Distributed Learning for Cooperative Inference

César A. Uribe

Collaboration with: **Alex Olshevsky** and **Angelia Nedić**

LCCC - Focus Period on Large-Scale and Distributed Optimization

June 5th, 2017

Distributed ~~Learning~~ for ~~Cooperative~~ ~~Inference~~

Optimization

Consensus–Based *Statistical Estimation*

César A. Uribe

Collaboration with: **Alex Olshevsky** and **Angelia Nedić**

LCCC - Focus Period on Large-Scale and Distributed Optimization

June 5th, 2017

The three components for estimation

Data: $X \sim P^*$ is a r.v. with a sample space $(\mathcal{X}, \mathscr{X})$.
P^* is **unknown**.

Model:

- $\mathscr{P}$ a collection of probability measures $P : \mathscr{X} \to [0, 1]$.
- Parametrized by Θ; $\exists$ an injective map $\Theta \to \mathscr{P} : \theta \to P_\theta$.
- Dominated: $\exists \lambda$ s.t. $P_\theta \ll \lambda$ with $p_\theta = dP_\theta / d\lambda$.

(Point) Estimator: A map $\hat{P} : \mathcal{X} \to \mathscr{P}$. The best guess $\hat{P} \in \mathscr{P}$ for P^* based on X, e.g.

$$\hat{\theta}(X) = \sup_{\theta \in \Theta} p_\theta(X)$$

Bayesian Methods

The parameter is a r.v. ϑ taking values in $(\Theta, \mathscr{T})$.

There is a probability measure on $\mathcal{X} \times \Theta$ with $\mathscr{F} = \sigma(\mathscr{X} \times \mathscr{T})$,

$$\Pi : \mathscr{F} \to [0, 1],$$

Model: The distribution of X conditioned on ϑ, $\Pi_{X|\vartheta}$.

Prior: The marginal of Π on ϑ, $\Pi : \mathscr{T} \to [0, 1]$.

Posterior: The distribution $\Pi_{\vartheta|X} : \mathscr{T} \times \mathscr{X} \to [0, 1]$. In particular,

$$\Pi(\vartheta \in B|X) = \frac{\int_B p_\theta(X) d\Pi(\theta)}{\int_\Theta p_\theta(X) d\Pi(\theta)}.$$

One can construct the MAP or MMSE estimators as:

$$\hat{\theta}_{\text{MAP}}(X) = \arg\max_{\theta \in \Theta} \Pi(\theta|X)$$

$$\hat{\theta}_{\text{MMSE}}(X) = \int_{\theta \in \Theta} \theta d\Pi(\theta|X)$$

The Belief Notation

We are interested in computing posterior distributions. Thus, lets define the belief density on a hypothesis $\theta \in \Theta$ at time k as

$$\begin{aligned} d\mu_k(\theta) &= d\Pi(\theta|X_1, \ldots, X_k) \\ &\propto \prod_{i=1}^{k} p_\theta(X_i) d\Pi(\theta) \\ &= p_\theta(X_k) d\mu_{k-1}(\theta) \end{aligned}$$

This defines a iterative algorithm

$$d\mu_{k+1}(\theta) \propto d\mu_k(\theta) p_\theta(x_{k+1})$$

We will say that, we *learn* a parameter θ^* if

$$\lim_{k \to \infty} \mu_k(\theta^*) = 1 \qquad a.s. \text{ (usually)}$$

We hope that P_{θ^*} is the closest to P^* (in a sense defined later).

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^*.

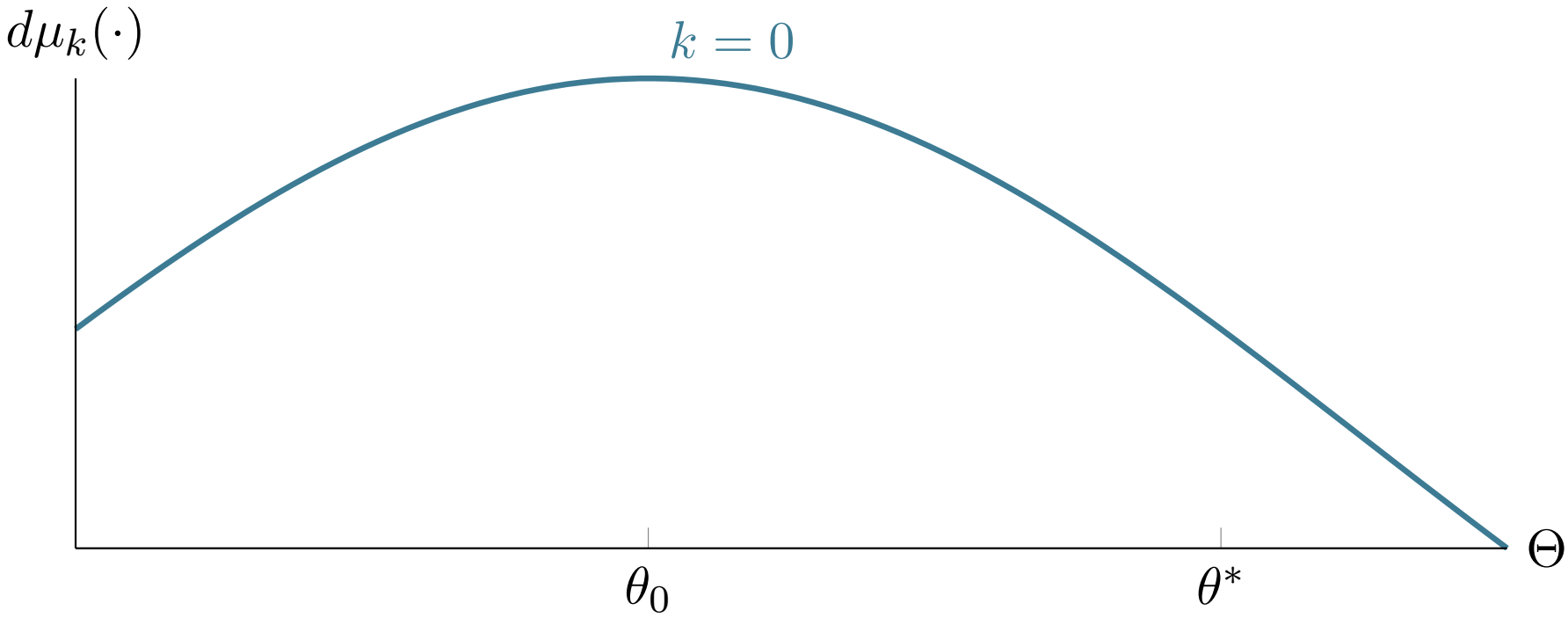
Model: The collection of all Normal distributions with variance σ^2, i.e. $\mathscr{P}_\theta = \{\mathcal{N}(\theta, \sigma^2)\}$.

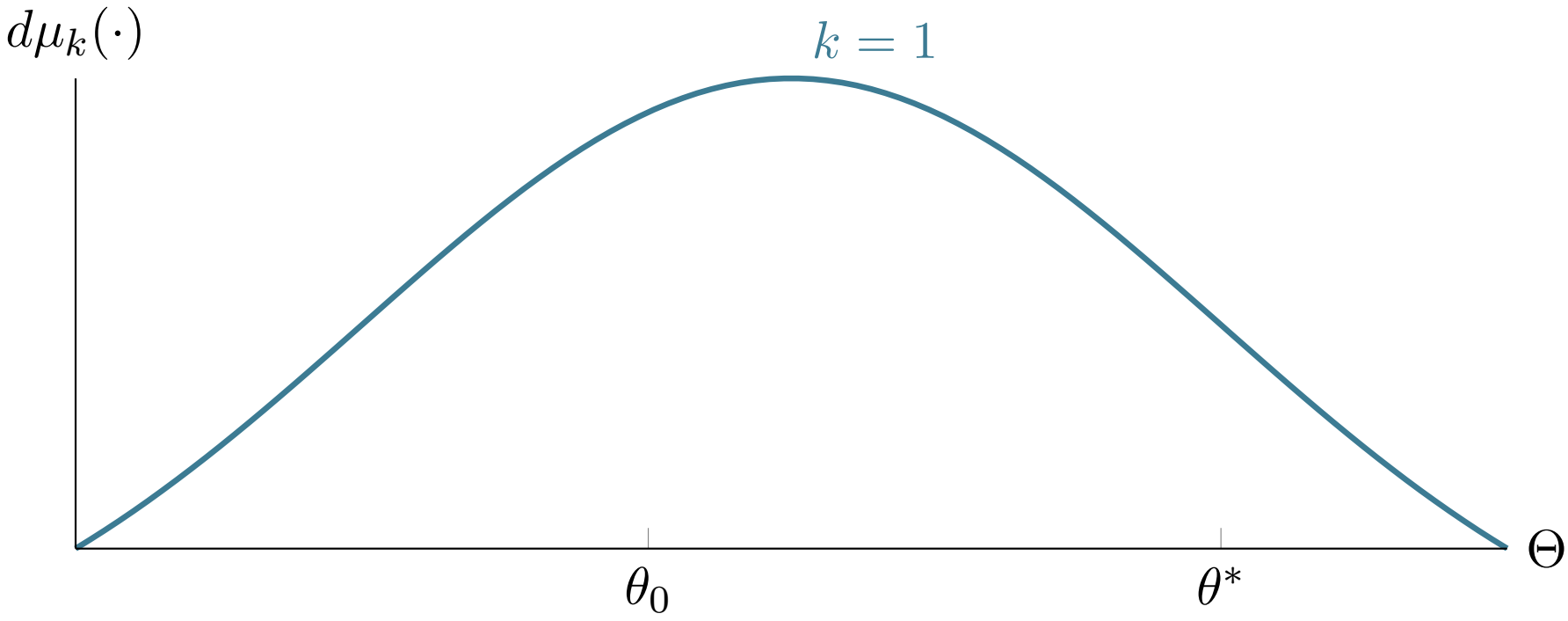
Prior: Our prior is the standard Normal distribution $d\mu_0(\theta) = \mathcal{N}(0, 1)$.

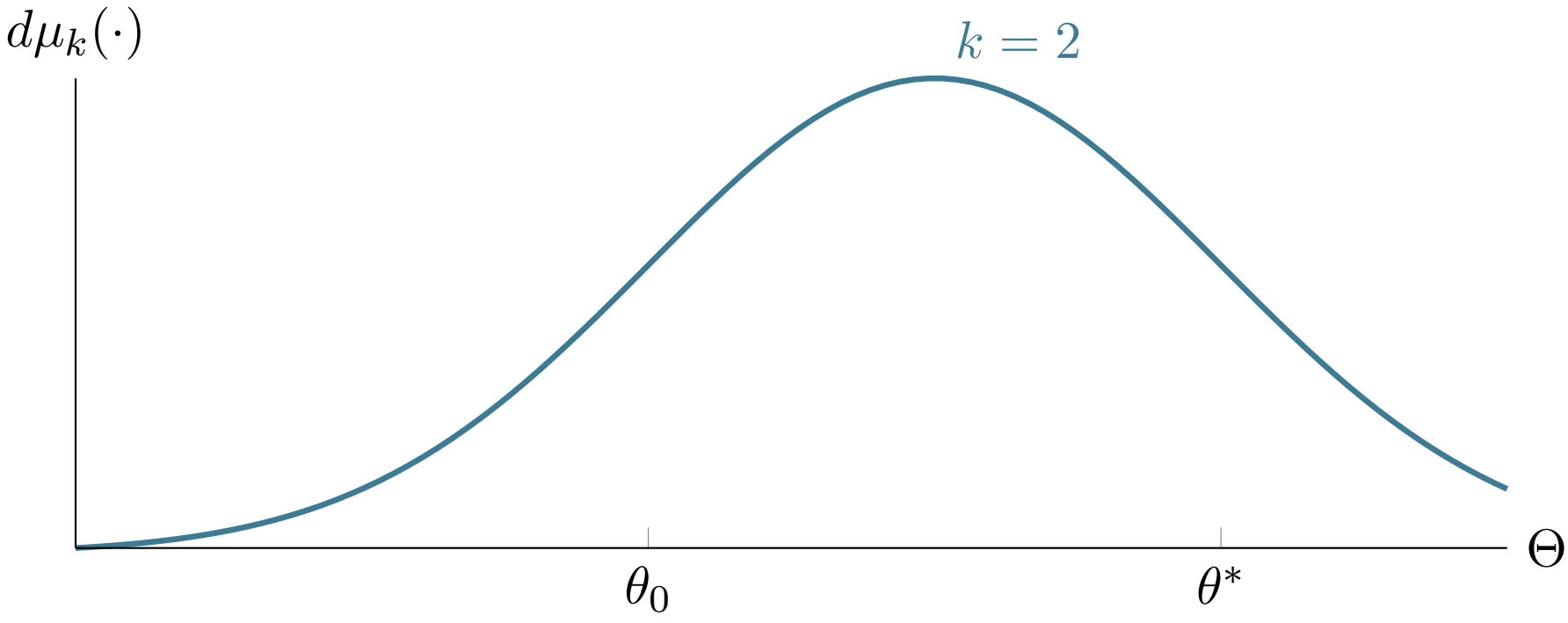
Posterior: The posterior is defined as

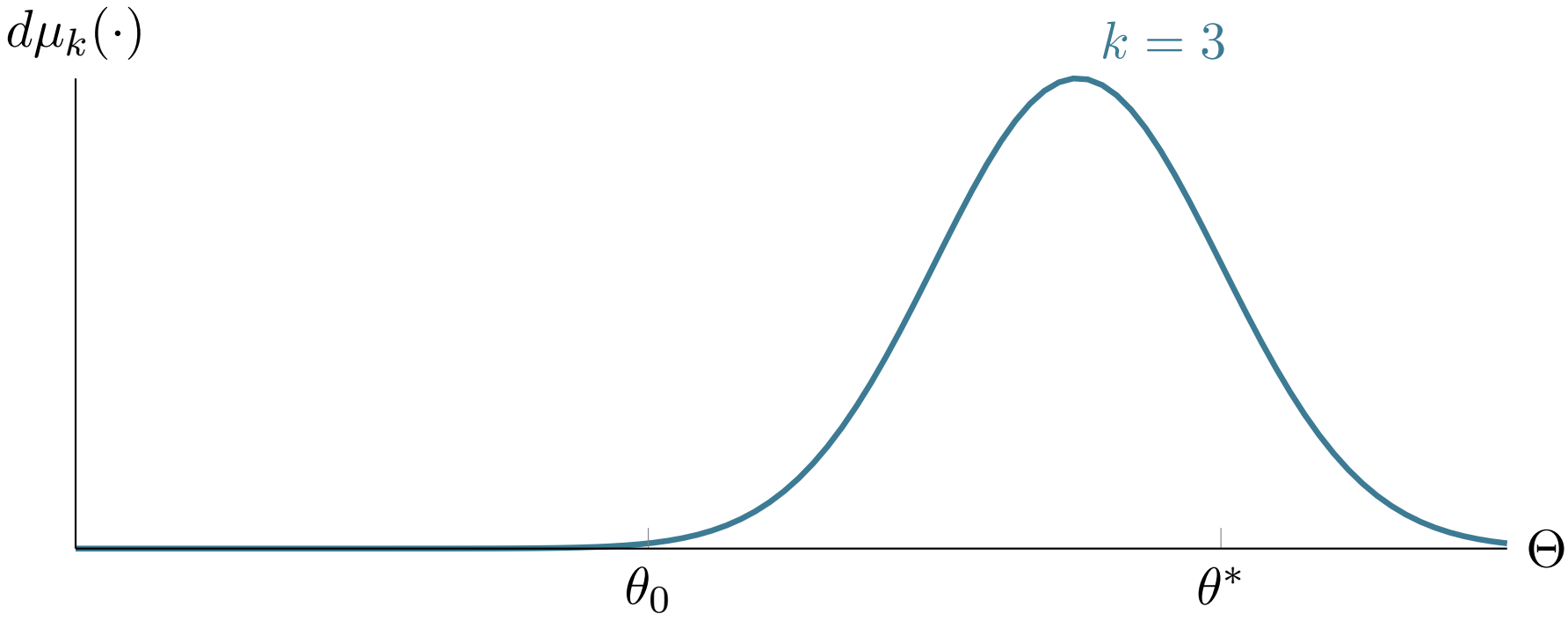
$$d\mu_k(\theta) \propto d\mu_0(\theta) \prod_{t=1}^{k} p_\theta(x_t)$$

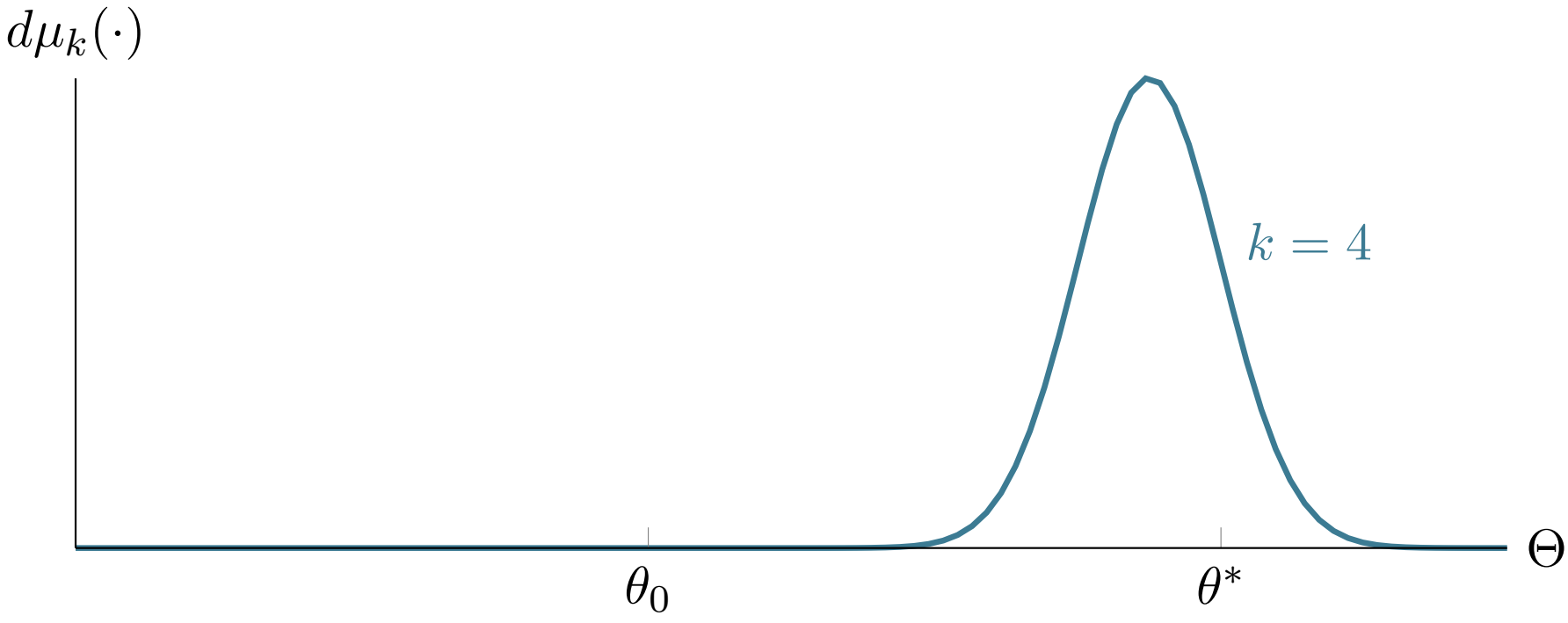
$$= \mathcal{N}\left(\frac{\sum_{t=1}^{k} x_t}{\sigma^2 + k}, \frac{\sigma^2}{\sigma^2 + k} \right)$$

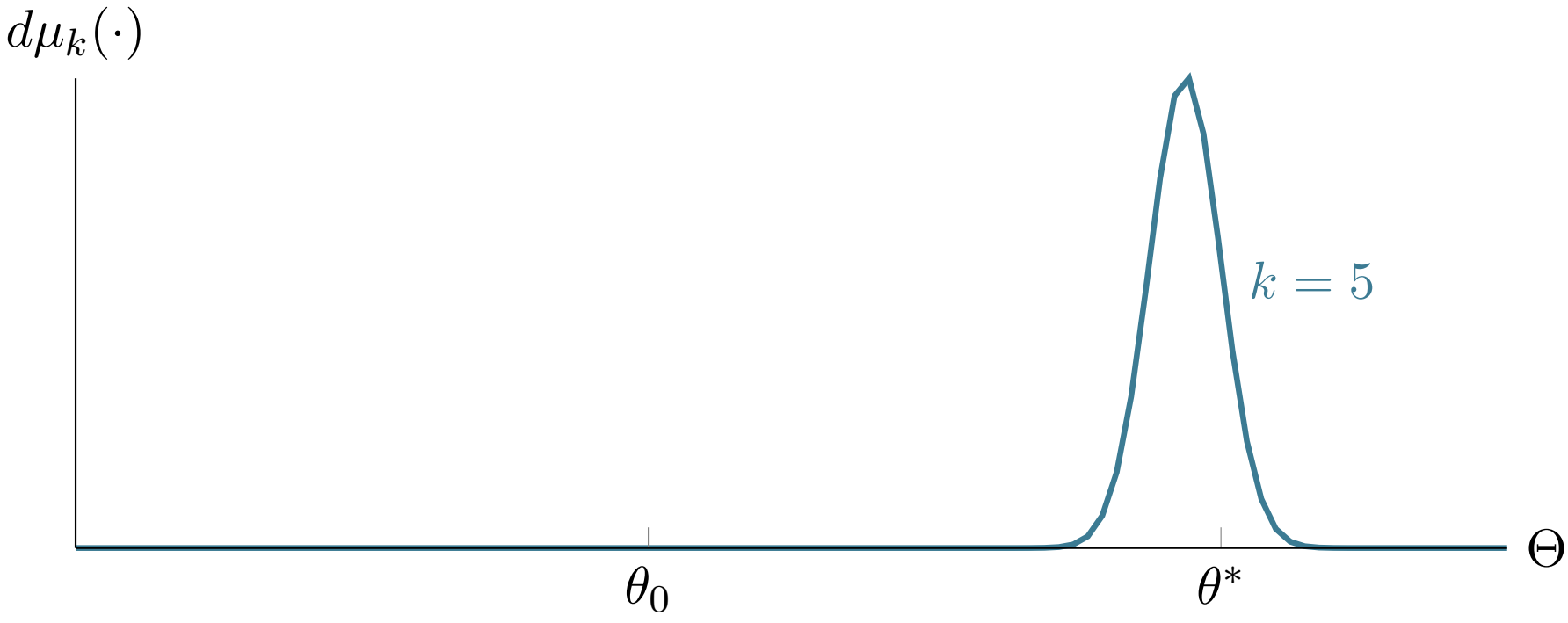

$d\mu_k(\cdot)$
$k = 0$
Θ
θ_0
θ^*

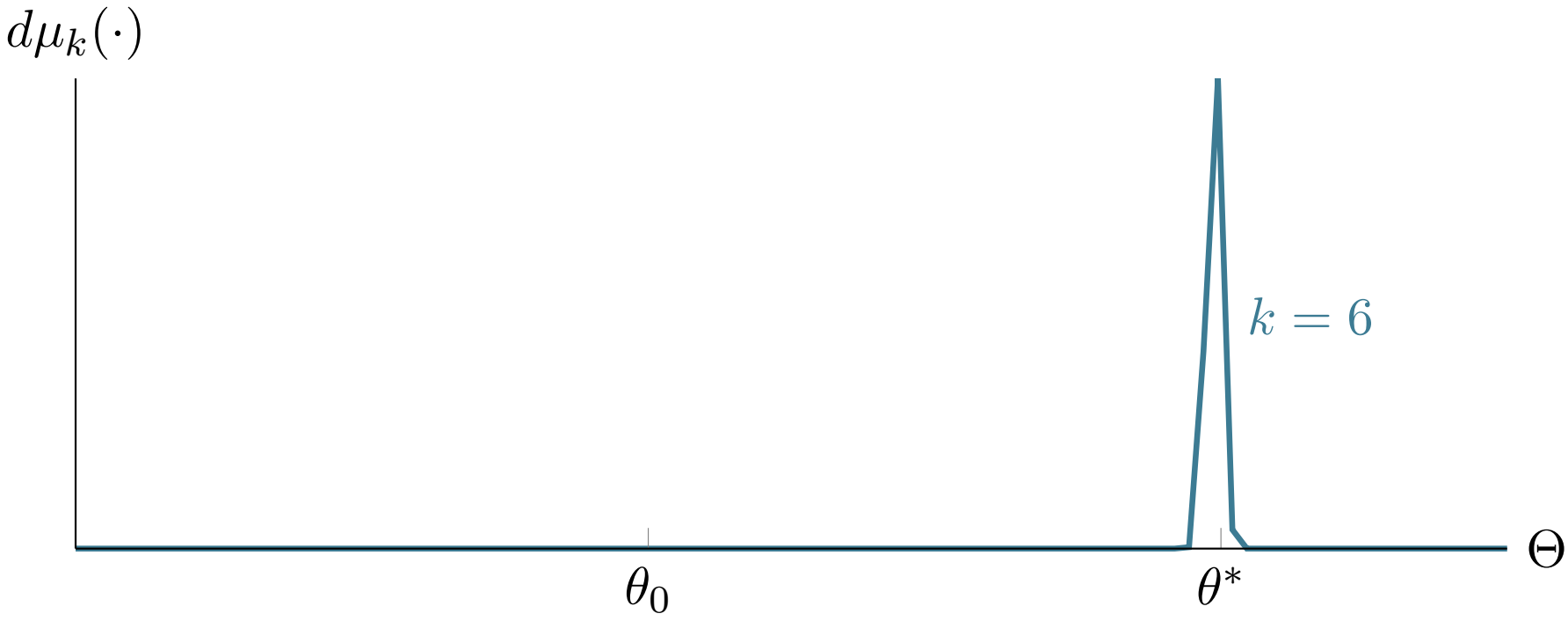

$d\mu_k(\cdot)$
$k = 1$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 2$
Θ
θ_0
θ^*

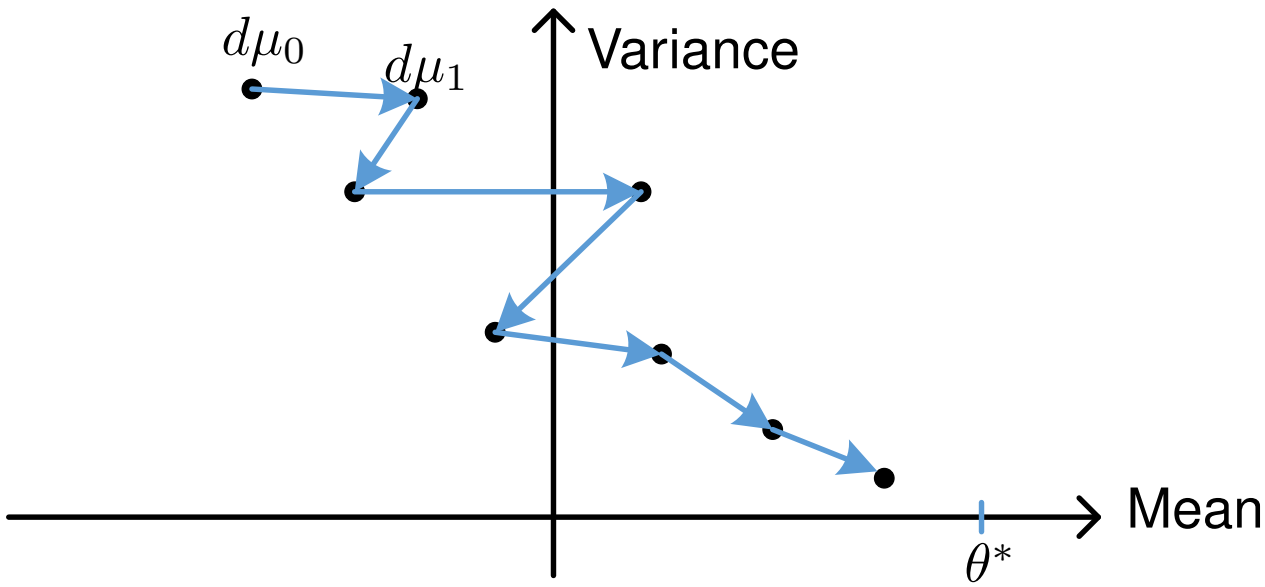

$d\mu_k(\cdot)$
$k = 3$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 4$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 5$
Θ
θ_0
θ^*


$d\mu_k(\cdot)$
$k = 6$
Θ
θ_0
θ^*

Geometric Interpretation for Finite Hypotheses



Bayes' Theorem Belongs to Stochastic Approximations

$\subset$

Consider the following optimization problem

$$\min_{\theta \in \Theta} F(\theta) = D_{KL}(P \| P_\theta), \tag{1}$$

We can rewrite Eq. (1) as

$$\begin{aligned} \min_{\theta \in \Theta} D_{KL}(P \| P_\theta) &= \min_{\pi \in \Delta_\Theta} \mathbb{E}_\pi D_{KL}(P \| P_\theta) \qquad \text{where } \theta \sim \pi \\ &= \min_{\pi \in \Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P \left[-\log \frac{dP_\theta}{dP} \right], \end{aligned}$$

Moreover,

$$\begin{aligned} \arg\min_{\theta \in \Theta} D_{KL}(P \| P_\theta) &= \arg\min_{\pi \in \Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P \left[-\log p_\theta(X) \right], \theta \sim \pi, X \sim P \\ &= \arg\min_{\pi \in \Delta_\Theta} \mathbb{E}_P \mathbb{E}_\pi \left[-\log p_\theta(X) \right], \theta \sim \pi, X \sim P. \end{aligned}$$

Consider the following optimization problem

$$\min_{x \in Z} \mathbb{E}\left[F(x, \Xi)\right],$$

The stochastic mirror descent approach constructs a sequence $\{x_k\}$ as follows:

$$x_{k+1} = \arg\min_{x \in Z} \left\{ \langle \nabla F(x, \xi_k), x \rangle + \frac{1}{\alpha_k} D_w(x, x_k) \right\},$$

Recall our original problem

$$\min_{\pi \in \Delta_\Theta} \mathbb{E}_P \mathbb{E}_\pi \left[-\log p_\theta(X) \right], \theta \sim \pi, X \sim P. \tag{2}$$

For Eq. (2), Stochastic Mirror Descent generates a sequence of densities $\{d\mu_k\}$, as follows:

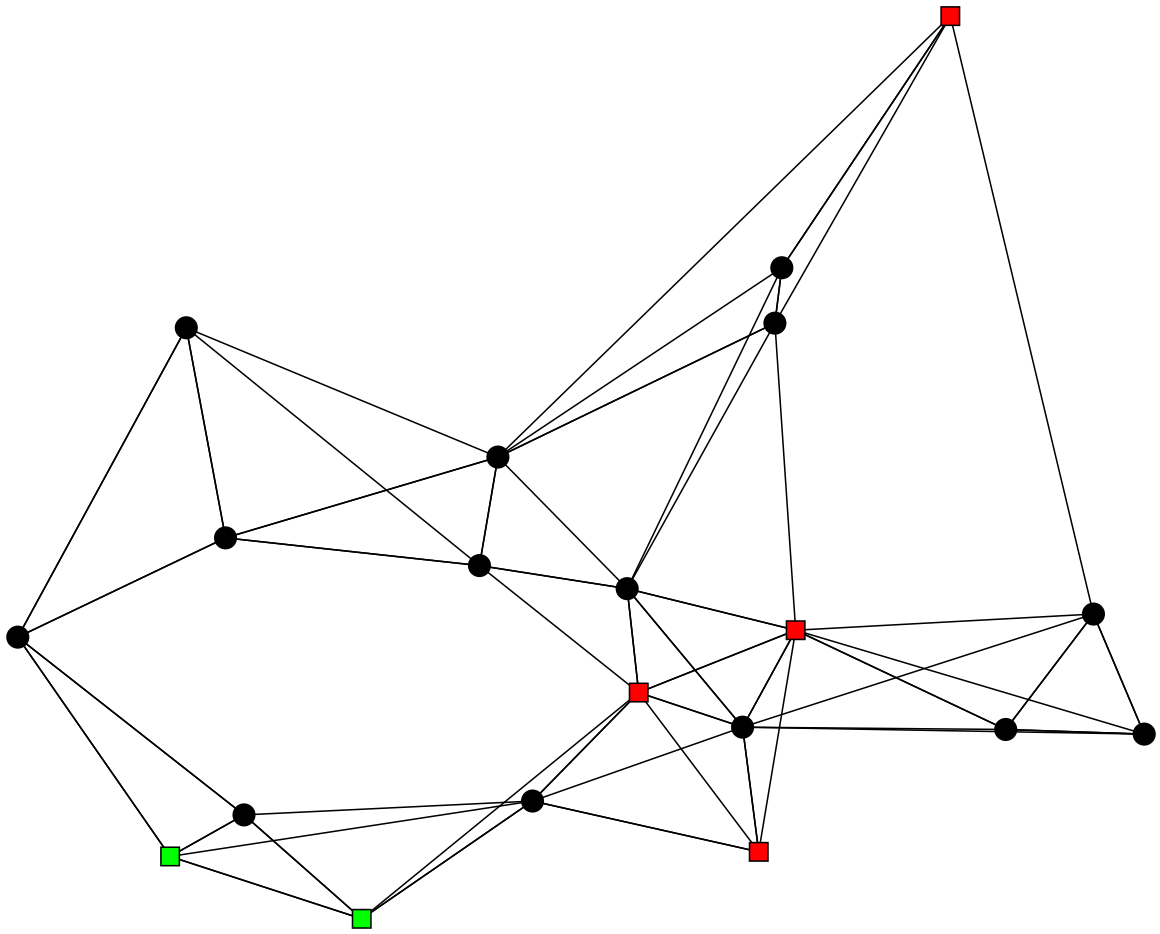
$$d\mu_{k+1} = \arg\min_{\pi \in \Delta_\Theta} \left\{ \langle -\log p_\theta(x_{k+1}), \pi \rangle + \frac{1}{\alpha_k} D_w(\pi, d\mu_k) \right\}, \theta \sim \pi. \tag{3}$$

$$d\mu_{k+1} = \underset{\pi \in \Delta_\Theta}{\arg\min} \left\{ \langle -\log p_\theta(x_{k+1}), \pi \rangle + D_{KL}(\pi \| d\mu_k) \right\}, \theta \sim \pi.$$

Choose $w(x) = \int x \log x$, then the corresponding Bregman distance is the Kullback-Leibler (KL) divergence D_{KL}. Additionally, by selecting $\alpha_k = 1$ then for each $\theta \in \Theta$,

$$\underbrace{d\mu_{k+1}(\theta) \propto p_\theta(x_{k+1}) d\mu_k(\theta)}_{Bayesian\ Posterior}$$

Distributed Inference Setup



Distributed Inference Setup

- n agents: $V = \{1, 2, \cdots, n\}$
- Agent i observes $X_k^i : \Omega \to \mathcal{X}^i$, $X_k^i \sim P^i$
- Agent i has a model about P^i, $\mathscr{P}^i = \{P_\theta^i | \theta \in \Theta\}$
- Agent i has a local belief density $d\mu_k^i(\theta)$
- Agents share beliefs over the network (connected, fixed, undirected)
- $a_{ij} \in (0, 1)$ is how agent i weights agent j information, $\sum a_{ij} = 1$

Agents want to collectively solve the following optimization problem

$$\min_{\theta \in \Theta} F(\theta) \triangleq D_{KL}(\boldsymbol{P} \| \boldsymbol{P}_\theta) = \sum_{i=1}^{n} D_{KL}(P^i \| P_\theta^i). \tag{4}$$

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

Our approach

Include beliefs of other agents in the regularization term:

Distributed Stochastic Entropic Mirror-descent

$$d\mu_{k+1}^i = \arg\min_{\pi \in \Delta_\Theta} \left\{ \sum_{j=1}^n a_{ij} D_{KL}\left(\pi \| d\mu_k^j\right) - \mathbb{E}_\pi \left[\log\left(p_\theta^i\left(x_{k+1}^i\right)\right)\right] \right\}$$

$$d\mu_{k+1}^i(\theta) \propto \prod_{j=1}^n d\mu_k^j(\theta)^{a_{ij}} \, p_\theta^i\left(x_{k+1}^i\right) \tag{5}$$

Q1. Does (5) achieves consensus learning?
Q2. If Q1 is positive, at what rate does this happens?

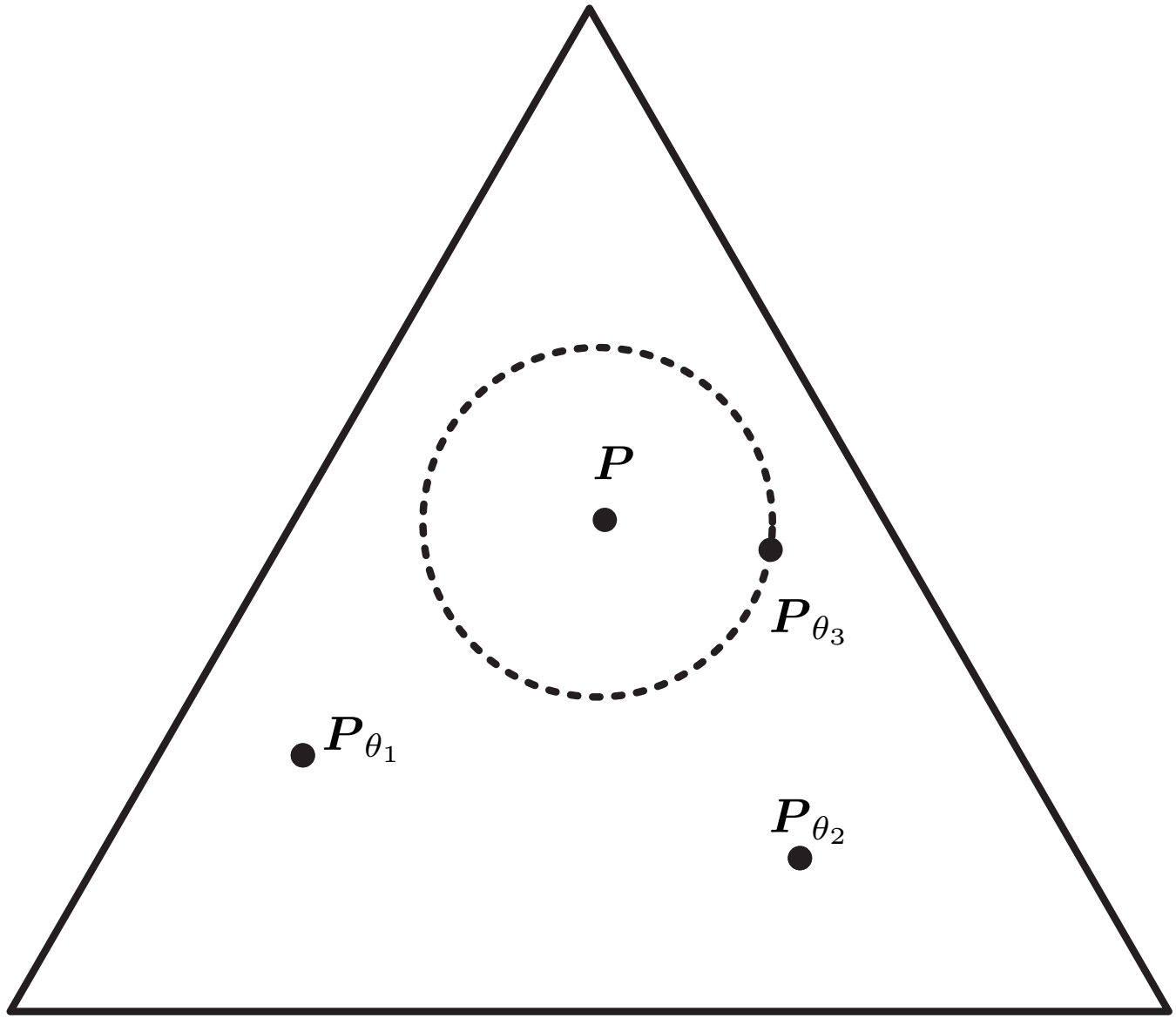
A finite set Θ

Extensive literature for finite parameter sets Θ

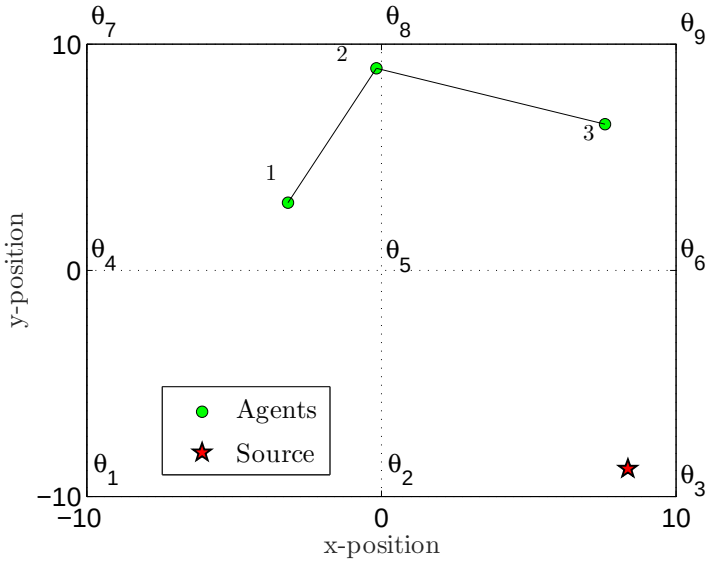
- The network is static/time-varying.
- The network is directed/undirected.
- Prove consistency of the algorithm.
- Prove asymptotic/non-asymptotic convergence rates.

Shahrampour, Rahimian, Jadbabaie, Lalitha, Sarwate, Javidi, Su, Vaidya, Qipeng, Bandyopadhyay, Sahu, Kar, Sayed, Chazelle, Olshevsky, Nedić, U.

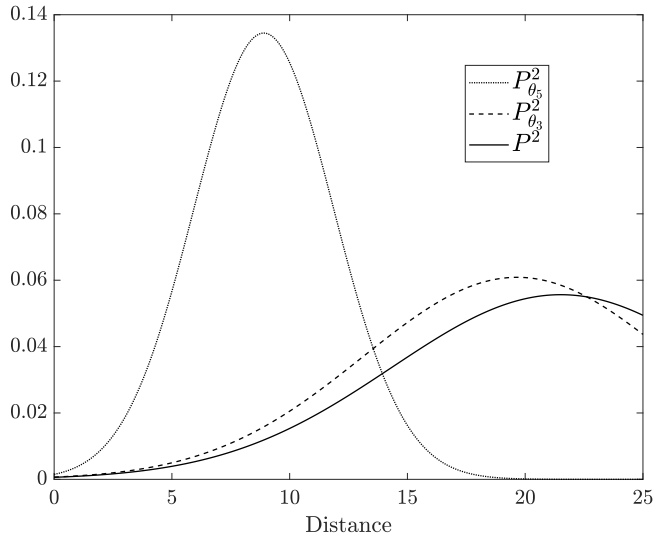
Geometric Interpretation for Finite Hypotheses



Distributed Source Localization

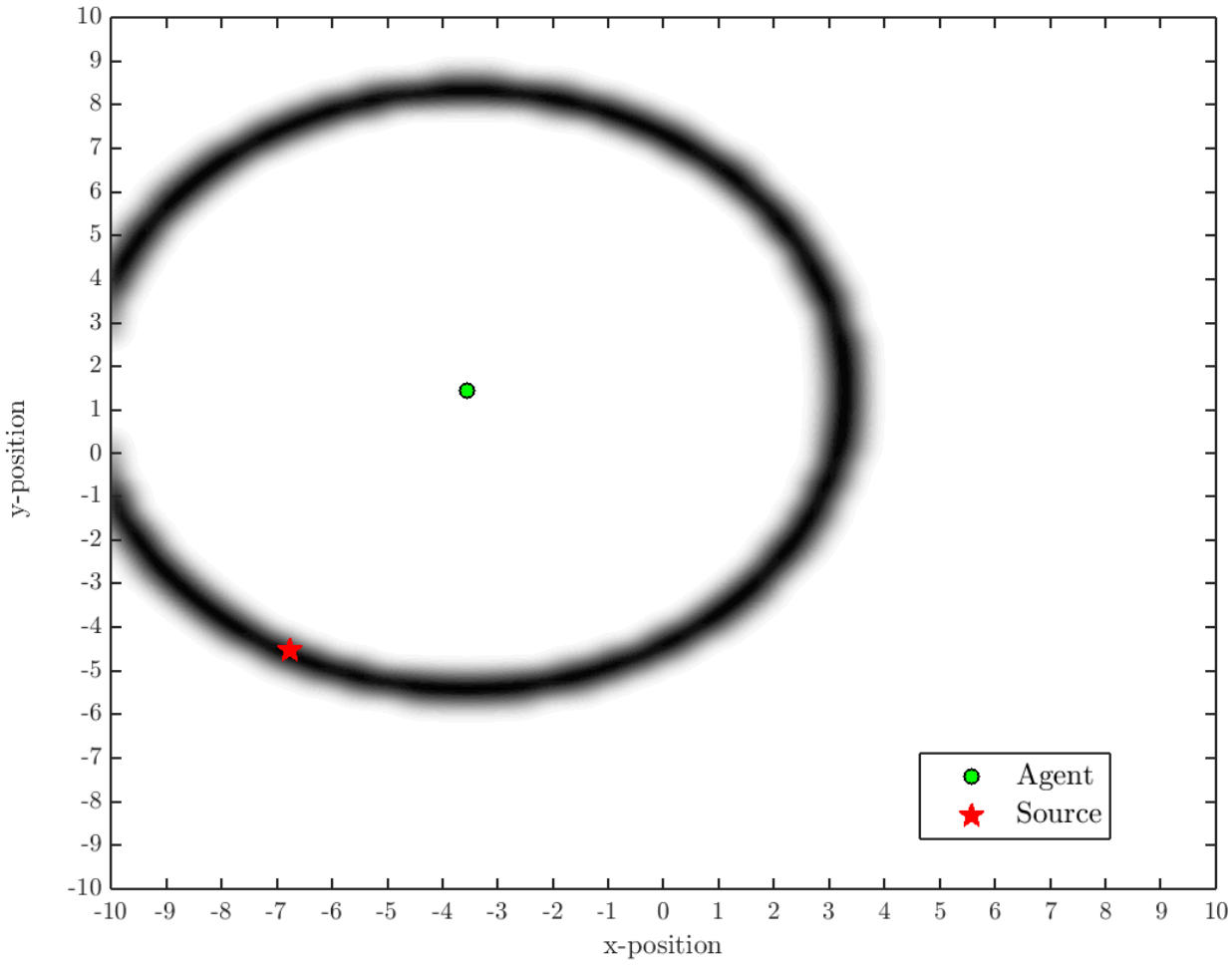


(a) Network of Agents

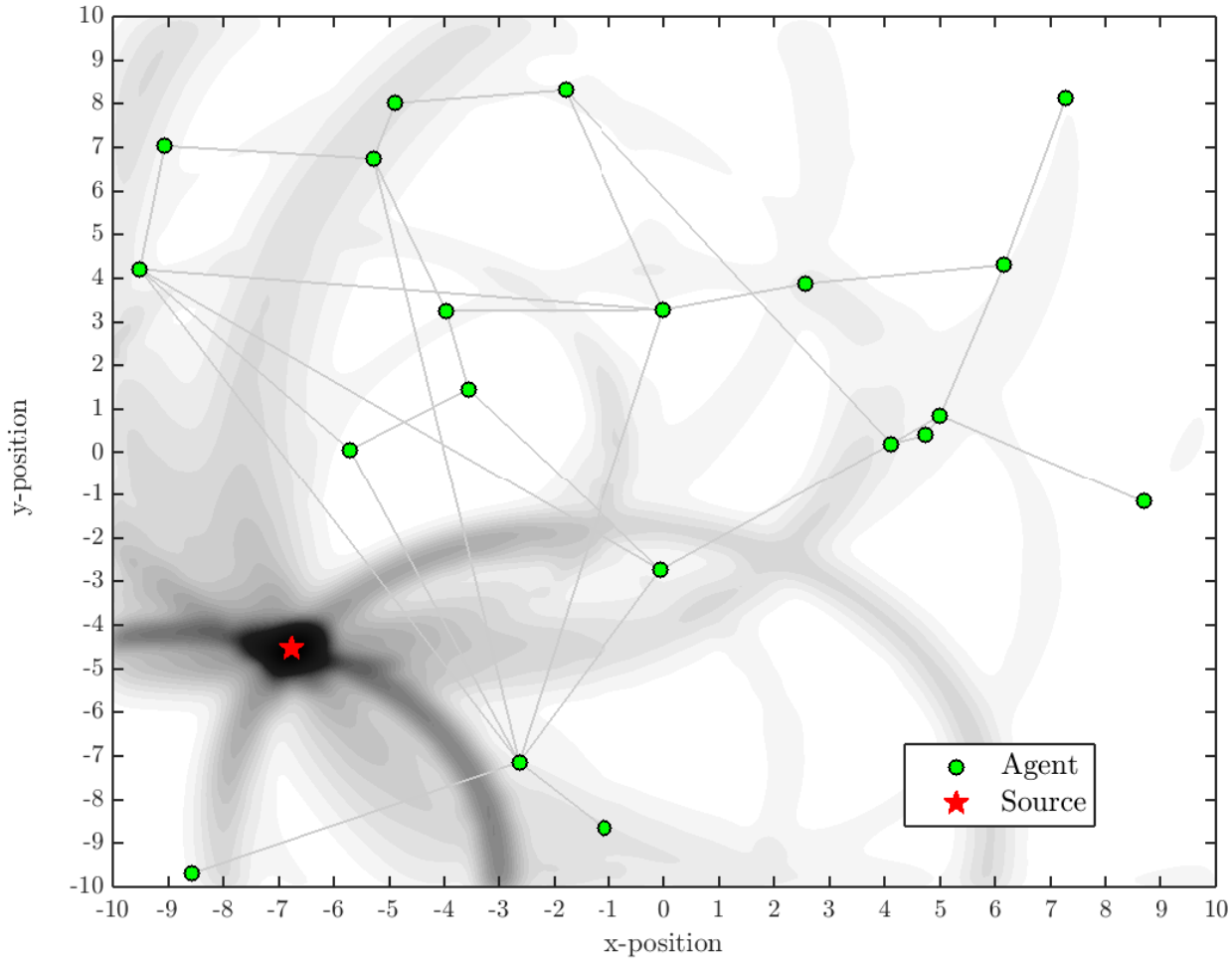


(b) Hypothesis Distributions

Distributed Source Localization



Distributed Source Localization



Our results for three different problems

1. Time-varying undirected graphs (Nedić,Olshevsky,U to appear TAC)
   - A_k is doubly-stochastic with $[A_k]_{ij} > 0$ if $(i,j) \in E_k$.
2. Time-varying directed graphs (Nedić,Olshevsky,U in ACC16)
   - $[A_k]_{ij} = \begin{cases} \frac{1}{d_k^j} & \text{if } j \in {}^{\text{in}}N_k^i \\ 0 & \text{if otherwise} \end{cases}$

     d_k^i is the out degree of node i at time k.
     ${}^{\text{in}}N_k^i$ is the set of in neighbors of node i.
3. Acceleration in static graphs (Nedić,Olshevsky,U to appear TAC)
   - $\bar{A}_{ij} = \begin{cases} \frac{1}{\max\{d^i, d^j\}} & \text{if } (i,j) \in E, \\ 0 & \text{if } (i,j) \notin E, \end{cases}$

     d^i degree of the node i.
     $A = \frac{1}{2}I + \frac{1}{2}\bar{A}$,

Time-Varying Undirected

$$\mu_{k+1}^i(\theta) \propto \prod_{j=1}^n \mu_k^j(\theta)^{[A_k]_{ij}} p_\theta^i(x_{k+1}^i)$$

Fixed Undirected

$$\mu_{k+1}^i(\theta) \propto \frac{\prod_{j=1}^n \mu_k^j(\theta)^{(1+\sigma)\bar{A}_{ij}} p_\theta^i(x_{k+1}^i)}{\prod_{j=1}^n \left(\mu_{k-1}^j(\theta) p_\theta^j(x_k^j)\right)^{\sigma \bar{A}_{ij}}}$$

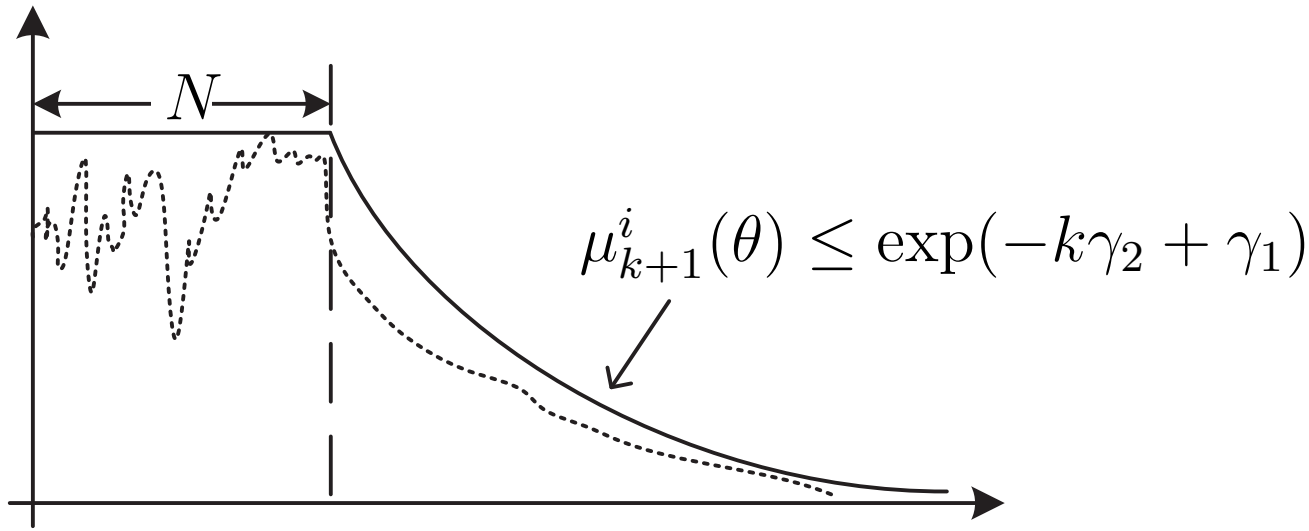
Time-Varying

$$y_{k+1}^i = \sum_{j \in N_k^i} \frac{y_k^j}{d_k^j}$$

Directed

$$\mu_{k+1}^i(\theta) \propto \left(\prod_{j \in N_k^i} \mu_k^j(\theta)^{\frac{y_k^j}{d_k^j}} p_\theta^i\left(x_{k+1}^i\right) \right)^{\frac{1}{y_{k+1}^i}}$$

General form of Theorems



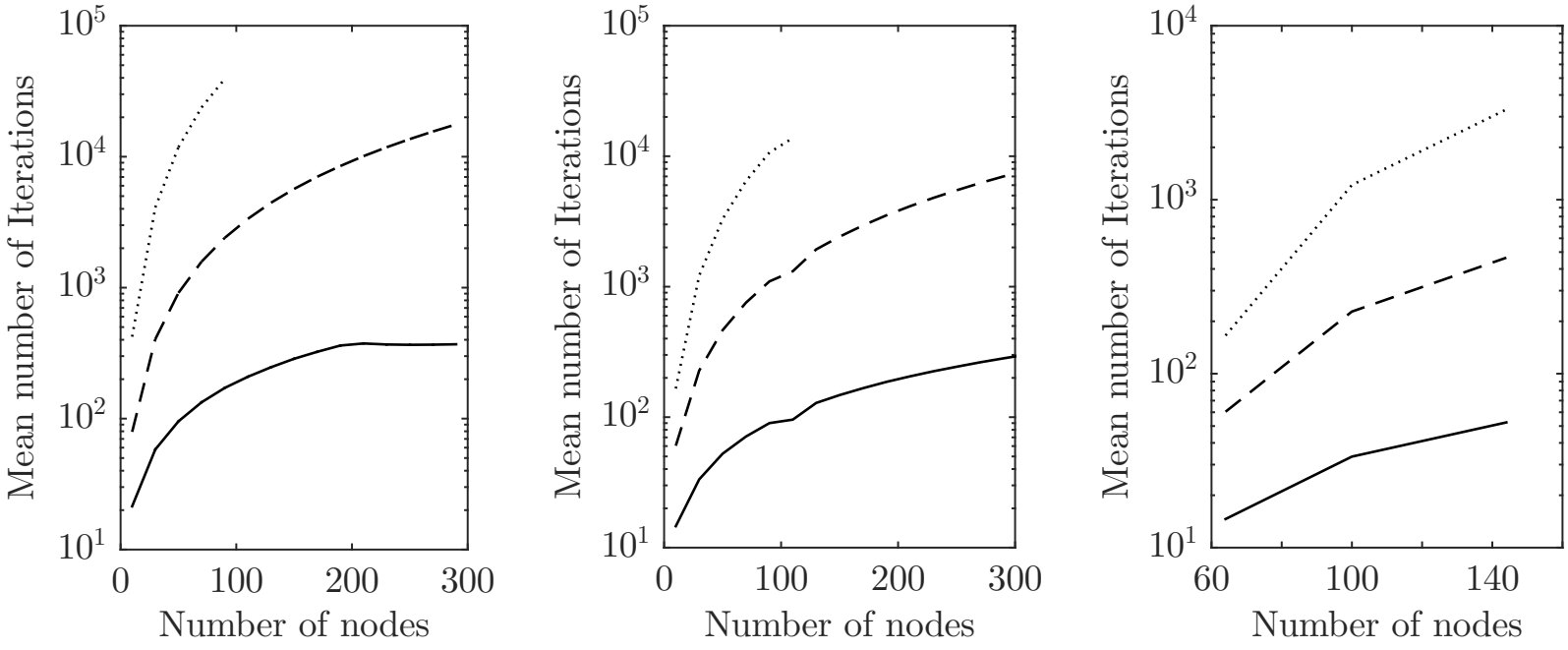
Under appropriate assumptions, a group of agents following algorithm X. There is a time $N(n, \lambda, \rho)$ such that with probability $1 - \rho$ for all $k \geq N(n, \lambda, \rho)$ for all $\theta \notin \Theta^*$,

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

After a time $N(n, \lambda, \rho)$ such that with probability $1 - \rho$ for all $k \geq N(n, \lambda, \rho)$, for all $\theta \notin \Theta^*$,

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

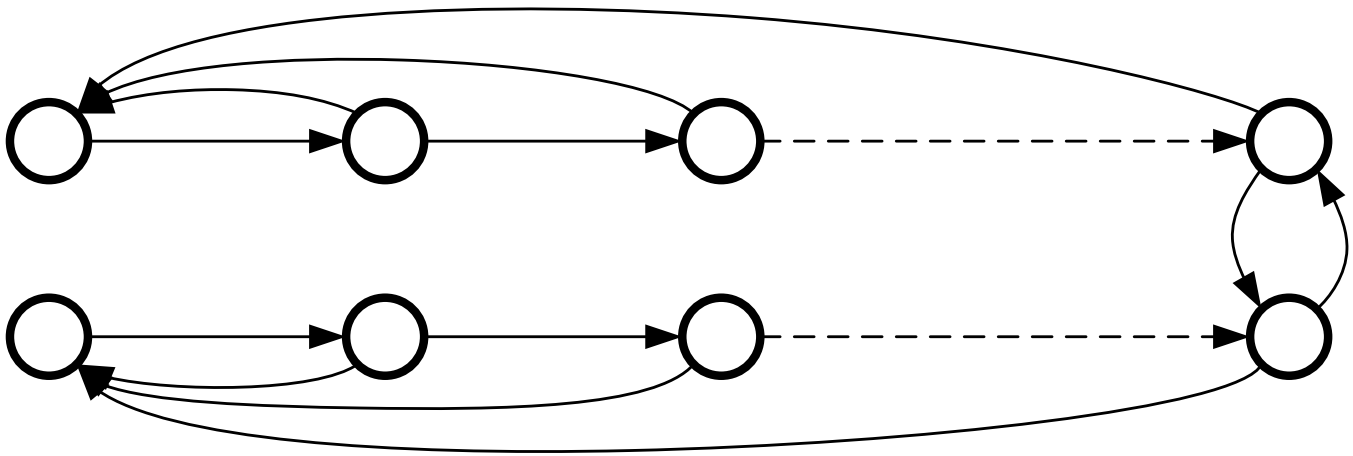
Graph	N	γ_1	γ_2	δ
Time-Varying Undirected	$O(\log 1/\rho)$	$O(\frac{n^2}{\eta} \log n)$	$O(1)$	
$\cdots$ + Metropolis	$O(\log 1/\rho)$	$O(n^2 \log n)$	$O(1)$	
Time-Varying Directed	$\frac{1}{\delta^2} O(\log 1/\rho)$	$O(n^n \log n)$	$O(1)$	$\delta \geq \frac{1}{n^n}$
$\cdots$ + regular	$O(\log 1/\rho)$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

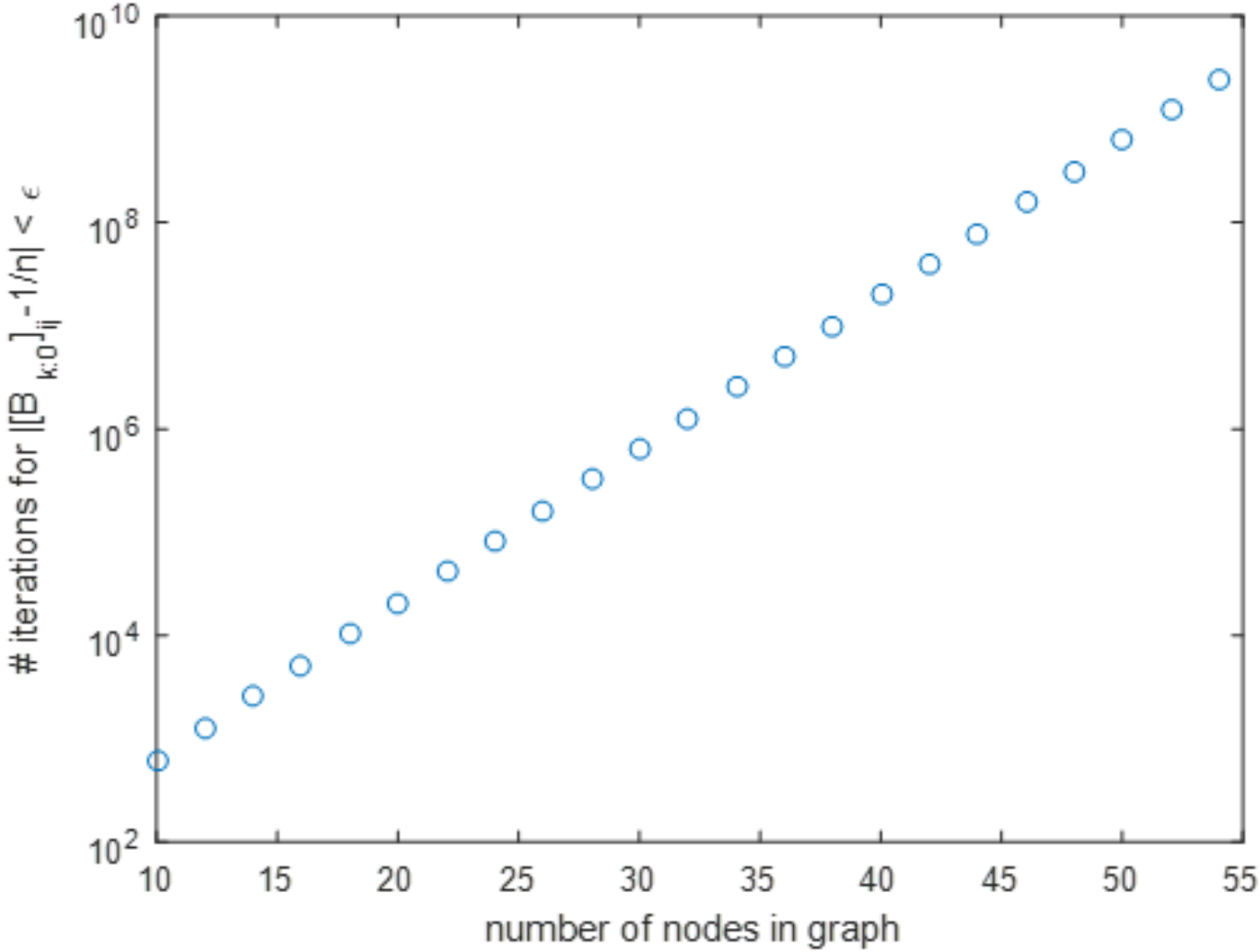


(a) Path Graph (b) Circle Graph (c) Grid Graph

Figure: Empirical mean over 50 Monte Carlo runs of the number of iterations required for $\mu_k^i(\theta) < \epsilon$ for all agents on $\theta \notin \Theta^*$. All agents but one have all their hypotheses to be observationally equivalent. Dotted line for the algorithm proposed by Jadbabaie et al. Dashed line no acceleration and solid line for acceleration.

A particularly bad graph




10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
iterations for $|[B_{\kappa:0}]_{ij}-1/n| < \epsilon$
10
15
20
25
30
35
40
45
50
55
number of nodes in graph

A problem with compact sets of Hypotheses

In particular, after a transient time depending on γ_1, the convergence rate is geometric with rate γ_2.

$$\gamma_2 = \frac{1}{n} \min_{\theta \notin \Theta^*} \sum_{i=1}^{n} \left(D_{KL}(P^i \| P^i_\theta) - D_{KL}((P^i \| P^i_{\theta^*}) \right)$$

γ_2 is the average distance between the second best hypotheses and the optimal one. This term goes to zero if there is a continuum of hypotheses, e.g. $\Theta \in \mathbb{R}^d$.

Q3. Can we derive nonasymptotic geometric concentration rates for the proposed learning rule?

$$\mu^i_{k+1}(B) \propto \int_{\theta \in B} \prod_{j=1}^{n} \left(d\mu^j_k(\theta) \right)^{a_{ij}} p^i_\theta(x^i_{k+1}) \tag{6}$$

A compact set of hypotheses: $\Theta \subset \mathbb{R}^d$

LeCam+Birgé

- Birgé, ”Model selection via testing: an alternative to (penalized) maximum likelihood estimators.”, 2006.
- Birgé, ”About the non-asymptotic behaviour of Bayes estimators.”, 2015.
- LeCam, ”Convergence of estimates under dimensionality restrictions.”, 1973.

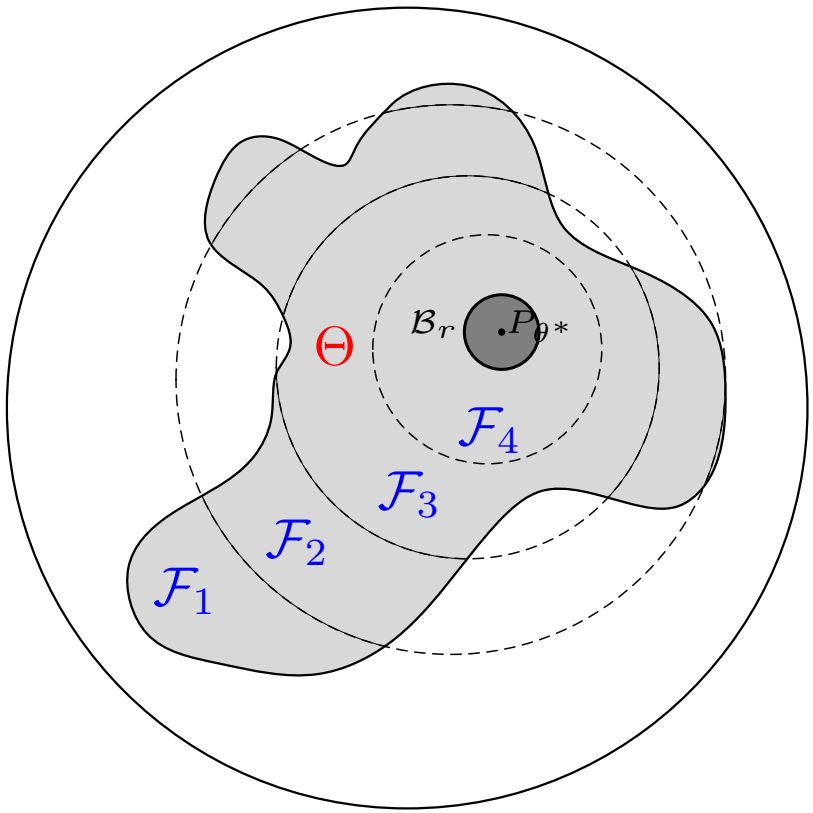
A couple of definitions first

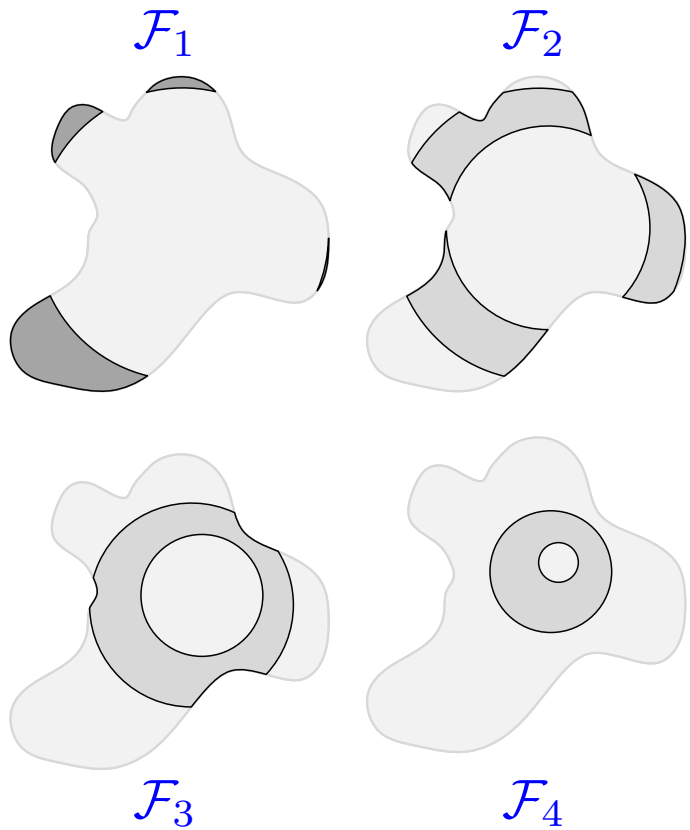
Define an n-Hellinger ball of radius r centered at θ as

$$\mathcal{B}_r(\theta) = \left\{ \hat{\theta} \in \Theta \left| \sqrt{\frac{1}{n} \sum_{i=1}^{n} h^2 \left(P_\theta^i, P_{\hat{\theta}}^i \right)} \leq r \right. \right\}$$

A covering for $\mathcal{B}_r^c \cap \Theta$

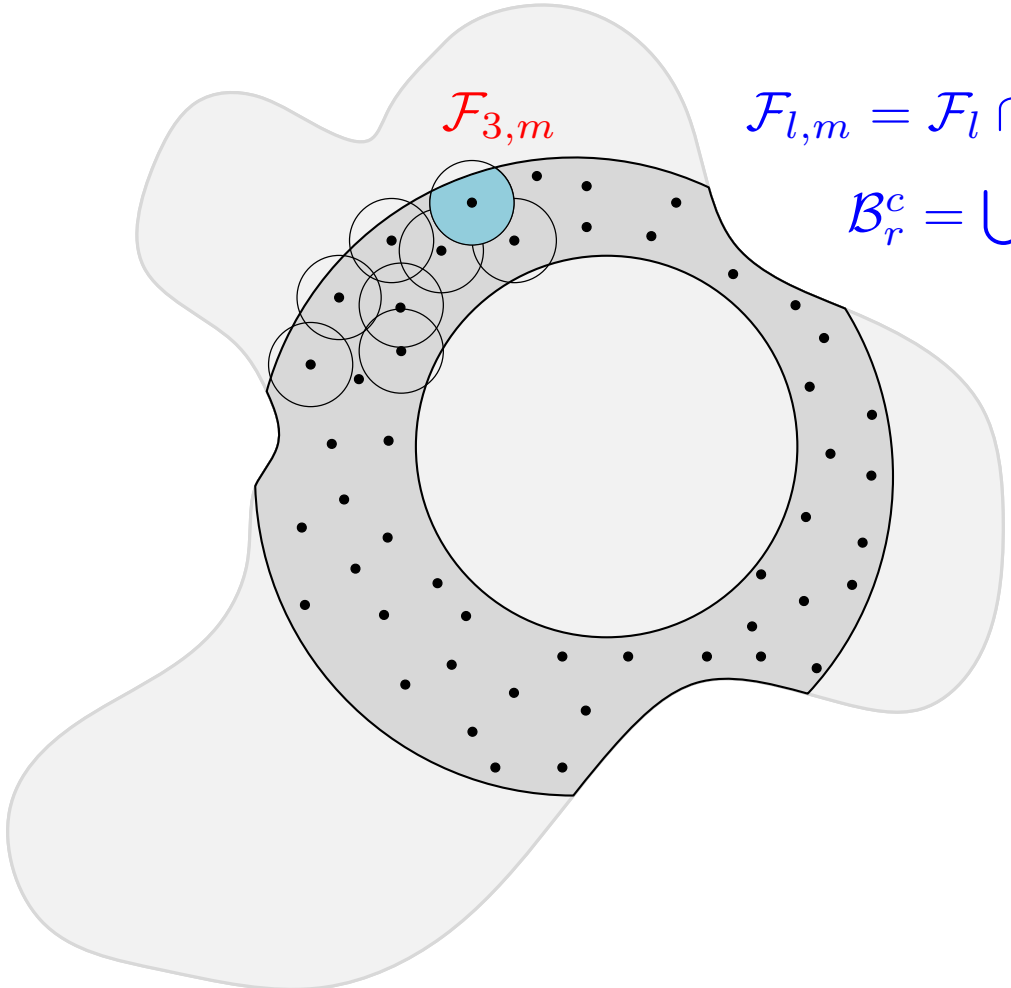
Let $r > 0$ and $\{r_l\}$ be a finite strictly decreasing sequence such that $r_1 = 1$ and $r_L = r$. Let $\mathcal{F}_l = \mathcal{B}_{r_l} \setminus \mathcal{B}_{r_{l+1}}$.





A covering for $\mathcal{B}_r^c \cap \Theta$

For each $\mathcal{F}_l$ find a maximal ε_l-separated set S_{ε_l}, with $K_l = |S_{\varepsilon_l}|$.



$$\mathcal{F}_{l,m} = \mathcal{F}_l \cap \mathcal{B}_{\varepsilon_l}(m \in S_{\varepsilon_l})$$

$$\mathcal{B}_r^c = \bigcup_{l=1}^{L-1} \bigcup_{m \in S_{\varepsilon_l}} \mathcal{F}_{l,m}$$

A condition on the initial beliefs

The initial beliefs of all agents are equal and have the following property:
For any constants $C \in (0, 1]$ and $r \in (0, 1]$ there exists a finite positive integer K, such that

$$\mu_0\left(\mathcal{B}_{\frac{C}{\sqrt{k}}}\right) \geq \exp\left(-k\frac{r^2}{32}\right) \qquad \text{for all } k \geq K.$$

Concentration Result for Compact Hypotheses sets

The beliefs $\{\mu_k^i\}$, generated by the update rule in Eq. (5) have the following property: with probability $1-\sigma$,

$$\mu_{k+1}^i(\mathcal{B}_r) \geq 1 - \chi \exp\left(-\frac{k}{16}r^2\right) \qquad \text{for all } i \text{ and all } k \geq \max\{N, K\}$$

where

$$N = \inf\left\{ t \geq 1 \middle| \exp\left(\log\frac{1}{\alpha}\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1} K_l \exp\left(-\frac{t}{32}r_{l+1}^2\right) < \frac{\sigma}{2} \right\},$$

with K as defined initial condition assumption,
$\chi = \sum_{l=1}^{L-1} \exp(-\frac{1}{16}r_{l+1}^2)$ and $\delta = 1 - \eta/n^2$, where η is the smallest positive element of the matrix A.

Distributed estimation for the exponential family

The exponential family, for a parameter $\theta = [\theta^1, \theta^2, \ldots, \theta^s]'$, is the set of probability distributions whose density can be represented as

$$p_{\chi,\nu}(\theta) = f(\chi, \nu) \exp(\theta' \chi - \nu C(\theta)),$$

We say, a prior is conjugate if for a likelihood of the form $p_\theta(x) = H(x) \exp(\theta' T(x) - C(\theta))$, the posterior distribution is

$$p_{\chi+T(x),\nu+1}(\theta|x) \propto p_\theta(x) p_{\chi,\nu}(\theta).$$

In our case, if all agents have conjugate beliefs to their corresponding models then, $d\mu_k^i(\theta) = p_{\chi_k^i,\nu_k^i}(\theta|x_k^i)$

$$\chi_{k+1}^i = \sum_{j=1}^n a_{ij} \chi_k^j + T^i(x^i), \qquad \nu_{k+1}^i = \sum_{j=1}^n a_{ij} \nu_k^j + 1$$

Gaussians: estimating the Mean with known variance

$$\min_{\theta} \sum_{i=1}^{n} D_{KL}(\mathcal{N}(\theta^i, (\sigma^i)^2) \| \mathcal{N}(\theta, (\sigma^i)^2))$$

which is equivalent to

$$\min_{\theta} \sum_{i=1}^{n} \frac{(\sigma^i)^{-2}(\theta^i - \theta)^2}{\sum_{j=1}^{n} (\sigma^j)^{-2}}$$

then

$$\tau_{k+1}^i = \sum_{j=1}^{n} a_{ij} \tau_k^j + \tau^i$$

$$\theta_{k+1}^i = \frac{\sum_{j=1}^{n} a_{ij} \tau_k^j \theta_k^j + x_{k+1}^i \tau^i}{\tau_{k+1}^i}$$

where $\tau_k^i = \frac{1}{(\sigma_k^i)^2}$.

Unknown Variance, known mean

$$\min_{\sigma^2} \sum_{i=1}^{n} D_{KL}(\mathcal{N}(\theta^i, (\sigma^i)^2) \| \mathcal{N}(\theta^i, (\sigma)^2))$$

which is equivalent to $\min_{\sigma^2} n \log \sigma^2 + \frac{\sum_{i=1}^{n} (\sigma^i)^2 + 4(\theta^i)^2}{2(\sigma)^2}$

then $\mu_k^i = \text{Inv}{-}\chi^2(\nu_k^i, (\tau_k^i)^2)$

$$\nu_{k+1}^i = \sum_{j=1}^{n} a_{ij} v_k^j + 1$$

$$(\tau_{k+1}^i)^2 = \frac{\sum_{j=1}^{n} a_{ij} v_k^j (\tau_k^j)^2 + (x_{k+1}^i - \theta^i)^2}{\nu_{k+1}^i}$$

Distributed Poisson Filter

$$\min_{\lambda} \sum_{i=1}^{n} D_{KL}(\text{Poisson}(\lambda^i) \| \text{Poisson}(\lambda))$$

which is equivalent to

$$\min_{\lambda} -\sum_{i=1}^{n} \lambda^i \log \lambda + \lambda$$

then $\mu_k^i = \text{Gamma}(\alpha_k^i, \beta_k^i)$

$$\alpha_{k+1}^i = \sum_{j=1}^{n} a_{ij} \alpha^j + x_{k+1}^i$$

$$\beta_{k+1}^i = \sum_{j=1}^{n} a_{ij} \beta^j + 1$$

Distributed Gaussian Filter: Unknown Mean, Unknown Variance

$$\min_{\theta,\sigma^2} \sum_{i=1}^{n} D_{KL}(\mathcal{N}(\theta^i, (\sigma^i)^2) \| \mathcal{N}(\theta, (\sigma)^2))$$

then

$$\tau_{k+1}^i = \sum_{j=1}^{n} a_{ij}\tau_k^j + 1, \qquad \theta_{k+1}^i = \frac{\sum_{j=1}^{n} a_{ij}\tau_k^j\theta_k^j + x_{k+1}^i}{\tau_{k+1}^i},$$

$$\alpha_{k+1}^i = \sum_{j=1}^{n} a_{ij}\alpha_k^j + 1/2, \quad \beta_{k+1}^i = \sum_{j=1}^{n} a_{ij}\beta_k^j + \frac{\sum_{j=1}^{n} a_{ij}\tau_k^j(x_{k+1}^i - \theta_k^j)^2}{2\tau_{k+1}^i}.$$

Conclusion

We studied the problem of distributed estimation. Starting from a variational interpretation of Bayes' Theorem, we propose a set of new algorithms with provable performance for a variety of graphs. We show non-asymptotic, explicit and geometric concentration rates around the correct hypotheses.

Questions?

If enough time, we can talk about two open problems on the relation with **Linear Regression** and the **Kalman filter** :).

Linear Observations and the Regression Problem

Consider two multivariate Normal distributions $P = \mathcal{N}(\theta_0, \Sigma_0)$ and $Q = \mathcal{N}(\theta_1, \Sigma_1)$

$$D_{KL}(P, Q) = \frac{1}{2}\left(\text{tr}(\Sigma_1^{-1}\Sigma_0) - (\theta_1 - \theta_0)'\Sigma_1^{-1}(\theta_1 - \theta_0) - k + \ln\left(\frac{\det\Sigma_1}{\det\Sigma_1}\right)\right)$$

In particular, the multivariate mean estimation problem is

$$\arg\min_{\theta\in\Theta} D_{KL}(P, P_\theta) = \arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P \|X - \theta\|^2_{\Sigma^{-1}}, \theta \sim \pi, X \sim P$$

This is the centralized problem where $X_k = \theta^* + \epsilon_k$, with $\epsilon_k \sim \mathcal{N}(0, \Sigma)$

Linear Observations and the Regression Problem

Now consider the network estimation problem were $X_k^i = C^{i\prime}\theta + \epsilon_k^i$, where $\theta \in \mathbb{R}^m$, $C^i \in \mathbb{R}^m$ and $\epsilon_k^i \sim \mathcal{N}(0, \Sigma)$. The optimization problem to be solved is then

$$\min_{\theta} \|\theta - \theta^*\|^2_{C\Sigma^{-1}C'}$$

and the resulting algorithm is

$$(\Sigma_{k+1}^i)^{-1} = \sum_{j=1}^{n} a_{ij}(\Sigma_k^j)^{-1} + C^i(\Sigma^i)^{-1}C^{i\prime}$$

$$\theta_{k+1}^i = \Sigma_{k+1}^i \left(\sum_{j=1}^{n} a_{ij}(\Sigma_k^j)^{-1}\theta_k^j + C^{i\prime}(\Sigma^i)^{-1}x_{k+1}^i \right)$$

Distributed Tracking and the Kalman Filter

Assume θ_k is a Markov process, and x_k are the observed states, then

$$\Pi(\theta_{k+1}|x^{k+1}) \propto p_{\theta_{k+1}}(x_{k+1}) \int_{\Theta} p(\theta_{k+1}|\theta_k) d\Pi(\theta_k|x^k)$$

From the belief update perspective, we can express this prediction+update procedure as

Prediction : $d\hat{\mu}_{k+1} = \int_{\Theta} p(\cdot|\theta) d\mu_k(\theta)$

Update : $d\mu_{k+1} = \underset{\pi \in \Delta_{\Theta}}{\arg\min} \{\langle -\log p_{\theta}(x_{k+1}), \pi \rangle + D_{KL}(\pi \| d\hat{\mu}_{k+1})\}$

One particular case is when θ_k and x_k evolve as a discrete-time Linear Gaussian system, where

$$\theta_{k+1} = A_k\theta_k + W_k$$
$$X_k = C_k\theta_k + V_k$$

with $W_k \sim \mathcal{N}(0, Q_k)$ and $V_k \sim \mathcal{N}(0, R_k)$.
Starting with a Gaussian prior $d\mu_0 = \mathcal{N}(\hat{\theta}_{0|0}, \Sigma_{0|0})$

$$d\hat{\mu}_k = \mathcal{N}(A_k\hat{\theta}_{k-1|k-1}, A_k\Sigma_{k-1|k-1}A_k' + Q_k)$$
$$d\mu_k = \mathcal{N}(\hat{\theta}_{k|k}, \Sigma_{k|k})$$

and

$$\tilde{x}_k = x_k - C_k\hat{\theta}_{k|k-1} \qquad S_k = H_k\Sigma_{k|k-1}H_k' + R_k$$
$$K_k = \Sigma_{k|k-1}H_k'S_k^{-1}$$
$$\hat{\theta}_{k|k} = \hat{\theta}_{k|k-1} + K_k\tilde{x}_k \qquad \Sigma_{k|k} = (I - K_kH_k)\Sigma_{k|k-1}$$

A distributed Kalman filter

If the predicted beliefs are shared, we can propose a distributed Kalman filter of the form

$$d\hat{\mu}^i_{k+1} = \int_\Theta p(\cdot|\theta) d\mu^i_k(\theta)$$

$$d\mu^i_{k+1} = \arg\min_{\pi \in \Delta_\Theta} \left\{ \langle -\log p_\theta(x_{k+1}), \pi \rangle + \sum_{j=1}^{n} a_{ij} D_{KL}(\pi \| d\hat{\mu}^j_{k+1}) \right\}$$