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
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
- $\theta$ is random variable representing data stream

$$
\mathbf{x}^* := \arg\min_{\mathbf{x}} F(\mathbf{x}) := \arg\min_{\mathbf{x}} \mathbb{E}_\theta[f(\mathbf{x}, \theta)]
$$
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

\( \Rightarrow \) \( \theta \) is random variable representing data stream

Suppose \( N \) i.i.d. samples \( \theta_n \) of stationary dist. of \( \theta \)

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

\[
\mathbf{x}^* := \arg\min_{\mathbf{x}} F(\mathbf{x}) := \arg\min_{\mathbf{x}} \frac{1}{N} \sum_{n=1}^{N} f_n(\mathbf{x})
\]

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
  $\Rightarrow \theta$ is random variable representing data stream
- Focus: data scattered across network (robot team, IoT, sensors)
Multi-Agent Optimization

- Network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  \[ |\mathcal{V}| = V, |\mathcal{E}| = E \]
- $\theta_{i,t}$ \(\Rightarrow\) data stream of agent $i$
- Wants to find $x_i^L = \text{argmin}_{x_i} F_i(x_i)$
  \[ \Rightarrow \text{local obj: } F_i(x_i) = \mathbb{E}_{\theta_i}[f(x_i, \theta_i)] \]
- Stacked prob: $x^L = \text{argmin}_x F(x)$
  \[ \Rightarrow \text{Global Obj: } F(x) = \sum_{i \in \mathcal{V}} F_i(x_i) \]
- **Hypothesis**: agents’ probs. related
  \[ \Rightarrow \text{e.g. seek same params. } x_i = x_j \]
  \[ \Rightarrow \text{agents exploit others’ obs.} \]
  \[ \Rightarrow \textbf{Consensus}: \text{Minimize global loss with equality constraints} \]
  \[ \min_{x \in \mathcal{X}^V} \sum_{i \in \mathcal{V}} F_i(x_i) \text{ s. t. } x_i = x_j \text{ for all } (i, j) \in \mathcal{E} \]
  \[ \Rightarrow \text{Implicitly only makes sense when info. is from common dist.} \]
Heterogeneous Multi-Agent Optimization

- **Hypothesis**: nearby nodes’ params.
  - close, not necessarily equal
  - e.g., estimate non-uniform field

- Local cvx. proximity func. $h_{ij}(x_i, x_j)$
  - tolerance $\gamma_{ij} \geq 0$ (prior $\rho(x_i, x_j)$)

⇒ **Proximity-Constrained Optimization**:

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

s. t. $h(x_i, x_j) \leq \gamma_{ij}$ for all $j \in n_i$

⇒ 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
Recall the problem
\[
\min_x \sum_{i \in V} F_i(x_i)
\]
s. t. \( h(x_i, x_j) \leq \gamma_{ij} \) for all \( j \in n_i \\

Let's consider the augmented Lagrangian relaxation:
\[
\mathcal{L}(x, \lambda) = \sum_{i=1}^{V} \left[ \mathbb{E}_{\theta_i} [f_i(x_i, \theta_i)] + \frac{1}{2} \sum_{j \in n_i} \left( \lambda_{ij} (h_{ij}(x_i, x_j) - \gamma_{ij}) - \frac{\delta \epsilon_t}{2} \lambda_{ij}^2 \right) \right],
\]
\[
\Rightarrow \text{dual regularizer } \frac{\delta \epsilon_t}{2} \lambda_{ij}^2 \text{ needed for convergence}
\Rightarrow \text{controls magnitude of dual var. while in unbounded set } \mathbb{R}_+^E
\]

To develop saddle pt. method, compute grads. of Lagrangian
\Rightarrow \text{Gradients depend on infinitely many realizations of } \theta
\Rightarrow \text{Therefore, consider stochastic approx. of } \mathcal{L}(x, \lambda):
\[
\hat{\mathcal{L}}_t(x, \lambda) = \sum_{i=1}^{V} \left[ f_i(x_i, \theta_{i,t}) + \frac{1}{2} \sum_{j \in n_i} \lambda_{ij} (h_{ij}(x_i, x_j) - \gamma_{ij}) - \frac{\delta \epsilon_t}{2} \lambda_{ij}^2 \right].
\]
Recall the problem

$$\min_x \sum_{i \in \mathcal{V}} F_i(x_i)$$

s. t. $h(x_i, x_j) \leq \gamma_{ij}$ for all $j \in n_i$

Apply Arrow-Hurwicz saddle point method to stoch. Lagrangian

$\Rightarrow$ Primal stochastic descent step:

$$x_{t+1} = P_{\mathcal{X}^N} \left[ x_t - \epsilon_t \nabla_x \hat{L}_t(x_t, \lambda_t) \right],$$

$\Rightarrow$ Dual stochastic ascent step:

$$\lambda_{t+1} = \left[ \lambda_t + \epsilon_t \nabla_{\lambda} \hat{L}_t(x_t, \lambda_t) \right]_+, $$
Projected stochastic saddle point yields an algorithm in which

Update of node $i$ only depends on local and neighbors’ info.

$$x_{i,t+1} = \mathcal{P}x\left[ x_{i,t} - \epsilon_t \left( \nabla x_i f_i(x_{i,t}; \theta_{i,t}) + \frac{1}{2} \sum_{j \in n_i} (\lambda_{ij,t} + \lambda_{ji,t}) \nabla_x h_{ij}(x_{i,t}, x_{j,t}) \right) \right]$$

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

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

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.
  \[ \mathbb{E} \| f(x, \theta) - f(\tilde{x}, \theta) \| \leq L_f \| x - \tilde{x} \| . \]
- Stacked constraint function $h(x)$ is $L_h$-Lipschitz continuous
  \[ \| h(x) - h(\tilde{x}) \| \leq L_h \| x - \tilde{x} \|. \]
- There exists feasible $(x, \lambda) \in \mathcal{X}^V \times \mathbb{R}^E_+$ that are optimal, i.e.,
  \[ (\mathcal{X}^* \times \Lambda^*) \cap (\mathcal{X}^V \times \mathbb{R}^E_+) \neq \emptyset \]  (Slater’s condition)
Theorem
(i) Denote \((x_t, \lambda_t)\) 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 \(T\):

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

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

\[
\sum_{(i,j) \in E} \mathbb{E} \left[ \sum_{t=1}^{T} \left( h_{ij}(x_{i,t}, x_{j,t}) - \gamma_{ij} \right) \right]_+ \leq O(T^{3/4}).
\]

- 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
Corollary

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

\[
\mathbb{E} \left[ F(\bar{x}_T) - F(x^*) \right] \leq O \left( \frac{1}{\sqrt{T}} \right)
\]

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

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

- 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 $\Rightarrow l_i \in A$ location of sensor $i$, field value at $l_i$: $x_i$
- Random field parameterized by correlation function $R_x$:
  $\Rightarrow$ Assumed to follow a spatial structure: $\rho(x_i, x_j) = e^{-\|l_i - l_j\|}$
  $\Rightarrow$ Sensors have unique SNR based upon location in region $A$
- Aggregate field value across network at time $t$: $x_t = \mu + C^T z_t$
  $\Rightarrow \mu$: fixed mean, $C$: Cholesky factorization of $R_x$, $z \sim \mathcal{N}(0, 1)$
- Sensors acquire obs. of field at respective positions $\theta_{i,t} \in \mathbb{R}^q$
  $\Rightarrow$ Noisy linear obs. model: $\theta_{i,t} = H_i x_{i,t} + w_{i,t}$
  $\Rightarrow$ Signal $x_i \in \mathbb{R}^p$ contaminated with i.i.d. noise $w_{i,t} \sim \mathcal{N}(0, \sigma^2 I)$
- Goal: sensors seek to minimize its local estimation error
Heterogeneous Estimation: Random Fields

- Instantaneous objective, ignoring neighbors’ obs.
  \[ f_i(x_i, \theta_i) = \|H_i x_i - \theta_i\|^2. \]

- Estimation improved via correlated info. of neighbors

- Hurt by making estimates uniformly equal across network

\[ x^* := \arg\min_{x \in \mathcal{X}} \sum_{i=1}^{\mathcal{V}} \mathbb{E}_{\theta_i} \left[ \|H_i x_i - \theta_i\|^2 \right] \]

\[ \text{s.t. } (1/2)\|x_i - x_j\|^2 \leq \gamma_{ij}, \quad \text{for all } j \in n_i. \]

- \((1/2)\|x_i - x_j\|^2 \leq \gamma_{ij} \Rightarrow \text{node } i\text{'s estimate } x^*_i \text{ close to neighbors}\]

- For this problem the primal update the form

\[ x_{i,t+1} = \mathcal{P}_\mathcal{X} \left[ x_{i,t} - \epsilon_t \left[ 2H_i^T (H_i x_{i,t} - \theta_i,t) + \frac{1}{2} \sum_{j \in n_i} (\lambda_{ij,t} + \lambda_{ji,t}) \left( x_{i,t} - x_{j,t} \right) \right] \right]. \]

- Likewise, the specific form of the dual update is

\[ \lambda_{ij,t+1} = \left[ (1 - \epsilon_t^2 \delta) \lambda_{ij,t} + (\epsilon_t/2)(\|x_{i,t} - x_{j,t}\|^2 - \gamma_{ij}) \right]_. \]
Heterogeneous Estimation: Random Fields

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

![Snapshot of random field](image)

![Objective over iteration](image)
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. $\|w_i - w_j\|^2 \leq \gamma_{ij}$
  $\Rightarrow \gamma_{ij} \Rightarrow$ sample correlation
- Comparable performance to (recursive) Weiner-Hopf estimator
  $\Rightarrow$ via proximity constraints
Heterogeneous Estimation: Source Localization

- $V$ sensors deployed in region $A$, $l_i$ is location of node $i$
  - seek location of a source location $x \in \mathbb{R}^p$
  - via access to sequential noisy range obs. $r_{i,t} = \|x - l_i\| + \epsilon_{i,t}$
  - $\epsilon_{i,t}$ is some unknown noise vector

- Square-range based least square source localization problem:

$$x^* := \arg\min_{x \in \mathbb{R}^p} \sum_{i=1}^N \mathbb{E}_{r_i}(\|l_i - x\|^2 - r_i^2)^2$$

  - Non-convex  ⇒  approx. convexification via change of vars.
  - We take convexification w/ constraint

$$\|x_i - x_j\|^2 \leq \min\{\|x_i - l_i\|^2, \|x_j - l_j\|^2\}$$

  - Estimates improve with smaller estimated distance to source
Expand the square inside expectation: \((\alpha - 2l_i^T x + ||l_i||^2 - r_i^2)^2\)

\[\Rightarrow \text{Introduce variable } \alpha \text{ as } ||x|| = \alpha.\]

Define matrix \(A \in \mathbb{R}^{N \times (p+1)} \Rightarrow \text{ith row is } A_i = [-2l_i^T; 1],\)

Vector \(b \in \mathbb{R}^N \Rightarrow \text{ith entry is } b_i = r_i^2 - ||l_i||^2, \ y = [x; \alpha] \in \mathbb{R}^{p+1}.\)

Non-convex problem becomes least-squares problem

\[\Rightarrow \text{Relax the constraint } ||x|| = \alpha.\]

\[y^* := \arg\min_{y \in \mathbb{R}^{p+1}} \sum_{i=1}^{N} \mathbb{E}_{b_i} \left( ||A_i y - b_i||^2 \right);\]

Approximate non-convex constraint with log-sum-exp function.
Heterogeneous Estimation: Source Localization

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

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

![Graphs showing Local Objective vs. iteration $t$ and Standard Error over iteration $t$.]
Heterogeneous Estimation: Source Localization

- \( N = 64 \ (8 \times 8) \) grid network
  - in 1000 sq. m. region
- \( \varepsilon_{i,t} \sim \mathcal{N}(0, 2\|l_i - \mathbf{x}^*\|) \)
  - dual regularization \( \delta = 10^{-7} \)
  - hybrid step-size
  - \( \epsilon_t = \min(\epsilon, \epsilon_t_0 / t), \ t_0 = 100 \)

- Consensus comparison:
  - DOGD and SP-Consensus
- Proximity constraint SP:
  - best (in terms of obj. and SE)
  - larger constraint violation
We considered multi-agent online opt. prob. (\( V \) parallel probs.)

**Consensus:** all nodes are trying to learn **common parameters**
\[ \Rightarrow \text{restrictive when latent correlation structure is present} \]

We handle this issue via **convex local proximity constraints**
\[ \Rightarrow \text{multi-agent stochastic program with inequality constraints} \]

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

- Applications to random field estimation and source localization
  \[ \Rightarrow \text{SP outperforms approaches based on consensus} \]


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