Distributed Approaches to Mirror Descent for Stochastic Learning over Rate-Limited Networks

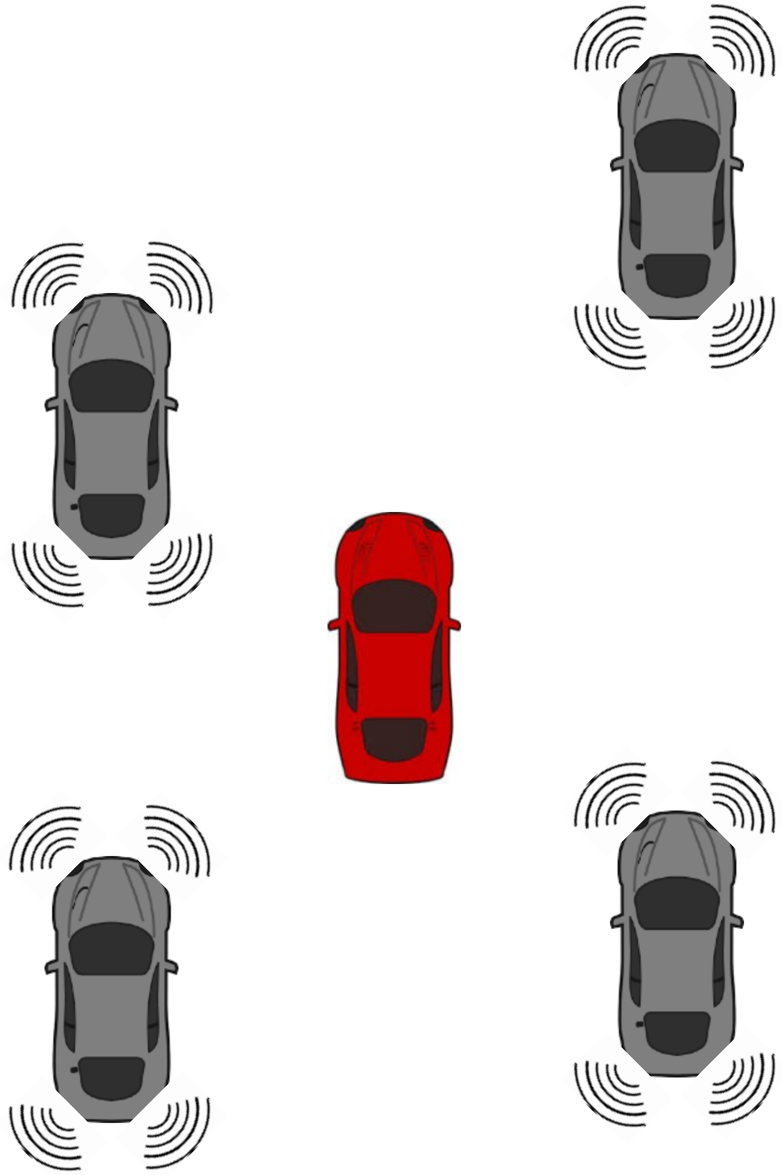
Matthew Nokleby, Wayne State University, Detroit MI
(joint work with Waheed Bajwa, Rutgers)

WAYNE STATE UNIVERSITY

RUTGERS

Motivation: Autonomous Driving

- Network of autonomous automobiles + one human-driven car
- Sensing for "anomalous" driving from human
- Want to jointly sense over communications links

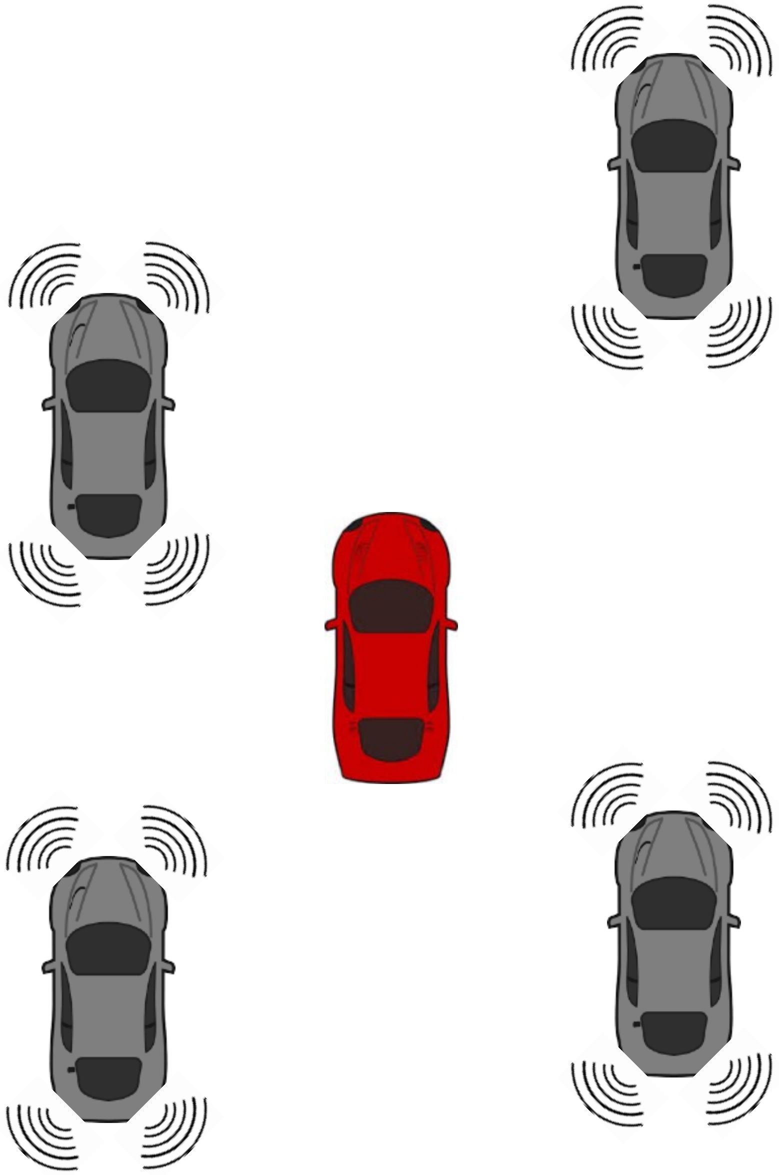


Motivation: Autonomous Driving

- Network of autonomous automobiles + one human-driven car
- Sensing for "anomalous" driving from human
- Want to jointly sense over communications links

Challenges:

- Need to detect/act quickly
- Wireless links have limited rate – can't exchange raw data



Motivation: Autonomous Driving

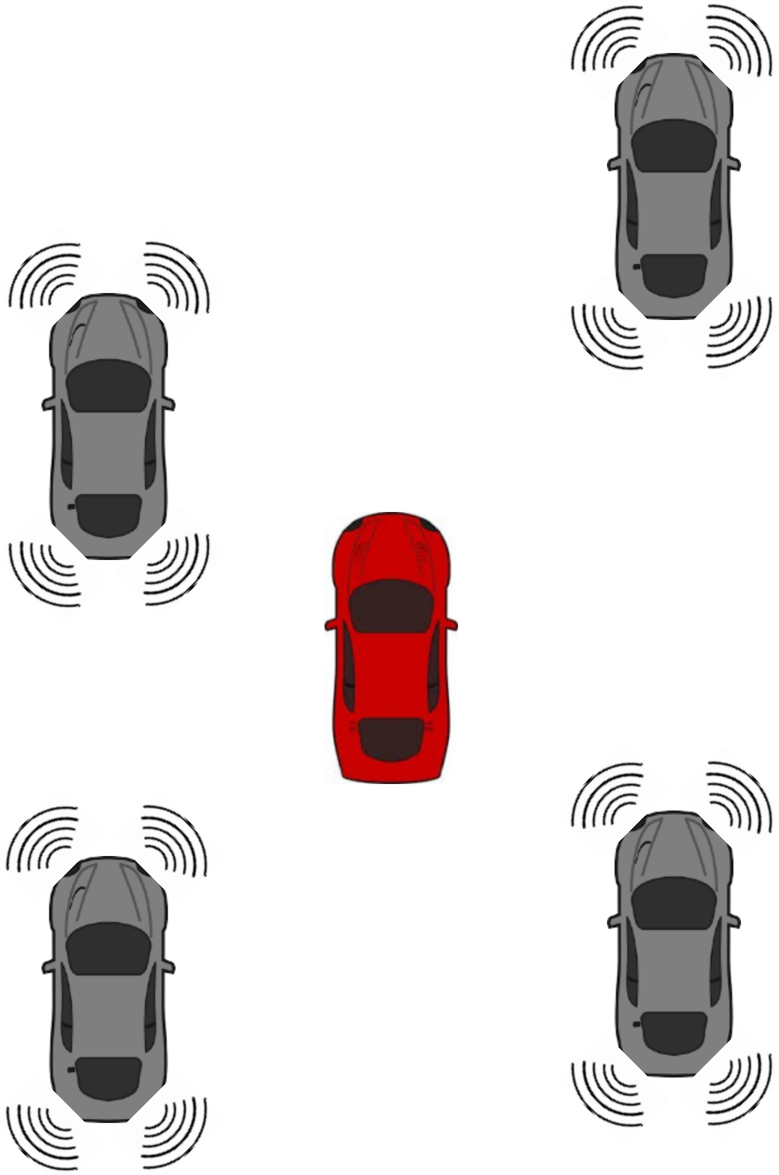
- Network of autonomous automobiles + one human-driven car
- Sensing for "anomalous" driving from human
- Want to jointly sense over communications links

Challenges:

- Need to detect/act quickly
- Wireless links have limited rate – can't exchange raw data

Questions:

- How well can devices jointly learn when links are slow(/not fast)?
- What are good strategies?



Contributions of This Talk

- Frame the problem as distributed **stochastic optimization**
- Network of devices trying to minimize an objective function from streams of noisy data

Contributions of This Talk

- Frame the problem as distributed **stochastic optimization**
- Network of devices trying to minimize an objective function from streams of noisy data

- Focus on communications aspect: how to collaborate when links have limited rates?
- Defining **two time scales**: one rate for data arrival, and one for message exchanges

Contributions of This Talk

- Frame the problem as distributed **stochastic optimization**
- Network of devices trying to minimize an objective function from streams of noisy data

- Focus on communications aspect: how to collaborate when links have limited rates?
- Defining **two time scales**: one rate for data arrival, and one for message exchanges

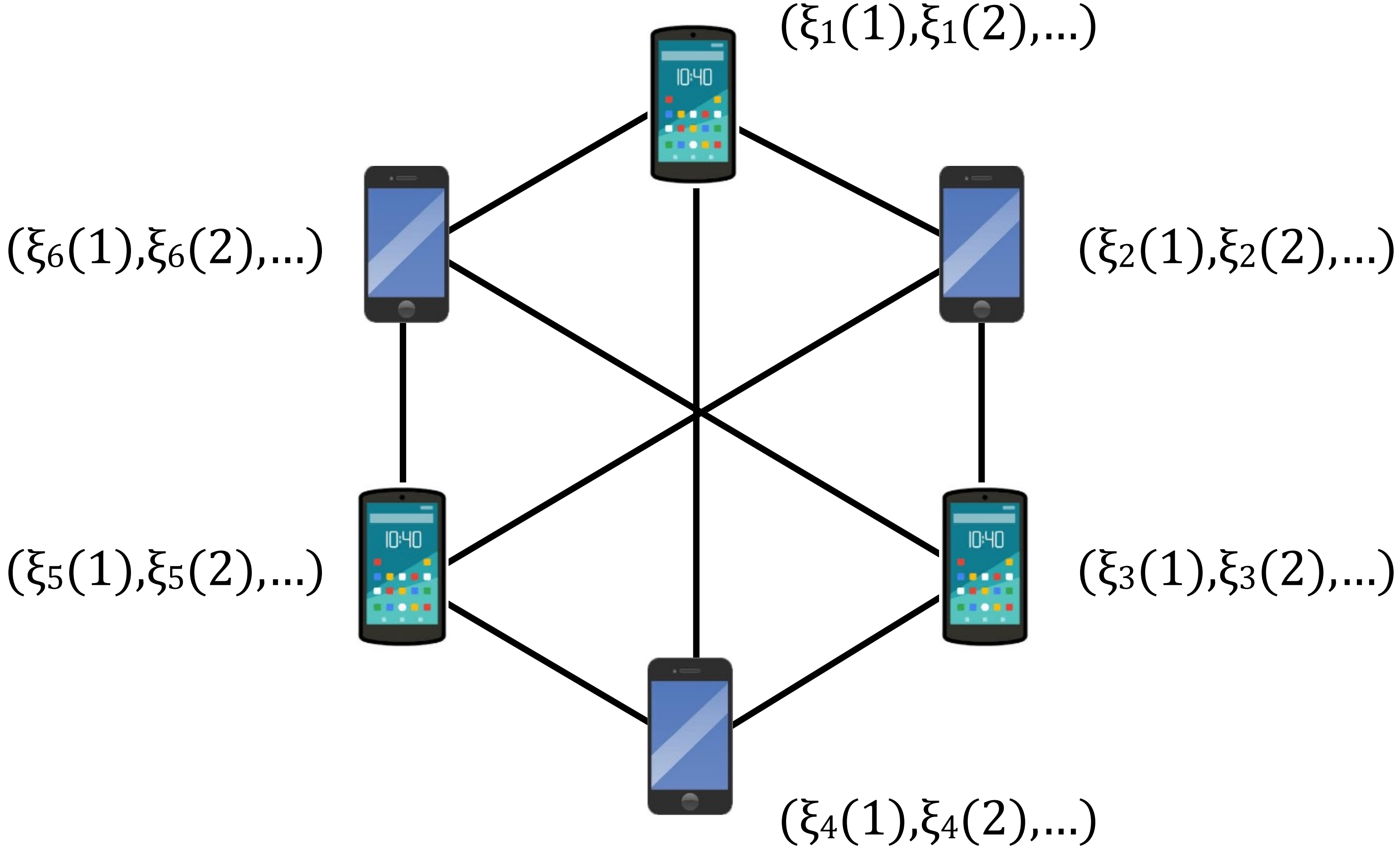
- Solution: distributed versions of **stochastic mirror descent** that carefully balance **gradient averaging** and **mini-batching**
- Derive network/rate conditions for near-optimum convergence
- **Accelerated** methods provide a substantial speedup

Distributed Stochastic Learning

- Network of *m* nodes, each with an i.i.d. data stream

$$\{\xi_i(t)\}, \text{ for sensor i at time t}$$

- Nodes communicate over wireless links, modeled by graph



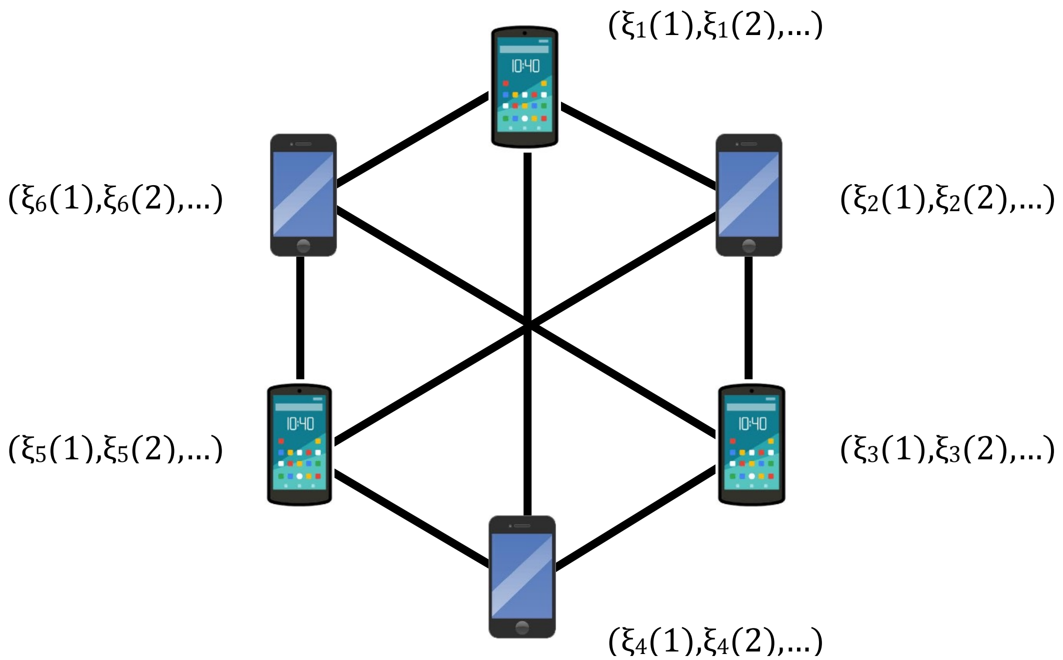
Stochastic Optimization Model

- Nodes want to solve the stochastic optimization problem:

$$\min_{x \in X} \psi(x) = \min_{x \in X} E_\xi[\phi(x,\xi)]$$

- ϕ is convex, $X \subset \mathbb{R}^d$ is compact and convex
- ψ has Lipschitz gradients: [composite optimization later!]

$$||\nabla\psi(x) - \nabla\psi(y)|| \leq L||x - y||, x,y \in X$$



Stochastic Optimization Model

- Nodes want to solve the stochastic optimization problem:

$$\min_{x \in X} \psi(x) = \min_{x \in X} E_\xi[\phi(x,\xi)]$$

- ϕ is convex, $X \subset \mathbb{R}^d$ is compact and convex
- ψ has Lipschitz gradients: [composite optimization later!]

$$||\nabla\psi(x) - \nabla\psi(y)|| \leq L||x - y||, \; x,y \in X$$

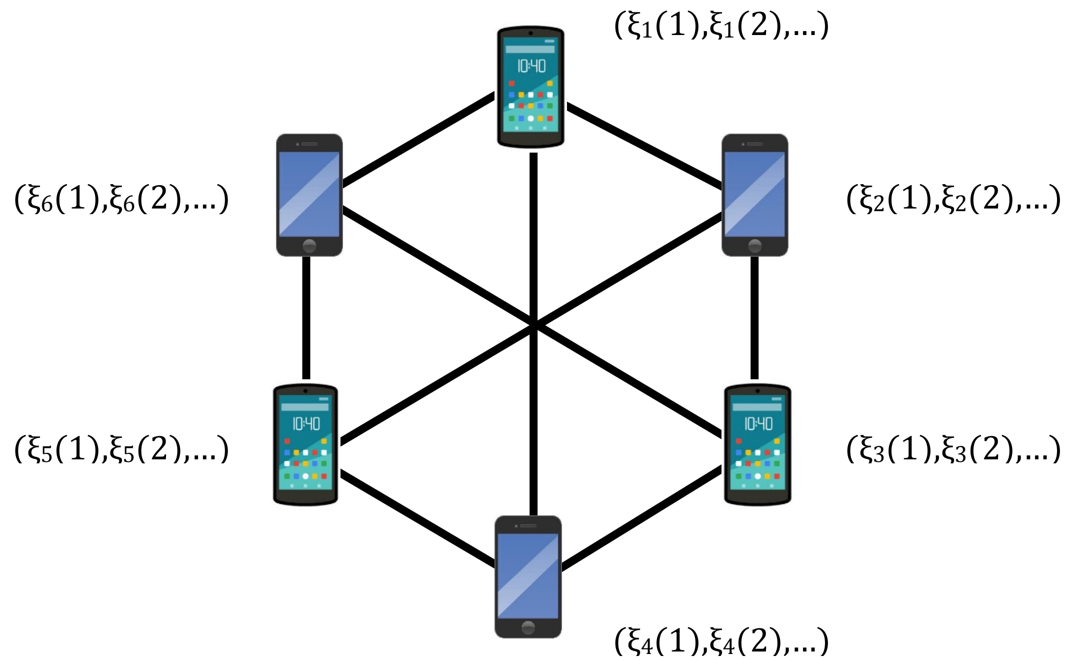
- Nodes have access to noisy gradients:

$$g_i(t) := \nabla\phi(x_i(t),\xi_i(t))$$

$$E_\xi[g_i(t)] = \nabla\psi(x_i(t))$$

$$E_\xi[||g_i(t) - \nabla\psi(x_i(t)||^2] \leq \sigma^2$$

- Nodes keep search points $x_i(t)$



Stochastic Mirror Descent

- (Centralized) SO is well understood
- Optimum convergence via **mirror descent**

Algorithm: Stochastic Mirror Descent

Initialize $x_i(0) \leftarrow 0$

for $t=1$ to T:

  $x_i(t) \leftarrow P_x[x_i(t-1) - \gamma_t\, g_i(t-1)]$

  $x^{av}_i(t) \leftarrow 1/t\, \boldsymbol{\Sigma}_\tau\, x_i(\tau)$

end for t

[Xiao, "Dual averaging methods for regularized stochastic learning and online optimization", 2010]

[Lan, "An Optimal Method for Stochastic Composite Optimization", 2012]

Stochastic Mirror Descent

- (Centralized) SO is well understood
- Optimum convergence via **mirror descent**

Algorithm: Stochastic Mirror Descent

Initialize $x_i(0) \leftarrow 0$

for t=1 to T:

$x_i(t) \leftarrow P_x[x_i(t\text{-}1) - \gamma_t\, g_i(t\text{-}1)]$

$x^{av}_i(t) \leftarrow 1/t\, \boldsymbol{\Sigma}_\tau\, x_i(\tau)$

end for t

- Extensions via Bregman divergences + prox mappings
- After T rounds:

$$E[\psi(\mathbf{x}_i^{\mathrm{av}}(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L}{T} + \frac{\sigma}{\sqrt{T}}\right]$$

[Xiao, "Dual averaging methods for regularized stochastic learning and online optimization", 2010]

[Lan, "An Optimal Method for Stochastic Composite Optimization", 2012]

Stochastic Mirror Descent

- Can speed up convergence via **accelerated stochastic mirror descent:**
- Similar SGD steps, but more complex iterate averaging
- After T rounds:

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\frac{L}{T^2} + \frac{\sigma}{\sqrt{T}} \right]$$

[Xiao, "Dual averaging methods for regularized stochastic learning and online optimization", 2010]

[Lan, "An Optimal Method for Stochastic Composite Optimization", 2012]

Stochastic Mirror Descent

- Can speed up convergence via **accelerated stochastic mirror descent:**
- Similar SGD steps, but more complex iterate averaging
- After T rounds:

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L}{T^2} + \frac{\sigma}{\sqrt{T}}\right]$$

- Optimum convergence order-wise
- Noise term dominates in general, but ASMD provides a **universal** solution to the SO problem

- Will prove significant in **distributed** stochastic learning

[Xiao, "Dual averaging methods for regularized stochastic learning and online optimization", 2010]
[Lan, "An Optimal Method for Stochastic Composite Optimization", 2012]

Back to Distributed Stochastic Learning

- With m nodes, after T rounds, the best possible performance is

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L}{(mT)^2} + \frac{\sigma}{\sqrt{mT}}\right]$$

Back to Distributed Stochastic Learning

- With m nodes, after T rounds, the best possible performance is

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L}{(mT)^2} + \frac{\sigma}{\sqrt{mT}}\right]$$

- Achievable with sufficiently fast communications
- In **distributed computing** environment, noise term is achievable via gradient averaging:
    1. Use `AllReduce` to average gradients over a spanning tree
    2. Take a SMD step
- Upshot: Averaging reduces gradient noise, provides speedup
- **Perfect** averages difficult to compute over wireless networks
- Approaches: average consensus, incremental methods, etc.

[Dekel et al., "Optimal distributed online prediction using mini-batches", 2012]

[Duchi et al., "Dual averaging for distributed optimization…", 2012]

[Ram et al., "Incremental stochastic sub-gradient algorithms for convex optimization", 2009]

Communications Model

- Nodes connected over an undirected graph $G = (V,E)$
- Every communications round, each node broadcasts a single gradient-like message $m_i(r)$ to its neighbors
- Rate limitations modeled by the **communications ratio** ρ
- ρ communications rounds for every data sample that arrives

Communications Model

- Nodes connected over an undirected graph $G = (V,E)$
- Every communications round, each node broadcasts a single gradient-like message $m_i(r)$ to its neighbors
- Rate limitations modeled by the **communications ratio** ρ
- ρ communications rounds for every data sample that arrives

<table>
<tr><td>$\xi_i(t=1)$</td><td>$\xi_i(t=2)$</td><td>$\xi_i(t=3)$</td><td>$\xi_i(t=4)$</td><td>data rounds</td></tr>
<tr><td colspan="2">$m_i(r=1)$</td><td colspan="2">$m_i(r=2)$</td><td>comms rounds</td></tr>
</table>

$\rho = 1/2$

<table>
<tr><td colspan="2">$\xi_i(t=1)$</td><td colspan="2">$\xi_i(t=2)$</td><td>data rounds</td></tr>
<tr><td>$m_i(r=1)$</td><td>$m_i(r=2)$</td><td>$m_i(r=3)$</td><td>$m_i(r=4)$</td><td>comms rounds</td></tr>
</table>

$\rho = 2$

Distributed Mirror Descent Outline

- Distribute stochastic MD via **averaging consensus**:
    1. Nodes obtain local gradients
    2. Compute distributed gradient averages via consensus
    3. Take MD step using the average gradients

$\xi_i(t=1)$		$\xi_i(t=2)$		data rounds
$m_i(r=1)$	$m_i(r=2)$	$m_i(r=3)$	$m_i(r=4)$	consensus rounds
$x_i(t=1)$		$x_i(t=2)$		search point updates

$\rho = 2$

Distributed Mirror Descent Outline

- Distribute stochastic MD via **averaging consensus**:
    1. Nodes obtain local gradients
    2. Compute distributed gradient averages via consensus
    3. Take MD step using the average gradients

$\xi_i(t=1)$		$\xi_i(t=2)$		data rounds
$m_i(r=1)$	$m_i(r=2)$	$m_i(r=3)$	$m_i(r=4)$	consensus rounds
$x_i(t=1)$		$x_i(t=2)$		search point updates

$\rho = 2$

- If links are slow (ρ small), there isn't much time for consensus
- New data samples arrives before the network can process the previous one

Mini-batching Gradients

- Solution: **mini-batch** together b gradients, batch size $b \geq 1$
- Hold search point constant for b rounds
- Average together b gradient evaluations:

$$\theta_i(s) = \frac{1}{b} \sum_{t=(s-1)b+1}^{sb} g_i(t)$$

- Reduces gradient noise: $E_\xi[||\Theta_i(s) - \nabla\psi(x_i(s)||^2] \leq \sigma^2/b$

Mini-batching Gradients

- Solution: **mini-batch** together b gradients, batch size $b \geq 1$
- Hold search point constant for b rounds
- Average together b gradient evaluations:

$$\theta_i(s) = \frac{1}{b} \sum_{t=(s-1)b+1}^{sb} g_i(t)$$

- Reduces gradient noise: $E_\xi[||\Theta_i(s) - \nabla\psi(x_i(s)||^2] \leq \sigma^2/b$
- Allows for more consensus rounds

$\xi_i(t=1)$	$\xi_i(t=2)$	$\xi_i(t=3)$	$\xi_i(t=4)$	$\xi_i(t=5)$	$\xi_i(t=6)$	$\xi_i(t=7)$	$\xi_i(t=8)$	data rounds
$m_i(r=1)$		$m_i(r=2)$		$m_i(r=3)$		$m_i(r=4)$		consensus rounds
$\Theta_i(s=1)$				$\Theta_i(s=2)$				mini-batch rounds
$x_i(t=1)$				$x_i(t=5)$				search points

$\rho = 1/2$, $b=4$

Mini-batching Gradients

- Solution: **mini-batch** together b gradients, batch size $b \geq 1$
- Hold search point constant for b rounds
- Average together b gradient evaluations:

$$\theta_i(s) = \frac{1}{b} \sum_{t=(s-1)b+1}^{sb} g_i(t)$$

- Reduces gradient noise: $E_\xi[||\Theta_i(s) - \nabla\psi(x_i(s)||^2] \leq \sigma^2/b$
- Allows for more consensus rounds

$\xi_i(t=1)$	$\xi_i(t=2)$	$\xi_i(t=3)$	$\xi_i(t=4)$	$\xi_i(t=5)$	$\xi_i(t=6)$	$\xi_i(t=7)$	$\xi_i(t=8)$	data rounds
$m_i(r=1)$		$m_i(r=2)$		$m_i(r=3)$		$m_i(r=4)$		consensus rounds
$\Theta_i(s=1)$				$\Theta_i(s=2)$				mini-batch rounds
$x_i(t=1)$				$x_i(t=5)$				search points

$\rho = 1/2$, $b=4$

- **However,** fewer search point updates

Gradient Averaging via Consensus

- **Averaging consensus**: nodes compute local averages with neighbors, which converge on the global average
- Choose a doubly-stochastic matrix $W \in \mathbb{R}^{m \mathrm{x} m}$ such that $w_{ij} \neq 0$ only if nodes are connected, i.e. $(i,j) \in E$
- At mini-batch round s and communications round r:

$$\theta_i^r(s) = \sum_{i,j} w_{ij} \theta_j^{r-1}(s)$$

- For mini-batch size b and communications ratio ρ, nodes can carry out bρ consensus rounds per mini-batch.
- Iterates converge on true average as # of rounds -> infinity

[Duchi et al., “Dual averaging for distributed optimization…”, 2012]

[Tsianos and Rabbat, “Efficient distributed online prediction and stochastic optimization”, 2016]

Gradient Averaging via Consensus

- At mini-batch round s and communications round r:

$$\theta_i^r(s) = \sum_{i,j} w_{ij}\theta_j^{r-1}(s)$$

Lemma: The equivalent gradient noise variance is bounded by

$$\sigma_{\text{eq}}^2 := E[||\theta_i^{\rho b}(s) - \nabla\psi(\mathbf{x}_i(s))||^2] \leq$$

$$O(1)\left[\lambda_2^{2\rho b}(W)||\mathbf{x}_i(s) - \mathbf{x}_j(s)||^2 + \frac{\lambda_2^{2\rho b}(W)\sigma^2}{b} + \frac{\sigma^2}{mb}\right]$$

Gradient Averaging via Consensus

- At mini-batch round s and communications round r:

$$\theta_i^r(s) = \sum_{i,j} w_{ij} \theta_j^{r-1}(s)$$

Lemma: The equivalent gradient noise variance is bounded by

$$\sigma_{\text{eq}}^2 := E[||\theta_i^{\rho b}(s) - \nabla\psi(\mathbf{x}_i(s))||^2] \leq$$

$$O(1)\left[\lambda_2^{2\rho b}(W)||\mathbf{x}_i(s) - \mathbf{x}_j(s)||^2 + \frac{\lambda_2^{2\rho b}(W)\sigma^2}{b} + \frac{\sigma^2}{mb}\right]$$

- Noise components: gap in nodes' search points, error due to imperfect consensus averaging, residual noise
- For ρ or b large, noise converges on perfect-average case

Distributed SA Mirror Descent

Algorithm: Distributed Stochastic Approximation Mirror Descent (D-SAMD)

Initialize $x_i(0) \leftarrow 0$, for all i

for $s=1$ to T/b: [iterate over mini-batches]

  $\theta^0{}_i(s) \leftarrow \theta_i(s)$

  for $r=1$ to ρb: [iterate over consensus rounds]

    $\theta^r{}_i(s) = \Sigma_j w_{ij} \theta^{r-1}{}_i(s)$, for all i

  end for r

  $x_i(sb+1) \leftarrow P_x[x_i(sb) - \gamma_s \theta^{\rho b}{}_i(s)]$

  $x^{av}{}_i(t) \leftarrow 1/s \Sigma_\tau x_i(\tau b)$

end for s

- Outer loop: nodes compute mini-batches, take MD steps
- Inner loop: nodes engage in average consensus

D-SAMD Convergence Analysis

- Recall that Mirror Descent has convergence rate:

$$E[\psi(\mathbf{x}_i^{\mathrm{av}}(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\frac{L}{T} + \frac{\sigma}{\sqrt{T}} \right]$$

D-SAMD Convergence Analysis

- Recall that Mirror Descent has convergence rate:

$$E[\psi(\mathbf{x}_i^{\text{av}}(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\frac{L}{T} + \frac{\sigma}{\sqrt{T}}\right]$$

- With mini-batch size b and equivalent gradient noise σ^2_{eq}, D-SAMD has

$$E[\psi(\mathbf{x}_i^{\text{av}}(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\frac{Lb}{T} + \sqrt{\frac{\sigma_{\text{eq}}^2 b}{T}}\right]$$

$$\sigma_{\text{eq}}^2 = O(1) \left[\lambda_2^{2\rho b}(W)||\mathbf{x}_i(s) - \mathbf{x}_j(s)||^2 + \frac{\lambda_2^{2\rho b}(W)\sigma^2}{b} + \frac{\sigma^2}{mb}\right]$$

D-SAMD Convergence Analysis

- Recall that Mirror Descent has convergence rate:

$$E[\psi(\mathbf{x}_i^{\mathrm{av}}(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L}{T} + \frac{\sigma}{\sqrt{T}}\right]$$

- With mini-batch size b and equivalent gradient noise σ^2_{eq}, D-SAMD has

$$E[\psi(\mathbf{x}_i^{\mathrm{av}}(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{Lb}{T} + \sqrt{\frac{\sigma_{\mathrm{eq}}^2 b}{T}}\right]$$

$$\sigma_{\mathrm{eq}}^2 = O(1)\left[\lambda_2^{2\rho b}(W)||\mathbf{x}_i(s) - \mathbf{x}_j(s)||^2 + \frac{\lambda_2^{2\rho b}(W)\sigma^2}{b} + \frac{\sigma^2}{mb}\right]$$

- Need to choose b big enough to ensure:
  1. Nodes' iterates don't diverge
  2. Equivalent noise variance is on par with residual noise variance

D-SAMD Convergence Analysis

Lemma: D-SAMD iterates are guaranteed to converge provided

$$b \geq O(1) \left[1 + \frac{\log(mT)}{\rho \log(1/\lambda_2(W))} \right]$$

Furthermore, this condition is sufficient to ensure that

$$\sigma_{\text{eq}}^2 \leq O(1) \sqrt{\frac{\sigma^2}{mT}}$$

D-SAMD Convergence Analysis

Lemma: D-SAMD iterates are guaranteed to converge provided

$$b \geq O(1)\left[1 + \frac{\log(mT)}{\rho \log(1/\lambda_2(W))}\right]$$

Furthermore, this condition is sufficient to ensure that

$$\sigma_{\text{eq}}^2 \leq O(1)\sqrt{\frac{\sigma^2}{mT}}$$

- Results in convergence rate

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L\log(mT)}{\rho \log(1/\lambda_2(W))T} + \sqrt{\frac{\sigma^2}{mT}}\right]$$

- When is this order optimum?

D-SAMD Convergence Analysis

Theorem: If

$$\rho \geq O(1) \left[\frac{m^{1/2} \log(mT)}{\sigma T^{1/2} \log(1/\lambda_2(W))} \right]$$

Then the conditions of the previous lemma ensure that

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\sqrt{\frac{\sigma^2}{mT}} \right]$$

D-SAMD Convergence Analysis

Theorem: If

$$\rho \geq O(1) \left[\frac{m^{1/2} \log(mT)}{\sigma T^{1/2} \log(1/\lambda_2(W))} \right]$$

Then the conditions of the previous lemma ensure that

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\sqrt{\frac{\sigma^2}{mT}} \right]$$

- Larger mini-batches decreases gradient noise, but also decreases the number of MD steps taken
- Eventually, the deterministic term dominates the convergence rate

- Natural idea: use **accelerated** mirror descent

Accelerated Distributed SA Mirror Descent

- Recall: accelerated MD takes similar projected gradient descent steps, uses more complicated averaging scheme

Algorithm: Accelerated Distributed Stochastic Approximation Mirror Descent (AD-SAMD) [simplified]

for s=1 to T/b: [iterate over mini-batches]
  compute mini-batch gradients
  for r=1 to ρb:
    perform consensus iterations on gradients
  end for r
  perform accelerated MD updates
end for s

AD-SAMD Convergence Analysis

- With mini-batch size b and equivalent gradient noise σ^2_{eq}, AD-SAMD has

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{Lb^2}{T^2} + \sqrt{\frac{\sigma_{\text{eq}}^2 b}{T}}\right]$$

- The equivalent gradient noise has approx. the same variance:

$$\sigma_{\text{eq}}^2 = O(1)\left[\lambda^{2\rho b}||\mathbf{x}_i(s) - \mathbf{x}_j(s)||^2 + \frac{\lambda^{2\rho b}\sigma^2}{b} + \frac{\sigma^2}{mb}\right]$$

AD-SAMD Convergence Analysis

- With mini-batch size b and equivalent gradient noise σ^2_{eq}, AD-SAMD has

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{Lb^2}{T^2} + \sqrt{\frac{\sigma_{\text{eq}}^2 b}{T}}\right]$$

- The equivalent gradient noise has approx. the same variance:

$$\sigma_{\text{eq}}^2 = O(1)\left[\lambda^{2\rho b}||\mathbf{x}_i(s) - \mathbf{x}_j(s)||^2 + \frac{\lambda^{2\rho b}\sigma^2}{b} + \frac{\sigma^2}{mb}\right]$$

Lemma: AD-SAMD iterates are guaranteed to converge, and σ^2_{eq} has optimum scaling, provided

$$b \geq O(1)\left[1 + \frac{\log(mT)}{\rho\log(1/\lambda_2(W))}\right]$$

AD-SAMD Convergence Analysis

- Results in a convergence rate

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L\log^2(mT)}{\rho^2\log^2(1/\lambda_2(W))T^2} + \sqrt{\frac{\sigma^2}{mT}}\right]$$

AD-SAMD Convergence Analysis

- Results in a convergence rate

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L\log^2(mT)}{\rho^2\log^2(1/\lambda_2(W))T^2} + \sqrt{\frac{\sigma^2}{mT}}\right]$$

Theorem: If

$$\rho \geq O(1)\left[\frac{m^{1/4}\log(mT)}{\sigma T^{3/4}\log(1/\lambda_2(W))}\right]$$

Then the conditions of the previous lemma ensure that

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\sqrt{\frac{\sigma^2}{mT}}\right]$$

AD-SAMD Convergence Analysis

- Results in a convergence rate

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\frac{L\log^2(mT)}{\rho^2\log^2(1/\lambda_2(W))T^2} + \sqrt{\frac{\sigma^2}{mT}}\right]$$

Theorem: If

$$\rho \geq O(1)\left[\frac{m^{1/4}\log(mT)}{\sigma T^{3/4}\log(1/\lambda_2(W))}\right]$$

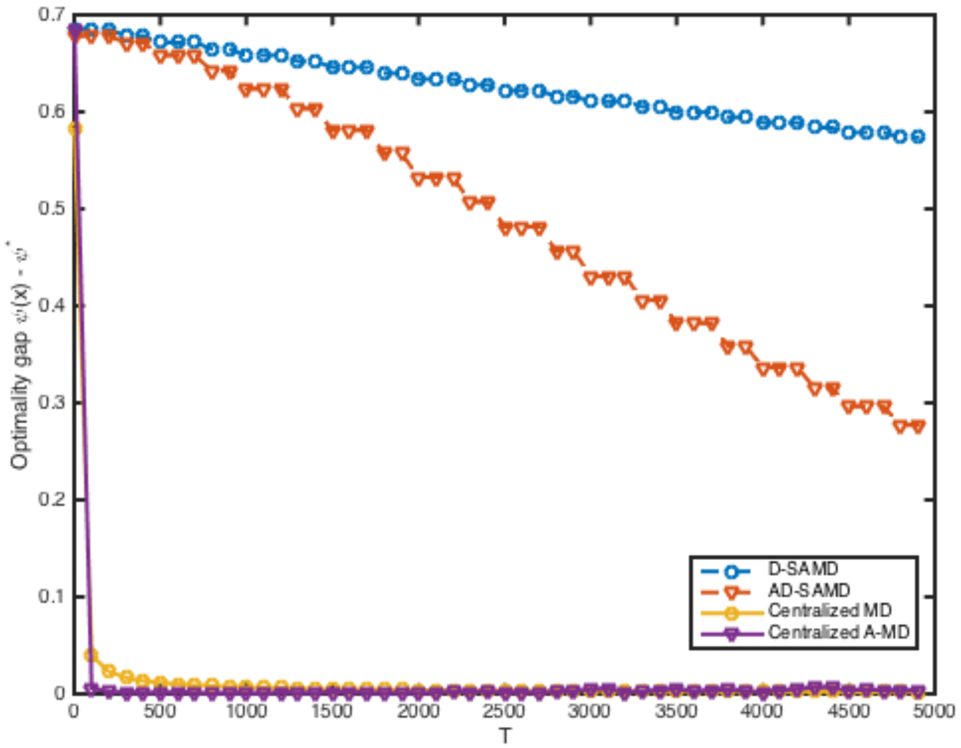
Then the conditions of the previous lemma ensure that

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1)\left[\sqrt{\frac{\sigma^2}{mT}}\right]$$

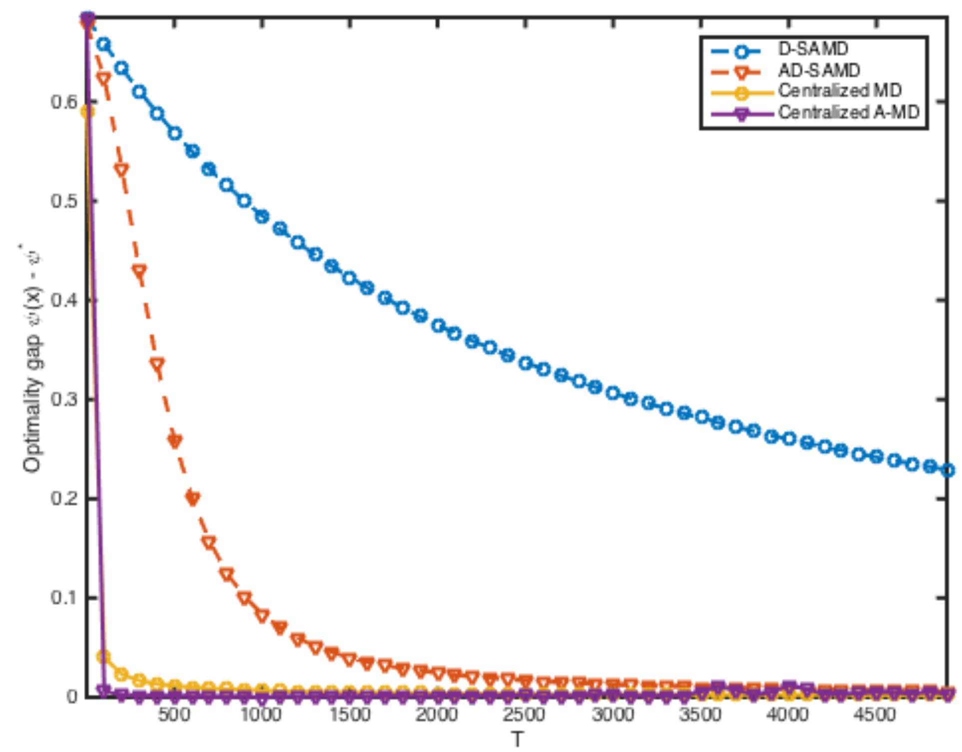
- AD-SAMD permits more aggressive mini-batching
- Improvement of 1/4 in the exponents of m and T

Numerical example: Logistic Regression

- Logistic regression: learn a binary classifier from streams of input data
- Measurements are Gaussian-distributed, unknown mean, d=50
- Network drawn from Erdos-Reyni model with m=20
- Log-loss cost function



(a) $\rho = 1$



(b) $\rho = 10$

Composite Optimization

- What if objective is not smooth?
- Composite convex optimization:

$$\psi(x) = f(x) + h(x)$$

- f(x) has Lipschitz gradients, but h(x) is only Lipschitz:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$
$$||h(x) - h(y)|| \leq \mathcal{M}||x - y||$$

- Accelerated MD via **subgradients** gives the optimum convergence

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\frac{L}{T^2} + \frac{\mathcal{M} + \sigma}{\sqrt{T}} \right]$$

Composite Optimization

- Small perturbations lead to significant deviations in subgradients
- Two new challenges:
    1. Mini-batching doesn't help – gradient noise variance doesn't matter!
    2. Imperfect average consensus results in a "noise floor"
- Results in sub-optimum convergence rates:

$$E[\psi(\mathbf{x}_i(T)) - \psi(\mathbf{x}^*)] \leq O(1) \left[\frac{Lb^2}{T^2} + \frac{\mathcal{M} + \sigma/\sqrt{mb}}{\sqrt{T/b}} + \mathcal{M} \right]$$

Conclusions

Summary:

- Investigated stochastic learning from the perspective of rate-limited, wireless links
- Developed two schemes, D-SAMD and AD-SAMD, that balance in-network gradient averaging and local mini-batching
- Derived conditions for order-optimum convergence

Future work:

- Optimum distributed SO for composite objectives
- Can we improve the convergence rates of AD-SAMD?
- Other communications issues: delay, quantization, etc.

Preprint available: https://arxiv.org/abs/1704.07888