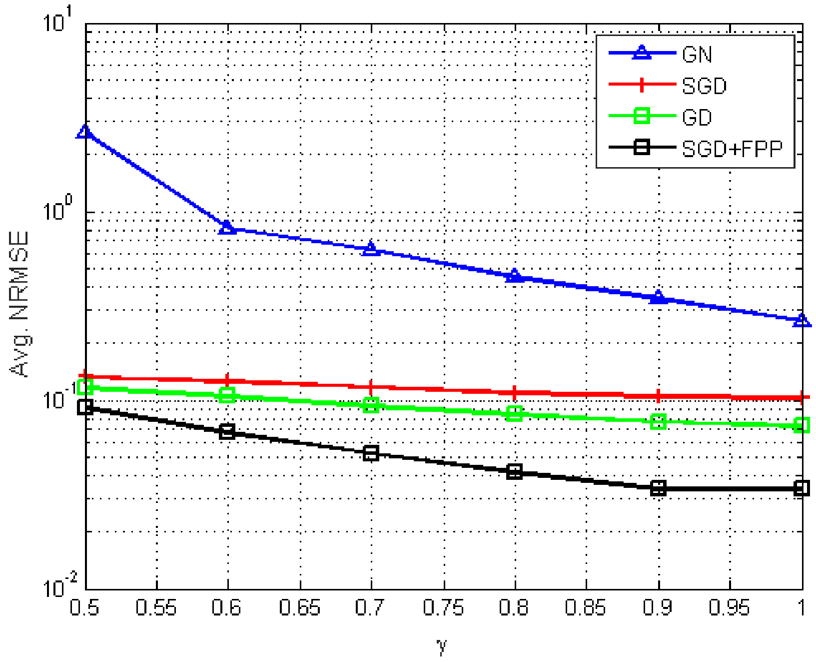
NRMSE vs. # Gradients/M



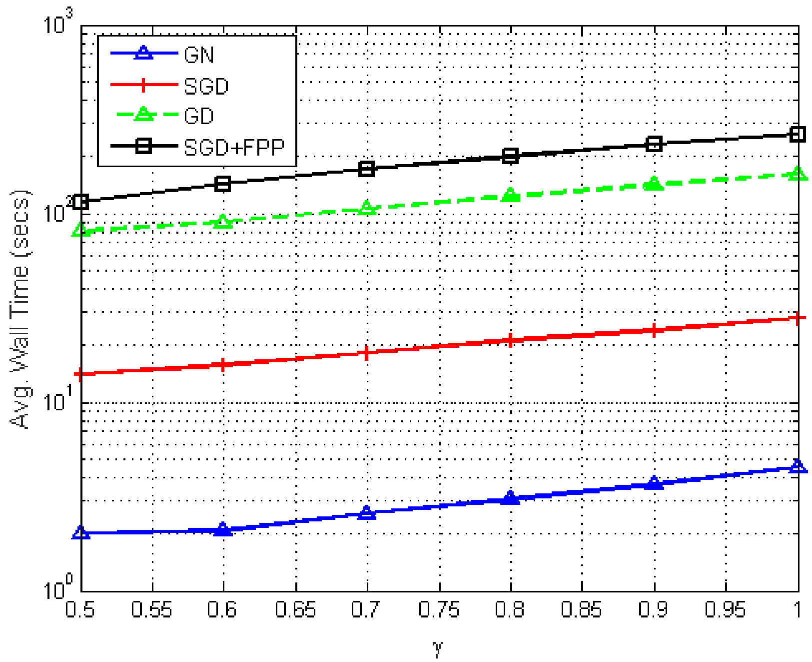
Detailed Experiments

PEGASE-89 bus network, 200 MC trials

NRMSE vs. Measurement Fraction



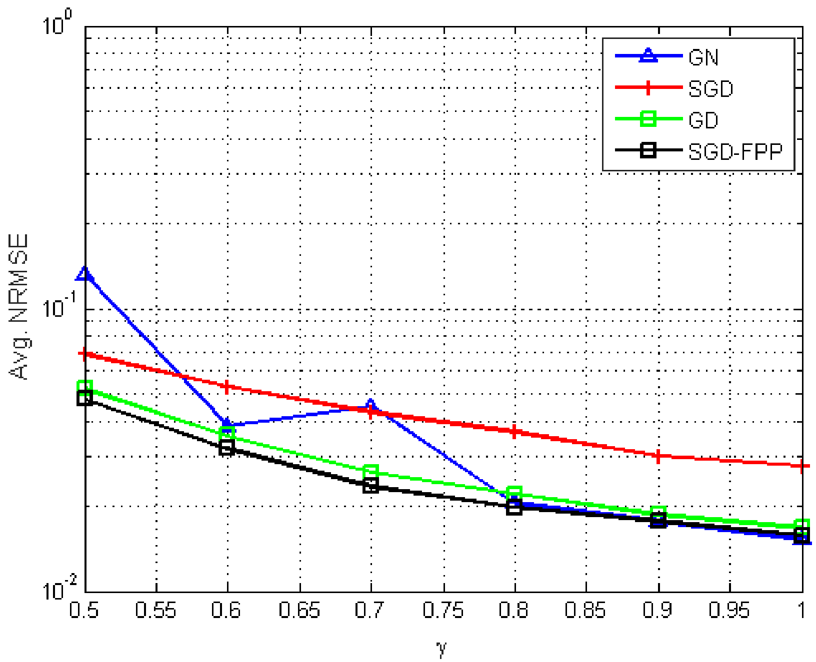
Wall Time vs. Measurement Fraction



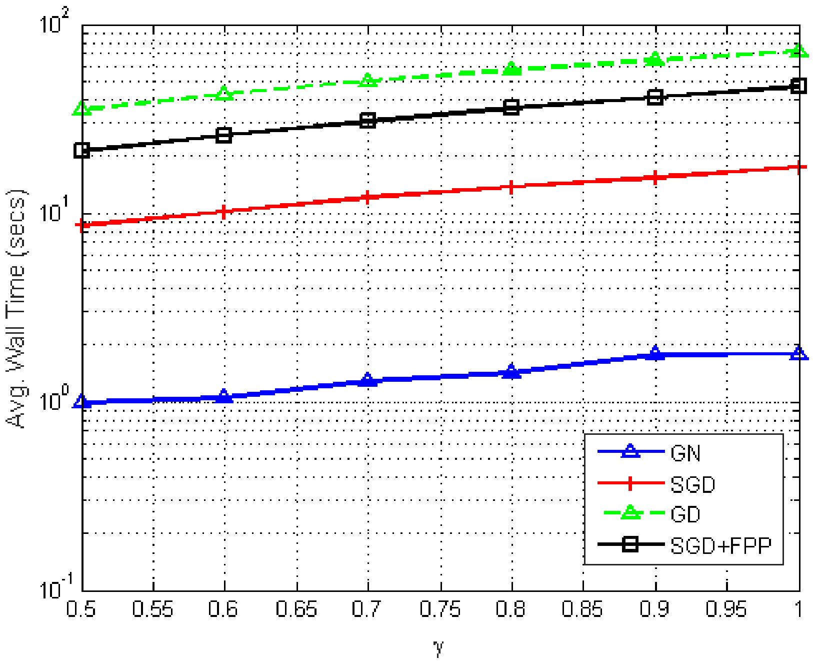
Detailed Experiments

IEEE-73 bus network, 200 MC trials

NRMSE vs. Measurement Fraction



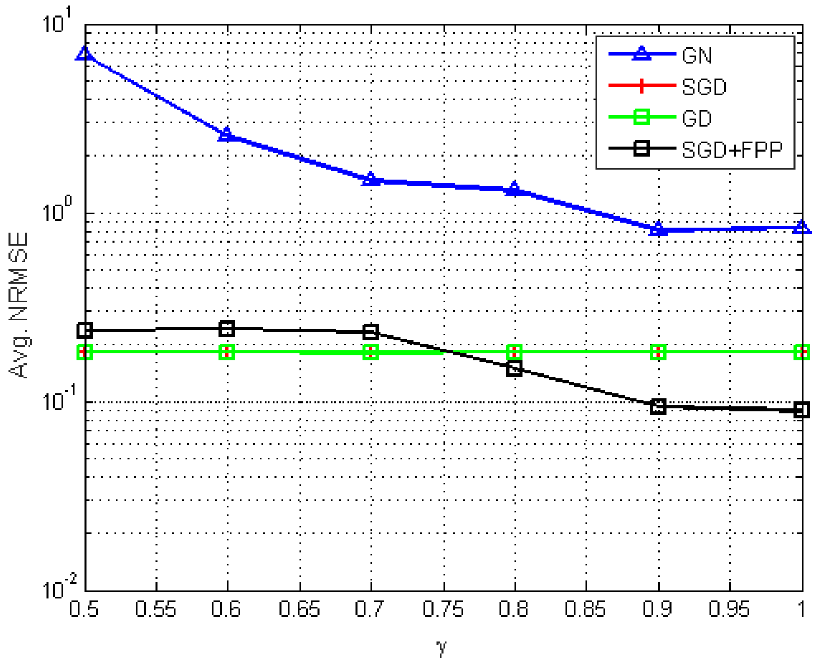
Wall Time vs. Measurement Fraction



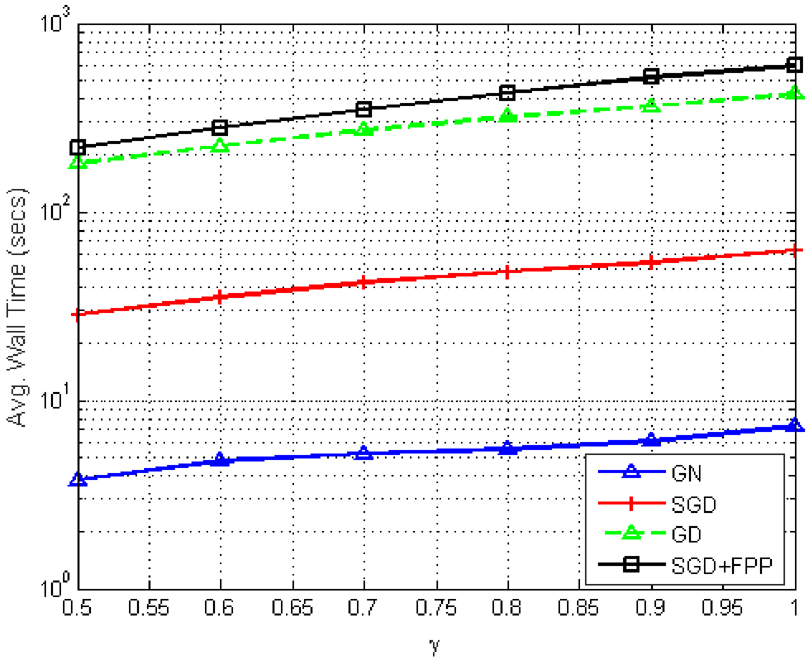
Detailed Experiments

EDIN-189 bus network, 200 MC trials

NRMSE vs. Measurement Fraction



Wall Time vs. Measurement Fraction



Conclusions and Future Work

- First Order Methods for nonconvex quadratic feasibility problems
  - Lightweight in terms of memory and computational resources; well-suited for large-scale problems
  - Stochastic Gradient Methods perform the best
    - Work very well for random problem instances
    - For PSSE, combined SGD + FPP meta-heuristic performs the best overall
- Future work
  - Develop general theoretical guarantees
    - Explain the behavior of algorithms for solving random systems of inequalities
  - SCA via SGD?