Decentralized Online Learning with Heterogeneous Data Sources

Alec Koppel[⋆], Brian M. Sadler[§], and Alejandro Ribeiro[⋆]

[⋆]University of Pennsylvania, Philadelphia, PA
[§]U.S. Army Research Laboratory, Adelphi, MD

Global Conference in Signal and Information Processing

Washington, DC, Dec., 8, 2016

Large-Scale Parameter Estimation

- Learning $\Rightarrow$ params $\mathbf{x}^* \in \mathbb{R}^p$ that minimize stat. avg. loss $F(\mathbf{x})$
- $f : \mathbb{R}^p \to \mathbb{R}$ $\Rightarrow$ convex loss, quantifies merit of statistical model

  $\Rightarrow$ $\boldsymbol{\theta}$ is random variable representing data stream

$$\mathbf{x}^* := \underset{\mathbf{x}}{\operatorname{argmin}} F(\mathbf{x}) := \underset{\mathbf{x}}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{\theta}}[f(\mathbf{x}, \boldsymbol{\theta})]$$

Large-Scale Parameter Estimation

- Learning $\Rightarrow$ params $\mathbf{x}^* \in \mathbb{R}^p$ that minimize stat. avg. loss $F(\mathbf{x})$
- $f : \mathbb{R}^p \to \mathbb{R}$ $\Rightarrow$ convex loss, quantifies merit of statistical model

  $\Rightarrow$ $\boldsymbol{\theta}$ is random variable representing data stream
- Suppose N i.i.d. samples $\boldsymbol{\theta}_n$ of stationary dist. of $\boldsymbol{\theta}$

  $\Rightarrow$ $f_n(\mathbf{x}) := f(\mathbf{x}, \boldsymbol{\theta}_n)$ loss associated with n-th sample

$$\mathbf{x}^* := \operatorname*{argmin}_{\mathbf{x}} F(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{x}} \frac{1}{N} \sum_{n=1}^{N} f_n(\mathbf{x})$$

- Example problems:

  $\Rightarrow$ support vector machines

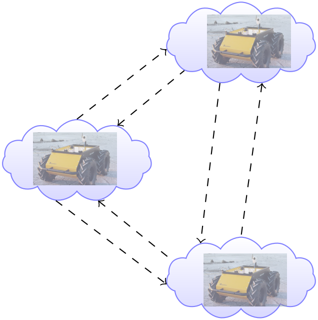
  $\Rightarrow$ logistic regression

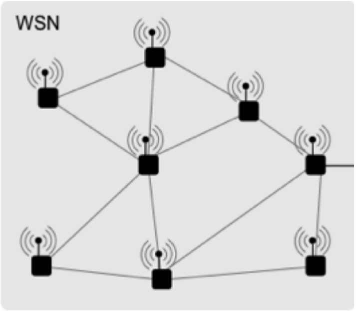
  $\Rightarrow$ matrix completion

Large-Scale Parameter Estimation

- Learning $\Rightarrow$ params $\mathbf{x}^* \in \mathbb{R}^p$ that minimize stat. avg. loss $F(\mathbf{x})$
- $f : \mathbb{R}^p \rightarrow \mathbb{R}$ $\Rightarrow$ convex loss, quantifies merit of statistical model

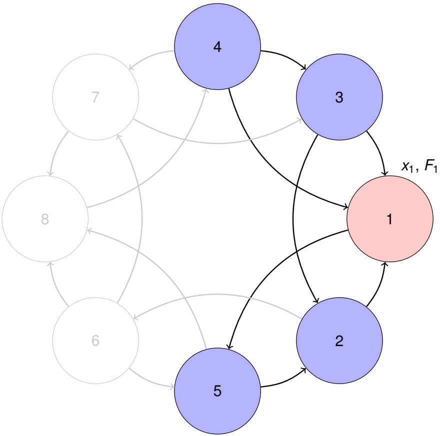
  $\Rightarrow$ $\boldsymbol{\theta}$ is random variable representing data stream
- Focus: data scattered across network (robot team, IoT, sensors)





Multi-Agent Optimization

- Network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - $\Rightarrow |\mathcal{V}| = V$, $|\mathcal{E}| = E$
- $\boldsymbol{\theta}_{i,t} \Rightarrow$ data stream of agent i
- Wants to find $\mathbf{x}_i^L = \text{argmin}_{\mathbf{x}_i} F_i(\mathbf{x}_i)$
  - $\Rightarrow$ local obj: $F_i(\mathbf{x}_i) = \mathbb{E}_{\boldsymbol{\theta}_i}[f(\mathbf{x}_i, \boldsymbol{\theta}_i)]$
- Stacked prob: $\mathbf{x}^L = \text{argmin}_{\mathbf{x}} F(\mathbf{x})$
  - $\Rightarrow$ Global Obj: $F(\mathbf{x}) = \sum_{i \in \mathcal{V}} F_i(\mathbf{x}_i)$



- **Hypothesis**: agents' probs. related
  - $\Rightarrow$ e.g. seek same params. $\mathbf{x}_i = \mathbf{x}_j$
  - $\Rightarrow$ agents exploit others' obs.

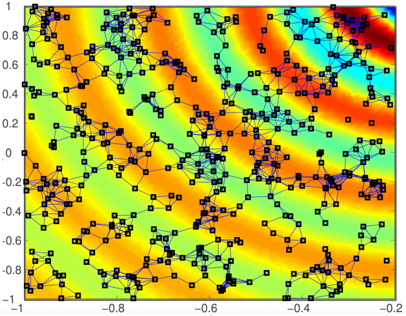
$\Rightarrow$ Consensus: Minimize global loss with equality constraints

$$\min_{\mathbf{x} \in \mathcal{X}^V} \sum_{i \in \mathcal{V}} F_i(\mathbf{x}_i) \text{ s. t. } \mathbf{x}_i = \mathbf{x}_j \text{ for all } (i, j) \in \mathcal{E}$$

$\Rightarrow$ Implicitly only makes sense when info. is from common dist.

Heterogeneous Multi-Agent Optimization

- **Hypothesis**: nearby nodes' params.
  - $\Rightarrow$ close, not necessarily equal
  - $\Rightarrow$ e.g., estimate non-uniform field
- Local cvx. proximity func. $h_{ij}(\mathbf{x}_i, \mathbf{x}_j)$
  - $\Rightarrow$ tolerance $\gamma_{ij} \geq 0$ (prior $\rho(\mathbf{x}_i, \mathbf{x}_j)$)



$\Rightarrow$ **Proximity-Constrained Optimization**:

$$\min_{\mathbf{x} \in \mathcal{X}^V} \sum_{i \in \mathcal{V}} F_i(\mathbf{x}_i)$$

$$\text{s. t. } h(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma_{ij} \text{ for all } j \in n_i$$

$\Rightarrow$ Multi-agent prob. with convex stoch. obj. and cvx. inequality cons.

Background

- Online consensus optimization
  - ⇒ primal (DGD): local SGD + weighted averaging (Nedich '07)
  - ⇒ dual (MM, ADMM): dual function + dual ascent step (Ling '14)
  - ⇒ primal-dual: primal-dual descent-ascent (Mateos-Nuez '16)

- Extensions to heterogeneous/correlated networks
  - ⇒ DGD + inequality constraints via penalty function (Towfic '14)
  - ⇒ square-loss + assumptions on correlation (Chen '14)

- This work: multi-agent stochastic opt. with inequality constraints
  - ⇒ Achieved via primal-dual methods (stochastic saddle point)
  - ⇒ Able to encode correlation information into opt. algorithm
  - ⇒ Want to use constant step-size ⇒ better practical estimation

Stochastic Saddle Point Method

- Recall the problem

$$\min_{\mathbf{x}} \sum_{i \in \mathcal{V}} F_i(\mathbf{x}_i)$$
$$\text{s. t. } h(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma_{ij} \text{ for all } j \in n_i$$

- Let's consider the *augmented* Lagrangian relaxation:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{V} \left[\mathbb{E}_{\boldsymbol{\theta}_i}[f_i(\mathbf{x}_i, \boldsymbol{\theta}_i)] + \frac{1}{2} \sum_{j \in n_i} \left(\lambda_{ij} \left(h_{ij}(\mathbf{x}_i, \mathbf{x}_j) - \gamma_{ij} \right) - \frac{\delta \epsilon_t}{2} \lambda_{ij}^2 \right) \right],$$

  $\Rightarrow$ dual regularizer $\frac{\delta \epsilon_t}{2} \lambda_{ij}^2$ needed for convergence
  $\Rightarrow$ controls magnitude of dual var. while in unbounded set $\mathbb{R}_+^E$

- To develop saddle pt. method, compute grads. of Lagrangian
  $\Rightarrow$ Gradients depend on infinitely many realizations of $\boldsymbol{\theta}$
  $\Rightarrow$ Therefore, consider stochastic approx. of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$:

$$\hat{\mathcal{L}}_t(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{V} \left[f_i(\mathbf{x}_i, \boldsymbol{\theta}_{i,t}) + \frac{1}{2} \sum_{j \in n_i} \lambda_{ij} \left(h_{ij}(\mathbf{x}_i, \mathbf{x}_j) - \gamma_{ij} \right) - \frac{\delta \epsilon_t}{2} \lambda_{ij}^2 \right].$$

Stochastic Saddle Point Method

- Recall the problem

$$\min_{\mathbf{x}} \sum_{i \in \mathcal{V}} F_i(\mathbf{x}_i)$$
$$\text{s. t. } h(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma_{ij} \text{ for all } j \in n_i$$

- Apply Arrow-Hurwicz saddle point method to stoch. Lagrangian

  $\Rightarrow$ Primal stochastic descent step:

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{X}^N}\Big[\mathbf{x}_t - \epsilon_t \nabla_{\mathbf{x}} \hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)\Big] \ ,$$

  $\Rightarrow$ Dual stochastic ascent step:

$$\boldsymbol{\lambda}_{t+1} = \Big[\boldsymbol{\lambda}_t + \epsilon_t \nabla_{\boldsymbol{\lambda}} \hat{\mathcal{L}}_t(\mathbf{x}_t, \boldsymbol{\lambda}_t)\Big]_+ \ ,$$

Decentralized Online Protocol

- Projected stochastic saddle point yields an algorithm in which
  $\Rightarrow$ Update of node i only depends on local and neighbors' info.

$$\mathbf{x}_{i,t+1} = \mathcal{P}_{\mathcal{X}}\Big[\mathbf{x}_{i,t} - \epsilon_t\Big(\nabla_{\mathbf{x}_i} f_i(\mathbf{x}_{i,t};\boldsymbol{\theta}_{i,t}) + \frac{1}{2}\sum_{j\in n_i}(\lambda_{ij,t}+\lambda_{ji,t})\nabla_{\mathbf{x}_i} h_{ij}(\mathbf{x}_{i,t},\mathbf{x}_{j,t})\Big)\Big]$$

  $\Rightarrow$ Dual variable updates along edges $(i,j)\in\mathcal{E}$ take the form

$$\lambda_{ij,t+1} = \Big[(1-\epsilon_t^2\delta)\lambda_{ij,t} + \epsilon_t\left(h_{ij}(\mathbf{x}_{i,t},\mathbf{x}_{j,t}) - \gamma_{ij}\right)\Big]_+ .$$

- Therefore, we can use this algorithm in a multi-agent system

Technical Conditions

- Network $\mathcal{G}$ $\Rightarrow$ symmetric, connected with diameter D.
- Stacked instantaneous obj. $\Rightarrow$ L_f-Lipschitz cont. on avg.

$$\mathbb{E}\|f(\mathbf{x},\theta)-f(\tilde{\mathbf{x}},\theta)\|\leq L_f\|\mathbf{x}-\tilde{\mathbf{x}}\| \ .$$

- Stacked constraint function $h(\mathbf{x})$ is L_h-Lipschitz continuous

$$\|h(\mathbf{x})-h(\tilde{\mathbf{x}})\|\leq L_h\|\mathbf{x}-\tilde{\mathbf{x}}\|.$$

- There exists feasible $(\mathbf{x},\boldsymbol{\lambda})\in\mathcal{X}^V\times\mathbb{R}_+^E$ that are optimal, i.e.,

$$(\mathcal{X}^*\times\boldsymbol{\Lambda}^*)\cap(\mathcal{X}^V\times\mathbb{R}_+^E)\neq\emptyset \qquad \text{(Slater's condition)}$$

Mean Convergence Rates

Theorem

(i) Denote $(\mathbf{x}_t, \boldsymbol{\lambda}_t)$ as the stochastic saddle pt. sequence. After T iterations with a constant step-size $\epsilon_t = \epsilon = 1/\sqrt{T}$, the average time aggregate objective error sequence is bounded sublinearly in T:

$$\sum_{t=1}^{T} \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq \mathcal{O}(\sqrt{T}).$$

The time-aggregate mean constraint violation grows sublinearly in T:

$$\sum_{(i,j)\in\mathcal{E}} \mathbb{E}\Big[\sum_{t=1}^{T} \Big(h_{ij}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) - \gamma_{ij}\Big)\Big]_+ \leq \mathcal{O}(T^{3/4}).$$

- Learning constants are extremely messy
  - $\Rightarrow$ depend on obj. & constraint Lipschitz constants L_f and L_h
  - $\Rightarrow$ diameter of primal set $\mathcal{X}^V$, initialization, network data

Mean Convergence Rates

Corollary

Let $\bar{\mathbf{x}}_T = (1/T)\sum_{t=1}^{T} \mathbf{x}_t$ be the vector formed by averaging the primal saddle point iterates $\mathbf{x}_t$ over times $t = 1, \ldots, T$ with constant step-size $\epsilon_t = 1/\sqrt{T}$. Then the following mean convergence results hold:

$$\mathbb{E}\big[F(\bar{\mathbf{x}}_T) - F(\mathbf{x}^*)\big] \leq \mathcal{O}(1/\sqrt{T})$$

The constraint violation evaluated at the average vector $\bar{\mathbf{x}}_T$ satisfies:

$$\mathbb{E}\Big[\sum_{(i,j)\in\mathcal{E}} \big[h_{ij}(\bar{\mathbf{x}}_{i,T}, \bar{\mathbf{x}}_{j,T}) - \gamma_{ij}\big]_+\Big] = \mathcal{O}(T^{-\frac{1}{4}}).$$

- Easy to establish by applying convexity to previous theorem
  $\Rightarrow$ same learning constant dependence on problem data as thm.

Heterogeneous Estimation: Random Fields

- Random field $\Rightarrow \mathbf{l}_i \in \mathcal{A}$ location of sensor i, field value at $\mathbf{l}_i$: $\mathbf{x}_i$
- Random field parameterized by correlation function $\mathbf{R}_x$
  - $\Rightarrow$ Assumed to follow a spatial structure: $\rho(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|l_i - l_j\|}$
  - $\Rightarrow$ Sensors have unique SNR based upon location in region $\mathcal{A}$
- Aggregate field value across network at time t: $\mathbf{x}_t = \boldsymbol{\mu} + \mathbf{C}^T \mathbf{z}_t$
  - $\Rightarrow$ $\boldsymbol{\mu}$: fixed mean, $\mathbf{C}$: Cholesky factorization of $\mathbf{R}_x$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, 1)$
- Sensors acquire obs. of field at respective positions $\boldsymbol{\theta}_{i,t} \in \mathbb{R}^q$
  - $\Rightarrow$ Noisy linear obs. model: $\boldsymbol{\theta}_{i,t} = \mathbf{H}_i \mathbf{x}_{i,t} + \mathbf{w}_{i,t}$
  - $\Rightarrow$ Signal $\mathbf{x}_i \in \mathbb{R}^p$ contaminated w/ i.i.d. noise $\mathbf{w}_{i,t} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- Goal: sensors seek to minimize its local estimation error

Heterogeneous Estimation: Random Fields

- Instantaneous objective, ignoring neighbors' obs.
  $\Rightarrow f_i(\mathbf{x}_i, \boldsymbol{\theta}_i) = \|\mathbf{H}_i \mathbf{x}_i - \boldsymbol{\theta}_i\|^2$.
  $\Rightarrow$ Estimation $\Rightarrow$ improved via correlated info. of neighbors
  $\Rightarrow$ hurt by making estimates uniformly equal across network

$$\mathbf{x}^* := \underset{\mathbf{x} \in \mathcal{X}^V}{\operatorname{argmin}} \sum_{i=1}^{V} \mathbb{E}_{\boldsymbol{\theta}_i}\Big[\|\mathbf{H}_i \mathbf{x}_i - \boldsymbol{\theta}_i\|^2\Big]$$

$$\text{s.t.} \qquad (1/2)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \gamma_{ij}, \qquad \text{for all } j \in n_i.$$

- $(1/2)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \gamma_{ij} \Rightarrow$ node i's estimate $\mathbf{x}_i^*$ close to neighbors
- For this problem the primal update the form

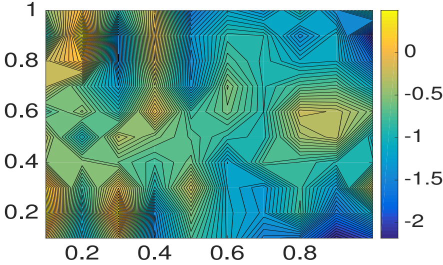
$$\mathbf{x}_{i,t+1} = \mathcal{P}_{\mathcal{X}}\Big[\mathbf{x}_{i,t} - \epsilon_t \Big[2\mathbf{H}_i^T(\mathbf{H}_i \mathbf{x}_{i,t} - \boldsymbol{\theta}_{i,t}) + \frac{1}{2}\sum_{j \in n_i}\Big(\lambda_{ij,t} + \lambda_{ji,t}\Big)\Big(\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\Big)\Big]\Big].$$

- Likewise, the specific form of the dual update is

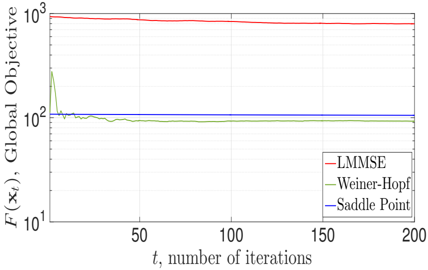
$$\lambda_{ij,t+1} = \Big[(1 - \epsilon_t^2 \delta)\lambda_{ij,t} + (\epsilon_t/2)(\|\mathbf{x}_{i,t} - \mathbf{x}_{j,t}\|^2 - \gamma_{ij})\Big]_+ .$$

Heterogeneous Estimation: Random Fields

- $N = 100$ grid sensor network
  - $\Rightarrow$ deployed in 200 sq. m. region
- Linear estimation w/ corr. obs.
  - $\Rightarrow$ distance corr. $\rho_{ij} = e^{-\|l_i - l_j\|}$
- Constant step-size $\epsilon = 10^{-2.75}$
  - $\Rightarrow$ Prox. func. $\|\mathbf{w}_i - \mathbf{w}_j\|^2 \leq \gamma_{ij}$
  - $\Rightarrow \gamma_{ij} \Rightarrow$ sample correlation
- Comparable performance to (recursive) Weiner-Hopf estimator
  - $\Rightarrow$ via proximity constraints



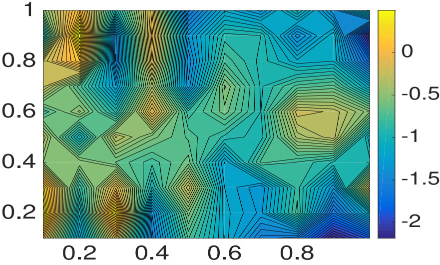
(a) Snapshot of random field



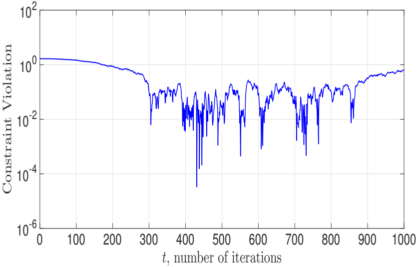
(b) Objective over iteration t

Heterogeneous Estimation: Random Fields

- $N = 100$ grid sensor network
  - $\Rightarrow$ deployed in 200 sq. m. region
- Linear estimation w/ corr. obs.
  - $\Rightarrow$ distance corr. $\rho_{ij} = e^{-\|l_i - l_j\|}$
- Constant step-size $\epsilon = 10^{-2.75}$
  - $\Rightarrow$ Prox. func. $\|\mathbf{w}_i - \mathbf{w}_j\|^2 \leq \gamma_{ij}$
  - $\Rightarrow \gamma_{ij} \Rightarrow$ sample correlation
- Comparable performance to (recursive) Weiner-Hopf estimator
  - $\Rightarrow$ via proximity constraints



(a) Snapshot of random field



(b) Constraint Violation over iteration t

Heterogeneous Estimation: Source Localization

- V sensors deployed in region $\mathcal{A}$, $\mathbf{l}_i$ is location of node i
  - $\Rightarrow$ seek location of a source location $\mathbf{x} \in \mathbb{R}^p$
  - $\Rightarrow$ via access to sequential noisy range obs. $r_{i,t} = \|\mathbf{x} - \mathbf{l}_i\| + \varepsilon_{i,t}$
  - $\Rightarrow$ $\varepsilon_{i,t}$ is some unknown noise vector
- Square-range based least square source localization problem:

$$\mathbf{x}^* := \underset{\mathbf{x} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{r}_i} \left(\|\mathbf{l}_i - \mathbf{x}\|^2 - r_i^2 \right)^2$$

$\Rightarrow$ Non-convex $\Rightarrow$ approx. convexification via change of vars.

$\Rightarrow$ We take convexification w/ constraint

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \min\{\|\mathbf{x}_i - \mathbf{l}_i\|^2, \|\mathbf{x}_j - \mathbf{l}_j\|^2\}$$

$\Rightarrow$ Estimates improve with smaller estimated distance to source

Heterogeneous Estimation: Source Localization

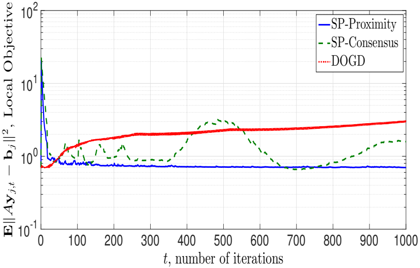
- Expand the square inside expectation: $(\alpha - 2\mathbf{l}_i^T\mathbf{x} + \|\mathbf{l}_i\|^2 - r_i^2)^2$
  $\Rightarrow$ Introduce variable α as $\|\mathbf{x}\| = \alpha$.
- Define matrix $\mathbf{A} \in \mathbb{R}^{N\times(p+1)}$ $\Rightarrow$ ith row is $\mathbf{A}_i = [-2\mathbf{l}_i^T; 1]$,
- Vector $\mathbf{b} \in \mathbb{R}^N$ $\Rightarrow$ ith entry is $\mathbf{b}_i = r_i^2 - \|\mathbf{l}_i\|^2$, $\mathbf{y} = [\mathbf{x}; \alpha] \in \mathbb{R}^{p+1}$.
- Non-convex problem becomes least-squares problem
  $\Rightarrow$ Relax the constraint $\|\mathbf{x}\| = \alpha$.

$$\mathbf{y}^* := \underset{\mathbf{y}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{b}_i}\left(\|\mathbf{A}_i\mathbf{y} - \mathbf{b}_i\|^2\right) ;$$

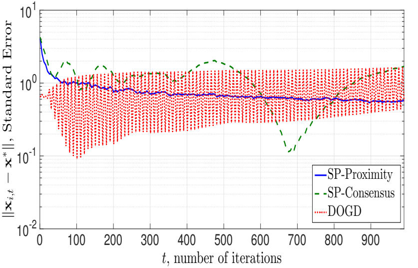
- Approximate non-convex constraint with log-sum-exp function.

Heterogeneous Estimation: Source Localization

- $N = 64$ (8×8) grid network
  - $\Rightarrow$ in 1000 sq. m. region
- $\varepsilon_{i,t} \sim \mathcal{N}(0, 2\|\mathbf{l}_i - \mathbf{x}^*\|)$
  - $\Rightarrow$ dual regularization $\delta = 10^{-7}$
  - $\Rightarrow$ hybrid step-size
  - $\Rightarrow$ $\epsilon_t = \min(\epsilon, \epsilon t_0/t)$, $t_0 = 100$
- Consensus comparison:
  - $\Rightarrow$ DOGD and SP-Consensus
- Proximity constraint SP:
  - $\Rightarrow$ best (in terms of obj. and SE)
  - $\Rightarrow$ larger constraint violation



(a) Local Objective vs. iteration t

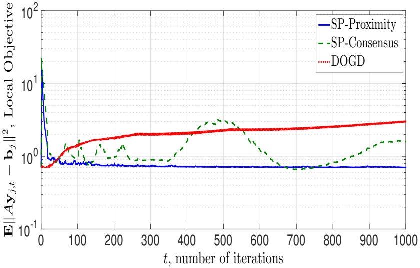


(b) Standard Error over iteration t

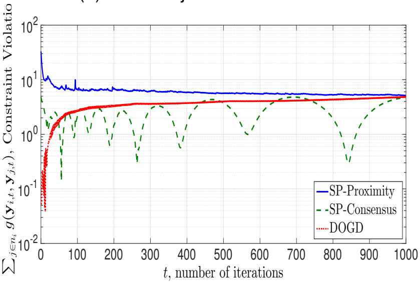
Heterogeneous Estimation: Source Localization

- $N = 64$ (8×8) grid network
  - $\Rightarrow$ in 1000 sq. m. region
- $\varepsilon_{i,t} \sim \mathcal{N}(0, 2\|\mathbf{l}_i - \mathbf{x}^*\|)$
  - $\Rightarrow$ dual regularization $\delta = 10^{-7}$
  - $\Rightarrow$ hybrid step-size
  - $\Rightarrow$ $\epsilon_t = \min(\epsilon, \epsilon t_0/t)$, $t_0 = 100$

- Consensus comparison:
  - $\Rightarrow$ DOGD and SP-Consensus
- Proximity constraint SP:
  - $\Rightarrow$ best (in terms of obj. and SE)
  - $\Rightarrow$ larger constraint violation



(a) Local Objective vs. iteration t



(b) Constraint Violation over iteration t

Conclusions

- We considered multi-agent online opt. prob. (V parallel probs.)
- Consensus: all nodes are trying to learn common parameters
  - $\Rightarrow$ restrictive when latent correlation structure is present
- We handle this issue via convex local proximity constraints
  - $\Rightarrow$ multi-agent stochastic program with inequality constraints

- Solve via primal-dual stochastic saddle point method
- Establish convergence in expectation (for average vectors)
  - $\Rightarrow$ primal mean sub-optimality, mean constraint slack over time

- Applications to random field estimation and source localization
  - $\Rightarrow$ SP outperforms approaches based on consensus

References

- A. Koppel, B. M. Sadler and A. Ribeiro, ”Proximity without consensus in online multi-agent optimization,” in Proc. Int. Conf. Accoustics Speech Signal Process., Shanghai, China, Mar. 20-25 2016.
- A. Koppel, B. Sadler, and A. Ribeiro, ”Proximity without Consensus in Online Multi-Agent Optimization,” in IEEE Trans. Signal Proc. (revised), June 2016.

`http://seas.upenn.edu/~akoppel/`