Towards Scalable Algorithms for Distributed Optimization and Learning

César A. Uribe

RICE ENGINEERING
Electrical and Computer Engineering

Motivation

(a) Sensor Networks in Agriculture

(b) (Mis)information Spread

(c) Camera Networks for Security

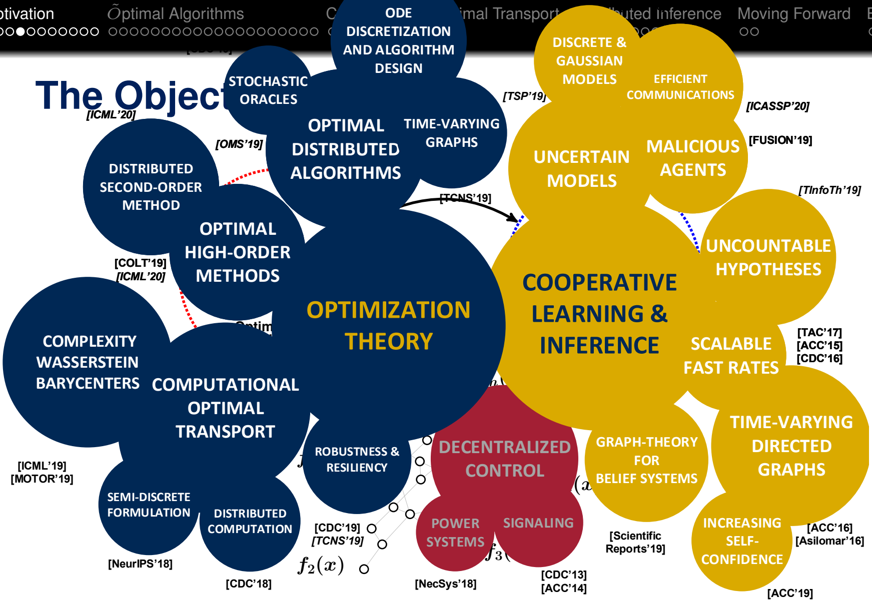
(d) Huge-scale ML

Characteristics and *Challenges*

- Characteristics
  - **Many** components/units (we call them *agents*).
  - **Connected** over networks.
  - Cyber and Physical **interactions**.
  - Distributed Storage.
- Challenges
  - **Decentralization**: distributed computations.
  - **Scalability**: Price of decentralization.
  - **Optimality**: Efficiency & Performance.

The Scalability Issue

The Object...



DECENTRALIZED **SCALABLE** **OPTIMAL**

The abstraction model

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x)$$

The abstraction model

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(\underbrace{x}_{\text{decision}})$$

The abstraction model

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \overbrace{f_i(x)}^{\text{cost of agent } i}$$

The abstraction model

$$\min_{x \in \mathbb{R}^n} \underbrace{\sum_{i=1}^{m} f_i(x)}_{\text{total cost}}$$

The abstraction model

$$\underbrace{\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x)}_{\text{make the best decision}}$$

Prototypical Problem: Risk Minimization

A general formulation of the learning problem, where, h_θ is some loss function.

$$\min_\theta R(h_\theta, P) \triangleq \mathbb{E}_{(X,Y)\sim P}\left[\ell(h_\theta(X), Y)\right]$$

However, in general we do not know the joint distribution P.

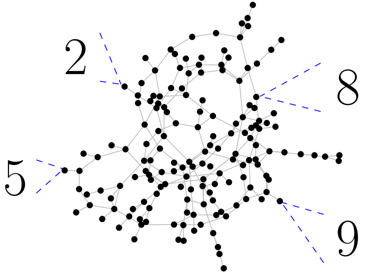
Empirical Risk Minimization

Assuming some finite number of data points m then we can solve the approximate problem assuming the empirical distribution.

$$\min_{\theta} R_m(h_\theta, \hat{P}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

Distributed Average Consensus 101

- There is a network of m agents, i.e., a graph $\mathcal{G} = \{V, E\}$.
- Agent i holds an initial value $x_0^i \in \mathbb{R}$.
- Each agent needs to distributedly compute $\frac{1}{m}\sum_{i=1}^{m} x_0^i$.



Equivalently, solve $\min_{x \in \mathbb{R}} \frac{1}{2} \sum_{i=1}^{m} \|x - x_i\|_2^2$

Enter the Consensus Algorithm

$$x_{k+1}^i = \sum_{j=1}^{m} [A]_{ij} x_k^j \tag{1}$$

FUNDAMENTAL RESULT: If $\mathcal{G}$ is connected, undirected and static, and A is doubly stochastic, where $[A]_{ij} > 0$ iff $(j,i) \in E$. Then, the iterates generated by (1) have the following property:

$$\lim_{k\to\infty} x_k^i = \frac{1}{m}\sum_{j=1}^{m} x_0^j \qquad \forall i \in V.$$

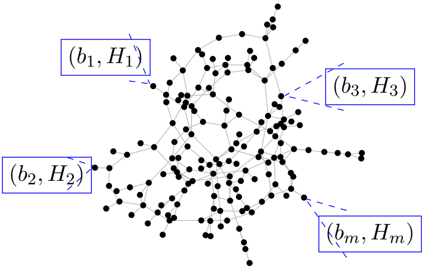
An Example: Distributed Ridge Regression

We want to estimate x assuming

$$b_i = H_i x + noise,$$

where

- $H_i \in \mathbb{R}^{d_i \times n}$: d_i data points of dimension n.
- $b_i \in \mathbb{R}^{d_i}$: d_i outputs.



$$\min_x \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} \|b_i - H_i x\|_2^2.$$

Today I'm going to talk about:

$\tilde{O}$ptimal Algorithms for (Distributed) Optimization

- **CAU**, S. Lee, A. Gasnikov, and A. Nedic, "A Dual Approach for Optimal Algorithms in Distributed Optimization over Networks," 2018
- A. Rogozin, **CAU**, A. Gasnikov, N. Malkovsky, and A. Nedic, "Optimal distributed convex optimization on slowly time-varying graphs," IEEE Transactions on Control of Network Systems, 2019
- A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and **CAU**, "Optimal tensor methods in smooth convex and uniformly convex optimization," in COLT 2019.

CASE 1: Computational Optimal Transport

- **CAU**, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedic, "Distributed computation of Wasserstein barycenters over networks," in IEEE Conference on Decision and Control, 2018.
- A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and **CAU**, "On the complexity of approximating Wasserstein barycenters," in ICML 2019.
- P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, **CAU**, and A. Nedić, "Decentralize and randomize: Faster algorithm for Wasserstein barycenters," Neurips 2018

CASE 2: Social Learning and Distributed Inference

- A. Nedic, A. Olshevsky, and **CAU**, "Fast Convergence Rates for Distributed Non-Bayesian Learning," IEEE Transactions on Automatic Control, 2017.
- A. Nedic, A. Olshevsky, and **CAU**, "Distributed learning for cooperative inference," 2017.
- J. Z. Hare, **CAU**, L. Kaplan, and A. Jadbabaie, "Non-Bayesian social learning with uncertain models," 2019

Oracle calls and complexity bounds

Consider the generic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

and assume that f is convex and

$$\|\nabla^p f(x) - \nabla^p f(y)\|_2 \leq M_p \|x - y\|_2 \qquad \forall x, y \in \mathbb{R}^n.$$

Calling the oracle: Query $\{f(x), \nabla f(x), \ldots, \nabla^p f(x)\}$ at a certain point x.

Oracle complexity: For a given $\varepsilon > 0$, how many oracle calls are required to obtain a point $\hat{x}$ such that

$$f(\hat{x}) - f^* \leq \varepsilon,$$

where f^* is an optimal function value.

Oracle calls and complexity bounds

Consider the generic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

and assume that f is convex and

$$||\nabla^p f(x) - \nabla^p f(y)||_2 \leq M_p ||x - y||_2 \qquad \forall x, y \in \mathbb{R}^n.$$

Calling the oracle: Query $\{f(x), \nabla f(x), \ldots, \nabla^p f(x)\}$ at a certain point x.

Oracle complexity: For a given $\varepsilon > 0$, how many oracle calls are required to obtain a point $\hat{x}$ such that

$$f(\hat{x}) - f^* \leq \varepsilon,$$

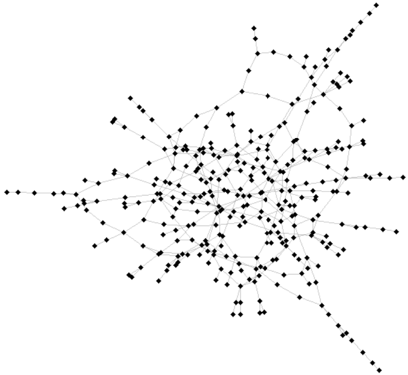
where f^* is an optimal function value.

The complexity of solving *smooth* optimization problems

	Lower Bound	Upper Bound
$p=1$	$\Omega\left(\left(\frac{M_1 R^2}{\varepsilon}\right)^{\frac{1}{2}}\right)$ [Nemirovski, Yudin (1983)]	$O\left(\left(\frac{M_1 R^2}{\varepsilon}\right)^{\frac{1}{2}}\right)$ [Nesterov (1983)]
$p=2$	$\Omega\left(\left(\frac{M_2 R^3}{\varepsilon}\right)^{\frac{2}{7}}\right)$ [Arjevani et al. (2018)]	$\tilde{O}\left(\left(\frac{M_2 R^3}{\varepsilon}\right)^{\frac{2}{7}}\right)$ [Monteiro, Svaiter (2013)]
$p\geq 3$	$\Omega\left(\left(\frac{M_p R^{p+1}}{\epsilon}\right)^{\frac{2}{3p+1}}\right)$ [Arjevani et al. (2018)] [Nesterov (2018a)]	$O\left(\left(\frac{M_p R^{p+1}}{\epsilon}\right)^{\frac{1}{p+1}}\right)$ [Baes (2009)] [Wibisono et al. (2016)] [Nesterov, (2018a)]
$p\geq 3$		$\tilde{O}\left(\left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}}\right)$ [Gasnikov et al. (2019)]

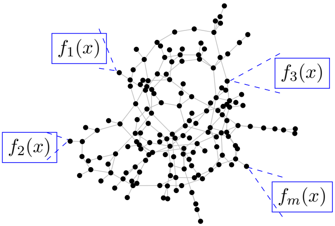
where $R = \|x_0 - x^*\|_2^2$.

How to take into account the distributed information and the network architecture?



The Distributed Optimization Setup

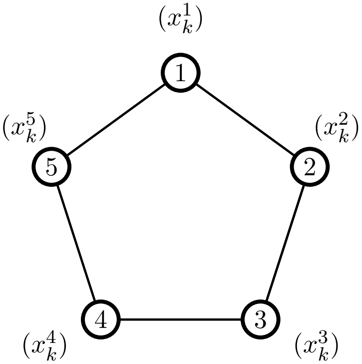
$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x) \qquad (2)$$



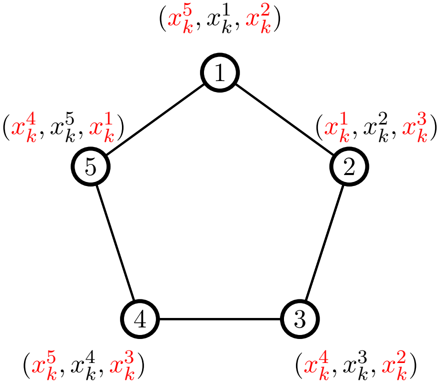
- Each node knows $f_i(x)$ (convex).
- Agents communicate over a graph $\mathcal{G} = (V, E)$.
- Agents $j \in V$ shares information with $i \in V$ if $(j, i) \in E$.

Objective: Solve (2) distributedly using local information only.

What does sharing information mean?



What does sharing information mean?

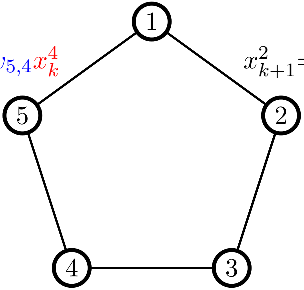


What does sharing information mean?

$$x^1_{k+1} = w_{1,5} x^5_k + w_{1,1} x^1_k + w_{1,2} x^2_k$$

$$x^5_{k+1} = w_{5,1} x^1_k + w_{5,5} x^5_k + w_{5,4} x^4_k$$

$$x^2_{k+1} = w_{2,1} x^1_k + w_{2,2} x^2_k + w_{2,3} x^3_k$$



$$x^4_{k+1} = w_{4,5} x^5_k + w_{4,4} x^4_k + w_{4,3} x^3_k$$

$$x^3_{k+1} = w_{3,4} x^4_k + w_{3,3} x^3_k + w_{3,2} x^2_k$$

What does sharing information mean?

$$\begin{bmatrix} x_{k+1}^1 \\ x_{k+1}^2 \\ x_{k+1}^3 \\ x_{k+1}^4 \\ x_{k+1}^5 \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & 0 & 0 & w_{1,5} \\ w_{1,1} & w_{2,2} & w_{2,3} & 0 & 0 \\ 0 & w_{3,2} & w_{3,3} & w_{3,4} & 0 \\ 0 & 0 & w_{4,3} & w_{4,4} & w_{4,5} \\ w_{5,1} & 0 & 0 & w_{5,4} & w_{5,5} \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \\ x_k^3 \\ x_k^4 \\ x_k^5 \end{bmatrix}$$

$$x_{k+1} = W x_k, \qquad \text{or} \qquad x_{k+1}^i = \sum_{i=1}^m w_{i,j} x_k^j$$

where W has the sparsity pattern of the graph.

(Lack of) Optimality in Distributed Optimization

Local oracles: Agent i queries $\{f_i(x^i), \nabla f_i(x^i), \dots, \nabla^p f_i(x^i)\}$ at a certain point x^i only.

E.g., No agent has access to a full gradient $\sum_{i=1}^m \nabla f_i(x^i)$

1. Each agent runs a local algorithm only,

$$x_{k+1}^i = x_k^i - \alpha_i \nabla f_i(x_k^i)$$

2. Rule of thumb, distributed gradient descent [Nedić-Ozdaglar, 2009]

$$x_{k+1}^i = \sum_{j=1}^m w_{ij} x_k^j - \alpha_i \nabla f_i(x_k^i)$$

(Lack of) Optimality in Distributed Optimization

Local oracles: Agent i queries $\{f_i(x^i), \nabla f_i(x^i), \ldots, \nabla^p f_i(x^i)\}$ at a certain point x^i only.

E.g., No agent has access to a full gradient $\sum_{i=1}^{m} \nabla f_i(x^i)$

1. Each agent runs a local algorithm only,

$$x_{k+1}^i = x_k^i - \alpha_i \nabla f_i(x_k^i), \qquad O\left(\varepsilon^{-1}\right)$$

2. Rule of thumb, distributed gradient descent [Nedić-Ozdaglar, 2009]

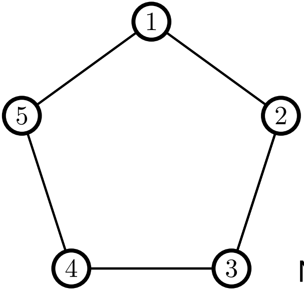
$$x_{k+1}^i = \sum_{j=1}^{m} w_{ij} x_k^j - \alpha_i \nabla f_i(x_k^i), \qquad O\left(\varepsilon^{-2}\right)$$

A map of Distributed Complexity Bounds

Approach	Reference	μ-strongly convex and L-smooth	μ-strongly convex	L-smooth	M-Lipschitz
Centralized	[Nemirovskii and Yudin, 1983]	$\sqrt{\frac{L}{\mu}}$	$\frac{M^2}{\mu\varepsilon}$	$\sqrt{\frac{L}{\varepsilon}}$	$\frac{M^2}{\varepsilon^2}$
Gradient Computations	[Qu and Li, 2017][b]	$m^3\left(\frac{L}{\mu}\right)^{5/7}$	–	$\frac{1}{\varepsilon^{5/7}}$	–
	[Olshevsky, 2014]	–	–	–	$m\frac{M^2}{\varepsilon^2}$
	[Duchi et al., 2012]	–	–	–	$m^2\frac{M^2}{\varepsilon^2}$
	[Doan and Olshevsky, 2017]	$m^2\frac{L}{\mu}$	–	–	–
	[Lakshmanan and De Farias, 2008]	–	–	$m^3\frac{L}{\varepsilon}$	–
	[Necoara, 2013]	$m^4\frac{L}{\mu}$	–	$m^4\frac{L}{\varepsilon}$	–
	[Jakovetic, 2017][c]	$m^2\sqrt{\frac{L}{\mu}}$	–	–	–
Communication Rounds	[Scaman et al., 2017]	$m\sqrt{\frac{L}{\mu}}$	–	–	–
	[Lan et al., 2017]	–	$m^2\sqrt{\frac{M^2}{\mu\varepsilon}}$	–	$m^2\frac{M}{\varepsilon}$
	[Uribe et al. 2018]	$m\sqrt{\frac{L}{\mu}}$	$m\sqrt{\frac{M^2}{\mu\varepsilon}}$	$m\sqrt{\frac{L}{\varepsilon}}$	$m\frac{M}{\varepsilon}$

[b] An iteration complexity of $\tilde{O}(\sqrt{1/\varepsilon})$ is shown if the objective is the composition of a linear map and a strongly convex and smooth function. Moreover, no explicit dependence on L and m is provided.
[c] A linear dependence on m is achieved if L is sufficiently close to μ.

Graph Laplacian



$$\bar{W} = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

Note that:

- $Wx = 0$ if and only if $x_1 = \ldots = x_m$.
- $\sqrt{W}x = 0$ if and only if $x_1 = \ldots = x_m$.

Problem Reformulation

$$x = \begin{bmatrix} x_1 \in R^n \\ x_2 \in R^n \\ \vdots \\ x_m \in R^n \end{bmatrix}$$

Rewrite problem (2) in an equivalent form as follows:

$$\min_{\sqrt{W}x=0} F(x) \qquad \text{where} \qquad F(x) \triangleq \sum_{i=1}^{m} f_i(x_i), \tag{3}$$

where $W = \bar{W} \otimes I_n$.

The analysis tools

Initially, consider the general problem

$$\min_{Ax=0} f(x). \tag{4}$$

We assume that the problem has optimal solutions.
Later, we will derive the specific results when

$$A = \sqrt{W} \qquad \text{and} \qquad f(x) = \sum_{i=1}^{m} f_i(x_i)$$

Approximate Solution Definition A point $x \in \mathbb{R}^{mn}$ is said to be an $(\varepsilon, \tilde{\varepsilon})$-solution of (9) if the following conditions are satisfied:

$$f(x) - f^* \leq \varepsilon \qquad \text{and} \qquad \|Ax\|_2 \leq \tilde{\varepsilon},$$

where f^* denotes the optimal value of (9).

Construction of the dual problem

The Lagrangian dual for the problem in (9) is given by

$$\min_{Ax=0} f(x) = \max_{y} \left\{ \min_{x} \left\{ f(x) - \langle A^T y, x \rangle \right\} \right\},$$

or equivalently

$$\min_{y} \varphi(y) \ \text{ where } \ \varphi(y) \triangleq \max_{x} \left\{ \langle A^T y, x \rangle - f(x) \right\},$$

where $\nabla\varphi(y) = Ax^*(A^T y)$ with

$$x^*(A^T y) = \arg\max_{x} \left\{ \langle A^T y, x \rangle - f(x) \right\}.$$

We say that f is **dual friendly** *when we can determine a solution of the preceding problem efficiently (in a closed form ideally)*

The duality of strong convexity and smoothness [Kakade et al., 2009]

- $f(x)$ is μ-strongly convex $\iff$ $\varphi(y)$ is L_φ-smooth with $L_\varphi = \lambda_{\max}(A^T A)/\mu$.
- $f(x)$ is L-smooth $\iff$ $\varphi(y)$ is μ_φ-strongly convex on the range space of A with $\mu_\varphi = \lambda^+_{\min}(A^T A)/L$.

The dual problem

$$\min_y \varphi(y) \text{ where } \varphi(y) \triangleq \max_x \left\{ \langle A^T y, x \rangle - f(x) \right\},$$

may have multiple solutions of the form $y^* + \ker(A^T)$ when the matrix A does not have a full row rank. When the solution is not unique, we *will use y^* to denote the smallest norm solution*, and we let R be its norm, i.e. $R = \|y^*\|_2$.

Remark

The dual problem

$$\min_y \varphi(y) \;\text{ where }\; \varphi(y) \triangleq \max_x \left\{ \left\langle A^T y, x \right\rangle - f(x) \right\},$$

is not strongly convex on the whole space.

Choosing $y_0 = \tilde{y}_0 = 0$ generates iterates that lie in the linear space of gradients $\nabla\varphi(y)$, which are of the form Ax.

The dual function $\varphi(y)$ is strongly convex when y is restricted to the linear space spanned by the range of the matrix A.

Nesterov's Fast Gradient Method (FGM) on the dual problem

Assume $\varphi(y)$ is μ-strongly convex and L-smooth.

$$x^*(A^T\tilde{y}_k) = \arg\max_x \left\{ \left\langle A^T\tilde{y}_k, x \right\rangle - f(x) \right\} \tag{5a}$$

$$y_{k+1} = \tilde{y}_k - \frac{1}{L_\varphi} A x^*(A^T\tilde{y}_k), \tag{5b}$$

$$\tilde{y}_{k+1} = y_{k+1} + \frac{\sqrt{L_\varphi} - \sqrt{\mu_\varphi}}{\sqrt{L_\varphi} + \sqrt{\mu_\varphi}}(y_{k+1} - y_k). \tag{5c}$$

and

$$\varphi(y_k) - \varphi^* \le L_\varphi \left(1 - \sqrt{\frac{\mu_\varphi}{L_\varphi}}\right)^k \|y_0 - y^*\|_2^2, \tag{6}$$

Distributed Nesterov's Fast Gradient Method: DFGM

Set $A = \sqrt{W}$, $z_k = \sqrt{W} y_k$ and $\tilde{z}_k = \sqrt{W} \tilde{y}_k$

$$x_i^*(\tilde{z}_k^i) = \arg\max_{x_i} \left\{ \left\langle \tilde{z}_k^i, x_i \right\rangle - f_i(x_i) \right\}$$

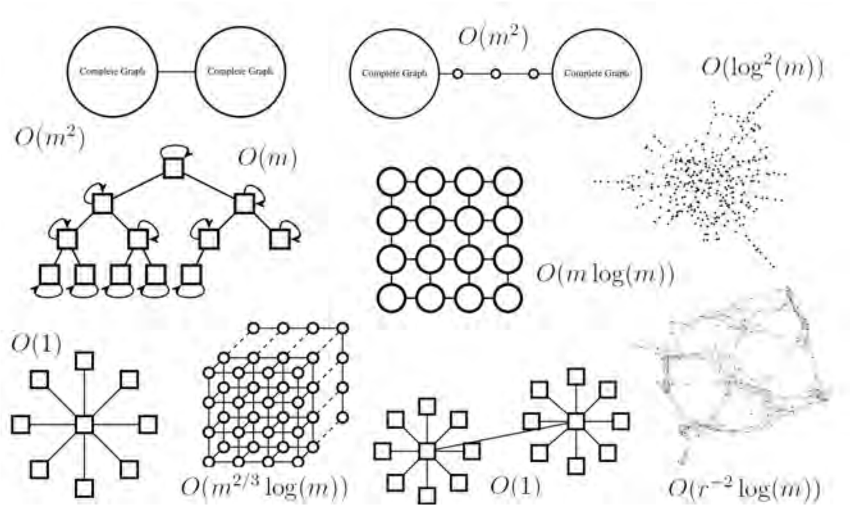
$$z_{k+1}^i = \tilde{z}_k^i - \frac{\mu}{\lambda_{\max}(W)} \sum_{j=1}^m W_{ij} x_j^*(\tilde{z}_k^j)$$

$$\tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{\lambda_{\max}(W)/\mu} - \sqrt{\lambda_{\min}^+(W)/L}}{\sqrt{\lambda_{\max}(W)/\mu} + \sqrt{\lambda_{\min}^+(W)/L}} (z_{k+1}^i - z_k^i)$$

A summary of results from [Uribe et al. 2018]

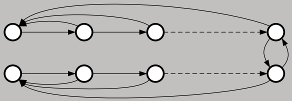
Property of $F(x)$	Oracle calls
μ-strongly convex and L-smooth	$\tilde{O}\left(\sqrt{\frac{L}{\mu}\chi(W)}\right)$
μ-strongly convex and M-Lipschitz*	$\tilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}\chi(W)}\right)$
L-smooth	$\tilde{O}\left(\sqrt{\frac{LR_x^2}{\varepsilon}\chi(W)}\right)$
M-Lipschitz	$\tilde{O}\left(\sqrt{\frac{M^2R_x^2}{\varepsilon^2}\chi(W)}\right)$

where $\chi(W) = \lambda_{\max}(W)/\lambda^{+}_{\min}(W)$.
The worst case for fixed undirected graphs is $\chi(W) = O(m^2)$ [Olshevsky, 2014].


Complete Graph
Complete Graph
Complete Graph
Complete Graph
$O(m^2)$
$O(\log^2(m))$
$O(m^2)$
$O(m)$
$O(m\log(m))$
$O(1)$
$O(m^{2/3}\log(m))$
$O(1)$
$O(r^{-2}\log(m))$

Challenges Moving Forward:

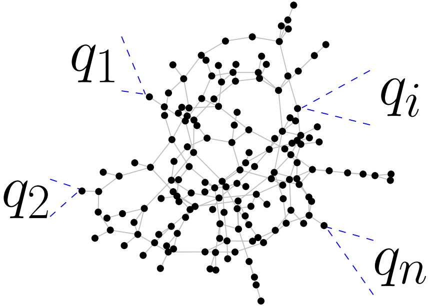
- **A search for an universal algorithm:** Typically, L, μ, R are unknown. Can we design an adaptive algorithm with optimal complexity with minimal information?
- **Scalable algorithms for directed graph:** The graph Laplacian is not symmetric, condition numbers can grow as $O(m^m)$ worst case.



- **Closer to real-world networks:** How to design optimal algorithms for **stochastic**, **asynchronous**, **time-varying**, **capacity-constrained** graphs.

Example: Distributed Computation of Wasserstein Barycenters

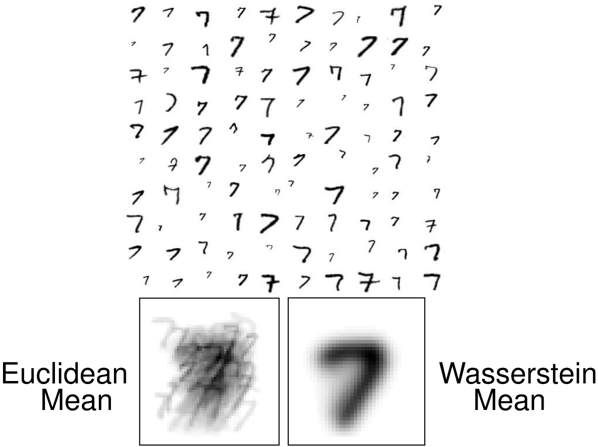
Now, what if each node holds a probability measure instead?

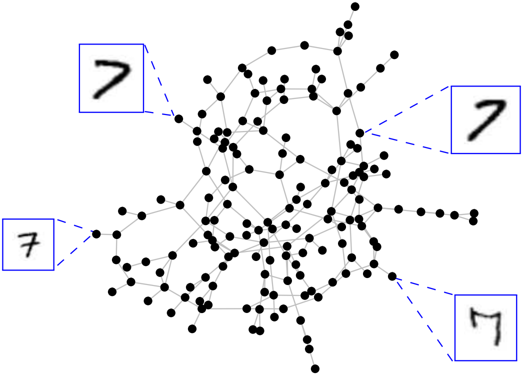


$$\min_{p \in S_1(n)} \sum_{i=1}^{m} \mathcal{W}_\gamma(p, q_i).$$

link

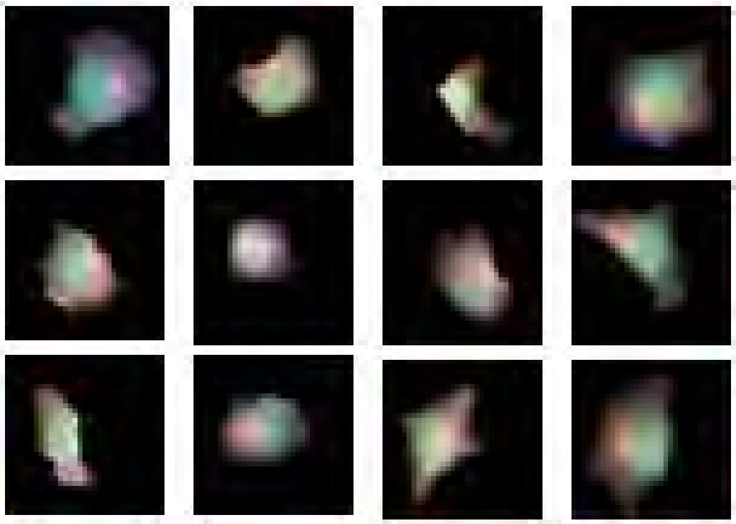
The Wasserstein Barycenters Problem:



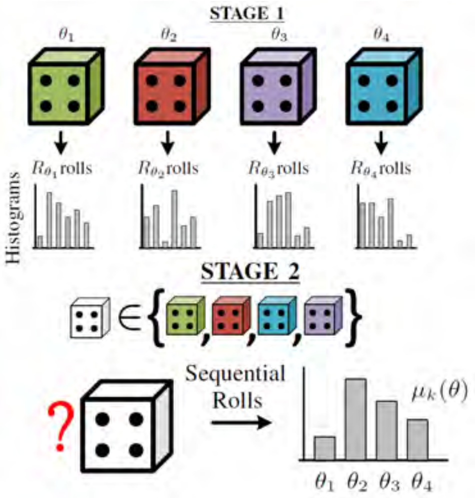


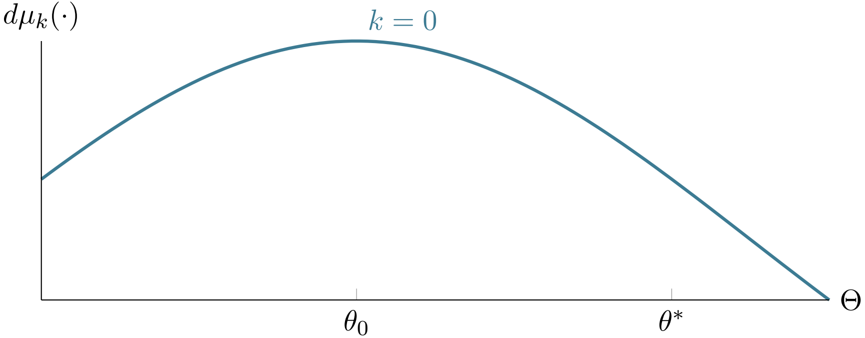
Click!

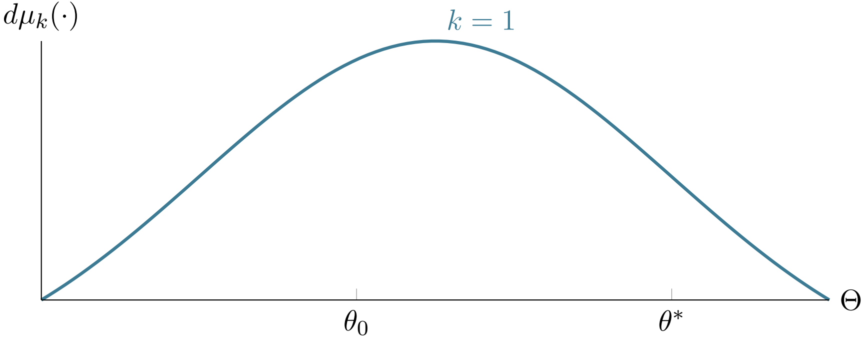
Click!

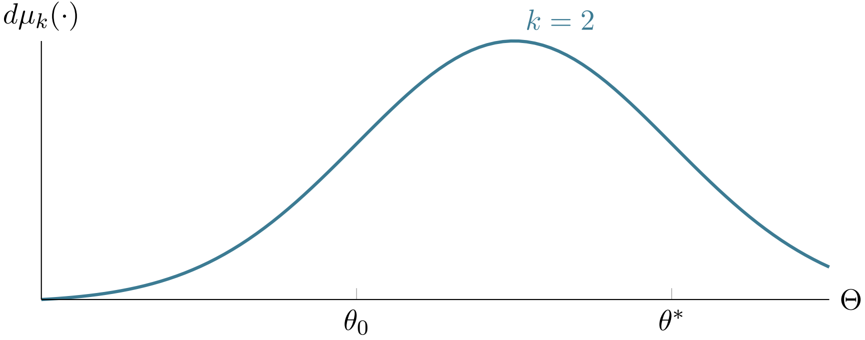


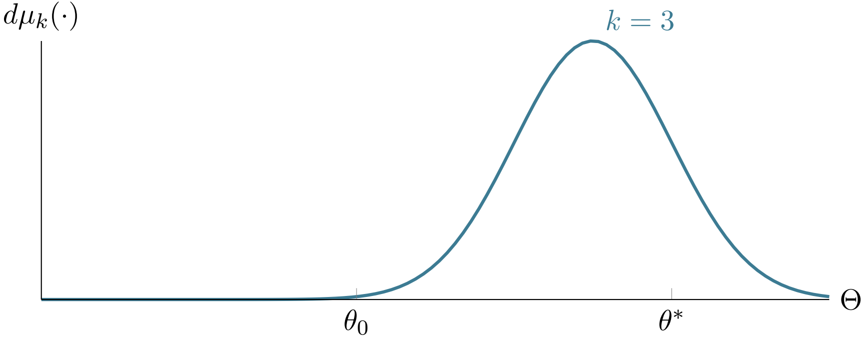
A toy problem for motivation

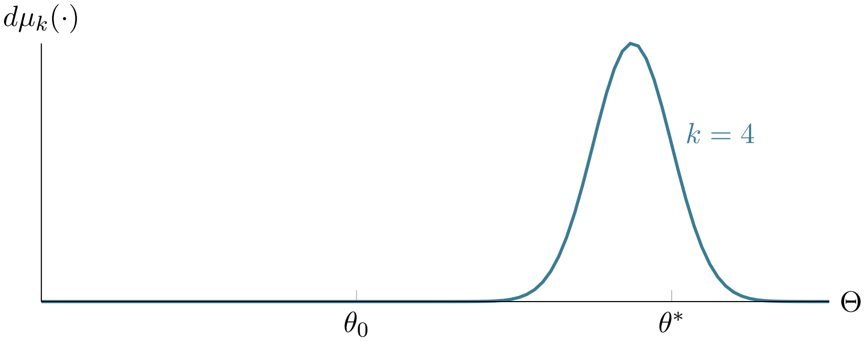


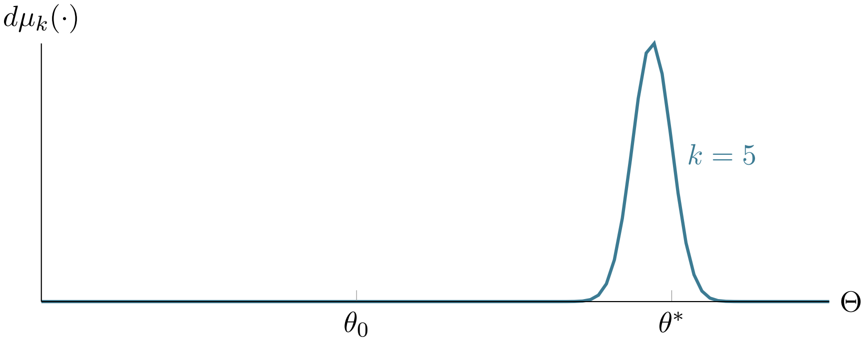

$d\mu_k(\cdot)$
$k = 0$
θ_0
θ^*
Θ

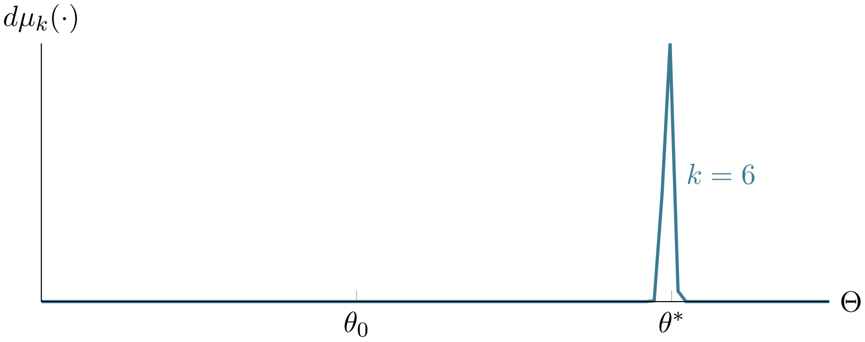

$d\mu_k(\cdot)$
$k = 1$
θ_0
θ^*
Θ


$d\mu_k(\cdot)$
$k = 2$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 3$
θ_0
θ^*
Θ


$d\mu_k(\cdot)$
$k = 4$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 5$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 6$
θ_0
θ^*
Θ

Information Exchange

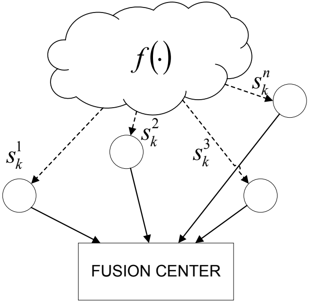


Figure: Distributed Observations Centralized Decision Making

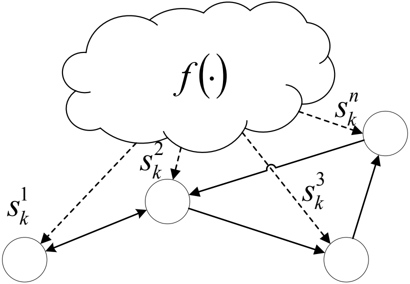


Figure: Distributed Observations, Distributed Decision Making

Information Exchange

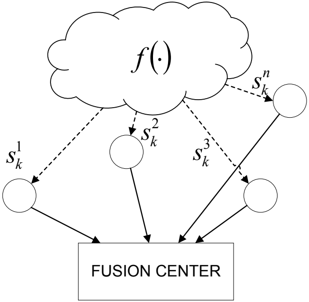


Figure: Distributed Observations Centralized Decision Making

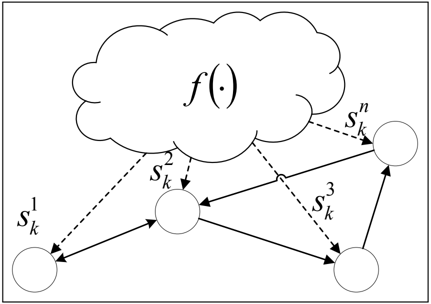


Figure: Distributed Observations, Distributed Decision Making

Problem Setup: Agent's Observations

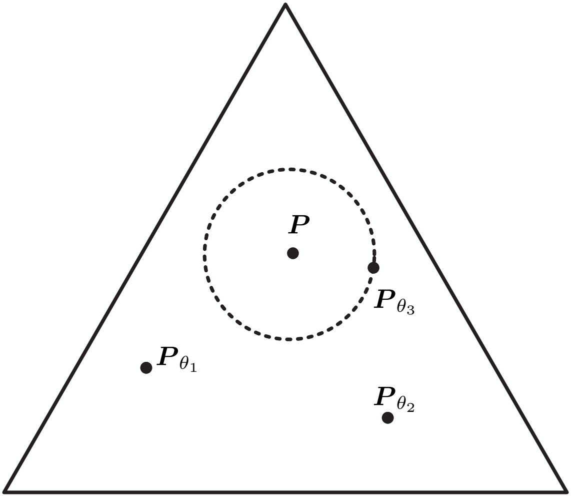
- m agents: $V = \{1, 2, \cdots, m\}$
- Agent i observes $X_k^i : \Omega \to \mathcal{X}^i$, $X_k^i \sim P^i$
- Agent i has an hypothesis set about P^i, $\{P_\theta^i\}$
- Probability distributions on Θ denoted as beliefs
- Agent i belief on hypothesis θ at time k denoted as $\mu_k^i(\theta)$

Agents want to collectively solve the following optimization problem

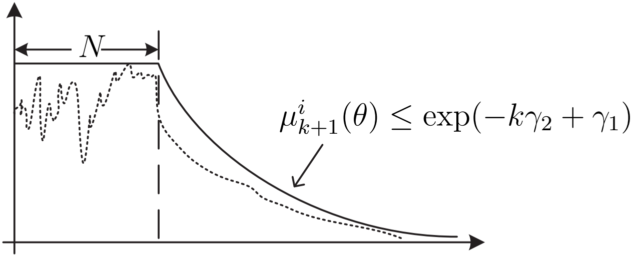
$$\min_{\theta \in \Theta} F(\theta) \triangleq D_{KL}(\boldsymbol{P} \| \boldsymbol{P}_\theta) = \sum_{i=1}^{m} D_{KL}(P^i \| P_\theta^i). \tag{7}$$

Consensus Learning: $d\mu_\infty^i(\theta^*) = 1$ for all i.

Geometric Interpretation for Finite Hypotheses



Informal Theorems from [Uribe et at. 2017]



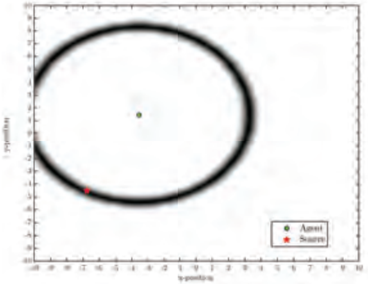
Under appropriate assumptions, the agents execute the distributed learning algorithm. Given a parameter $\rho \in (0,1)$, there is a time $N(m,\lambda,\rho)$ such that with probability $1-\rho$ for all $k \geq N(m,\lambda,\rho)$ for all $\theta \notin \Theta^*$,

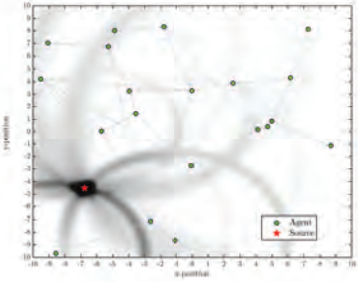
$$\mu^i_k(\theta) \leq \exp\left(-k\gamma_2 + \gamma_1\right) \quad \text{for all } i = 1,\ldots,n,$$

$$\mu_{k+1}^{i}(\theta) \leq \exp(-k\gamma_2 + \gamma_1) \quad \text{for all } i = 1, \ldots, m.$$

Graph Class	N	γ_1	γ_2	δ
Time-Varying Undirected	$O(\log 1/\rho)$	$O(m^3 \log m)$	$O(1)$	
$\cdots$ + Metropolis	$O(\log 1/\rho)$	$O(m^2 \log m)$	$O(1)$	
Time-Varying Directed	$\frac{1}{\delta^2} O(\log 1/\rho)$	$O(m^m \log m)$	$O(1)$	$\delta \geq \frac{1}{m^m}$
$\cdots$ + regular	$O(\log 1/\rho)$	$O(m^3 \log m)$	$O(1)$	1
Fixed Undirected	$O(\log 1/\rho)$	$O(m \log m)$	$O(1)$	

Distributed Source Localization

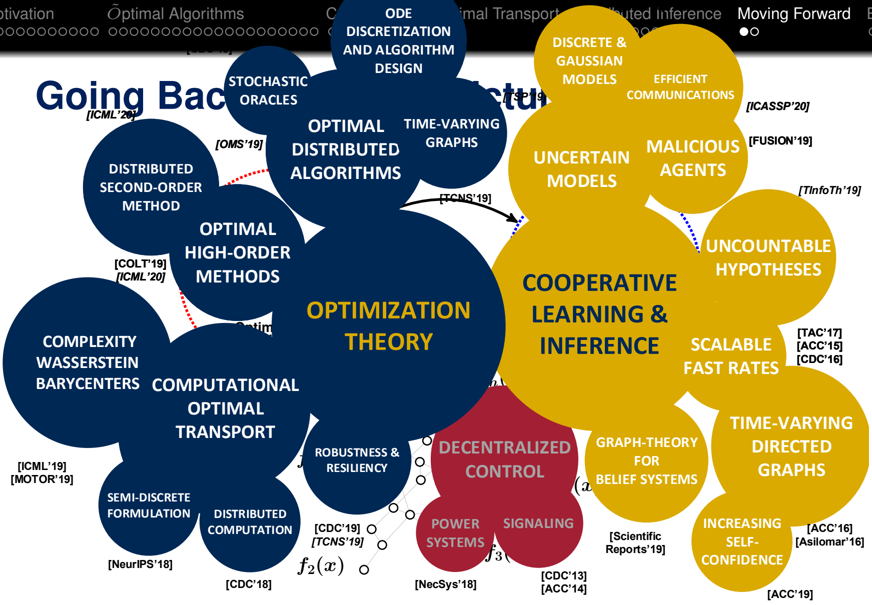




Click!

Challenges Moving Forward: Data-Driven Distributed Inference

- **Efficient belief communications:** How to communicate beliefs in when the number of hypothesis is large (maybe uncountably many)?
- **Non-parametric distributed learning:** How to define beliefs in non-parametric spaces? how to learn?
- **Distributed online learning and filtering:** Design “correct by definition” distributed algorithms for filtering and learning, e.g., what is the correct formulation of distributed Kalman filter?



DECENTRALIZED

SCALABLE

OPTIMAL

Towards Scalable Algorithms for Distributed Optimization and Learning

César A. Uribe

RICE ENGINEERING
Electrical and Computer Engineering

The Entropy-Regularized 2-Wasserstein Barycenter Problem: Discrete Distributions

$$\min_{p \in S_1(n)} \sum_{i=1}^{m} \mathcal{W}_\gamma(p, q_i).$$

$$\mathcal{W}_\gamma(p, q) \triangleq \min_{X \in U(p,q)} \left\{ \langle M, X \rangle - \gamma E(X) \right\},$$

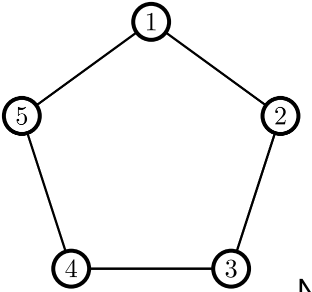
$$[M]_{ij} = \|x_i - x_j\|_2^2, \qquad \langle M, X \rangle \triangleq \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} X_{ij},$$

$$E(X) \triangleq -\sum_{i=1}^{n} \sum_{j=1}^{n} h(X_{ij}),$$

$$U(p, q) \triangleq \left\{ X \in \mathbb{R}_+^{n \times n} \mid X\mathbf{1} = p, X^T\mathbf{1} = q \right\}.$$

where $\gamma > 0$, and $h(x) \triangleq x \log x$.

A Dual Approach based on the Graph Laplacian



$$\bar{W} = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

Note that:

- $Wx = 0$ if and only if $x_1 = \ldots = x_m$.

Example: Estimating the Mean of a Gaussian Model

Data: Assume we receive a sample $x_1, \ldots, x_k$, where $X_k \sim \mathcal{N}(\theta^*, \sigma^2)$. σ^2 is known and we want to estimate θ^*.

Model: The collection of all Normal distributions with variance σ^2, i.e. $\mathscr{P}_\theta = \{\mathcal{N}(\theta, \sigma^2)\}$.

Prior: Our prior is the standard Normal distribution $d\mu_0(\theta) = \mathcal{N}(0, 1)$.

Posterior: The posterior is defined as

$$
\begin{aligned}
d\mu_k(\theta) &\propto d\mu_0(\theta) \prod_{t=1}^{k} p_\theta(x_t) \\
&= \mathcal{N}\left(\frac{\sum_{t=1}^{k} x_t}{\sigma^2 + k}, \frac{\sigma^2}{\sigma^2 + k} \right)
\end{aligned}
$$

Problem Reformulation

$$x = \begin{bmatrix} x_1 \in \mathbb{R}^n \\ x_2 \in \mathbb{R}^n \\ \vdots \\ x_m \in \mathbb{R}^n \end{bmatrix}$$

Rewrite problem (2) in an equivalent form as follows:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^m f_i(x) \qquad \text{equivalent to} \qquad \min_{Wx=0} \sum_{i=1}^m f_i(x_i), \qquad (8)$$

where $W = \bar{W} \otimes I_n$.

Some analysis tools

Initially, consider the general problem

$$\min_{Ax=0} f(x). \tag{9}$$

We assume that the problem has optimal solutions.
Later, we will derive the specific results when

$$A = \sqrt{W} \qquad \text{and} \qquad f(x) = \sum_{i=1}^{m} f_i(x_i)$$

Approximate Solution Definition A point $x \in \mathbb{R}^{mn}$ is said to be an $(\varepsilon, \tilde{\varepsilon})$-solution of (9) if the following conditions are satisfied:

$$f(x) - f^* \leq \varepsilon \qquad \text{and} \qquad \|Ax\|_2 \leq \tilde{\varepsilon},$$

where f^* denotes the optimal value of (9).

Construction of the dual problem

The Lagrangian dual for the problem in (9) is given by

$$\min_{Ax=0} f(x) = \max_{y} \left\{ \min_{x} \left\{ f(x) - \langle A^T y, x \rangle \right\} \right\},$$

or equivalently

$$\min_{y} \varphi(y) \ \text{ where } \ \varphi(y) \triangleq \max_{x} \left\{ \langle A^T y, x \rangle - f(x) \right\},$$

where $\nabla \varphi(y) = Ax^*(A^T y)$ (Demyanov-Danskin) with

$$x^*(A^T y) = \arg\max_{x} \left\{ \langle A^T y, x \rangle - f(x) \right\}.$$

We say that f is **dual friendly** *when we can determine a solution of the preceding problem efficiently (in a closed form ideally)*

The duality of strong convexity and smoothness, [Kakade et al., 2009] and others

- $f(x)$ is μ-strongly convex $\iff$ $\varphi(y)$ is L_φ-smooth with $L_\varphi = \lambda_{\max}(A^T A)/\mu$.
- $f(x)$ is L-smooth $\iff$ $\varphi(y)$ is μ_φ-strongly convex on the range space of A with $\mu_\varphi = \lambda^+_{\min}(A^T A)/L$.

The dual problem $\min_y \varphi(y)$ may have multiple solutions of the form $y^* + \ker(A^T)$.

Informally: If $f(x)$ has condition number $\frac{L}{\mu}$.

Then, $\varphi(y)$ has condition number $\frac{\lambda_{\max}(A^T A)}{\lambda^+_{\min}(A^T A)} \frac{L}{\mu}$

A proof sketch

Lets recall Nesterov's fast gradient method for

$$\min_{y} \varphi(y) \tag{10}$$

$$y_{k+1} = \tilde{y}_k - \frac{1}{L_\varphi}\nabla\varphi(\tilde{y}_k), \tag{11a}$$

$$\tilde{y}_{k+1} = y_{k+1} + \frac{\sqrt{L_\varphi} - \sqrt{\mu_\varphi}}{\sqrt{L} + \sqrt{\mu_\varphi}}(y_{k+1} - y_k). \tag{11b}$$

and

$$\varphi(y_k) - \varphi^* \leq L_\varphi \left(1 - \sqrt{\frac{\mu_\varphi}{L_\varphi}}\right)^k \|y_0 - y^*\|_2^2, \tag{12}$$

A proof sketch

Lets recall Nesterov's fast gradient method for

$$\min_y \varphi(y) \tag{10}$$

$$x^*(A^T\tilde{y}_k) = \arg\max_x \left\{ \langle A^T\tilde{y}_k, x\rangle - f(x) \right\} \tag{11a}$$

$$y_{k+1} = \tilde{y}_k - \frac{1}{L_\varphi} Ax^*(A^T\tilde{y}_k), \tag{11b}$$

$$\tilde{y}_{k+1} = y_{k+1} + \frac{\sqrt{L_\varphi} - \sqrt{\mu_\varphi}}{\sqrt{L_\varphi} + \sqrt{\mu_\varphi}}(y_{k+1} - y_k). \tag{11c}$$

and

$$\varphi(y_k) - \varphi^* \leq L_\varphi \left(1 - \sqrt{\frac{\mu_\varphi}{L_\varphi}}\right)^k \|y_0 - y^*\|_2^2, \tag{12}$$

What do agents do locally?

Set $A = \sqrt{W}$, $z_k = \sqrt{W} y_k$ and $\tilde{z}_k = \sqrt{W} \tilde{y}_k$

$$x_i^*(\tilde{z}_k^i) = \arg\max_{x_i} \left\{ \left\langle \tilde{z}_k^i, x_i \right\rangle - f_i(x_i) \right\}$$

$$z_{k+1}^i = \tilde{z}_k^i - \frac{\mu}{\lambda_{\max}(W)} \sum_{j=1}^{m} W_{ij} x_j^*(\tilde{z}_k^j)$$

$$\tilde{z}_{k+1}^i = z_{k+1}^i + \frac{\sqrt{\lambda_{\max}(W)/\mu} - \sqrt{\lambda_{\min}^+(W)/L}}{\sqrt{\lambda_{\max}(W)/\mu} + \sqrt{\lambda_{\min}^+(W)/L}} (z_{k+1}^i - z_k^i)$$