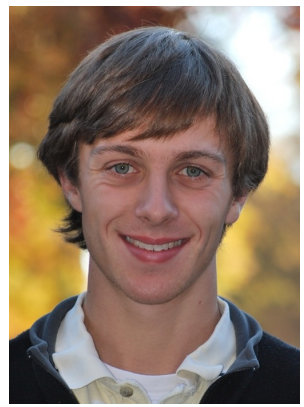
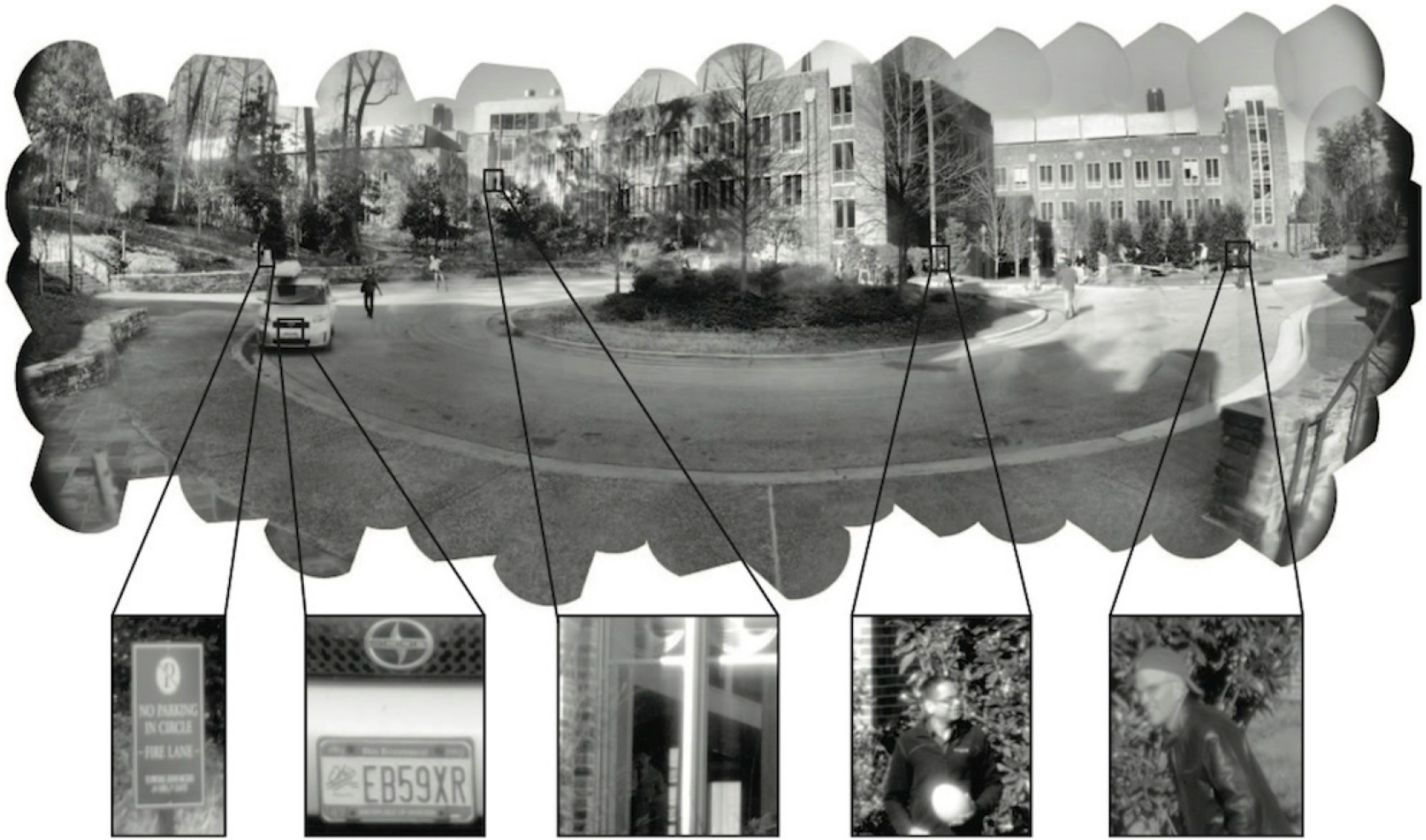
Dynamical models and tracking regret in online convex programming

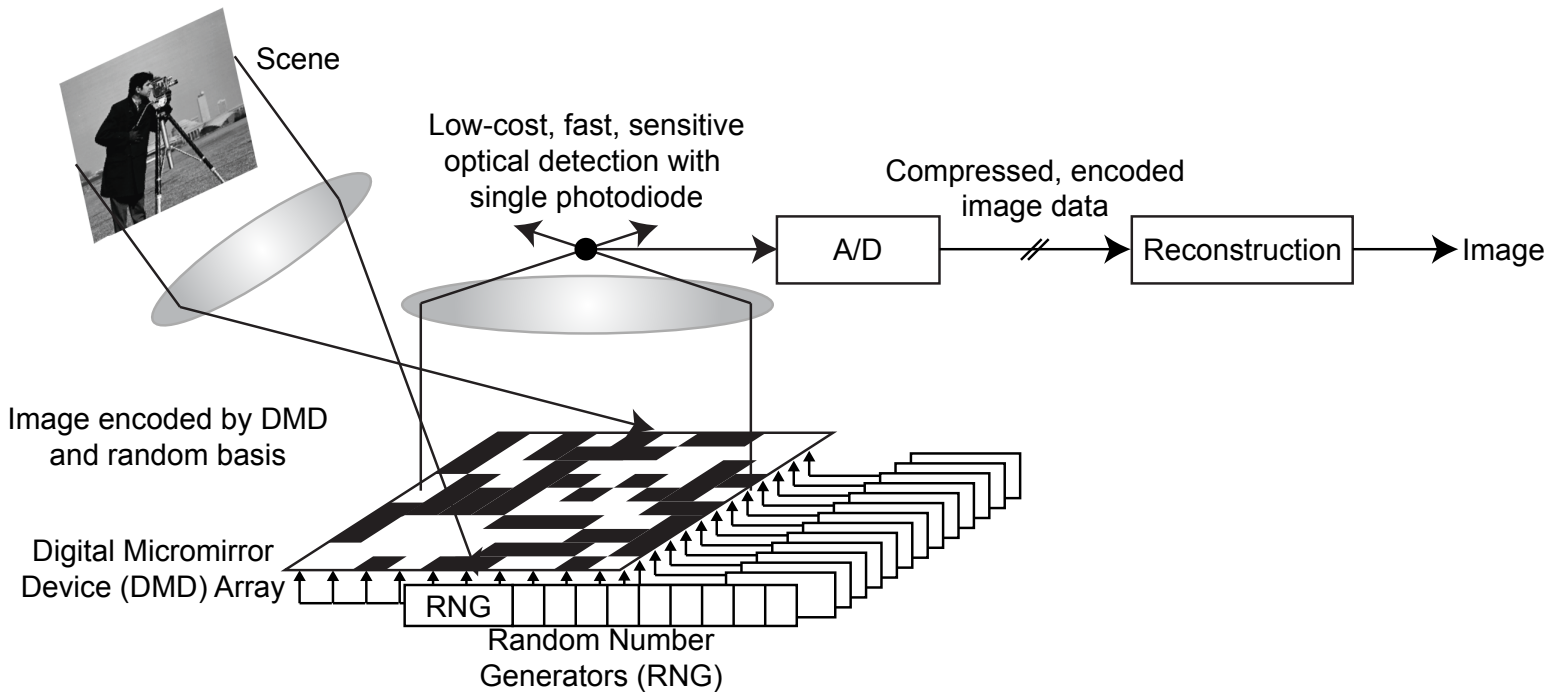
Rebecca Willett, Duke University



Joint work
with Eric Hall


NO PARKING
IN CIRCLE
FIRE LANE
EB59XR

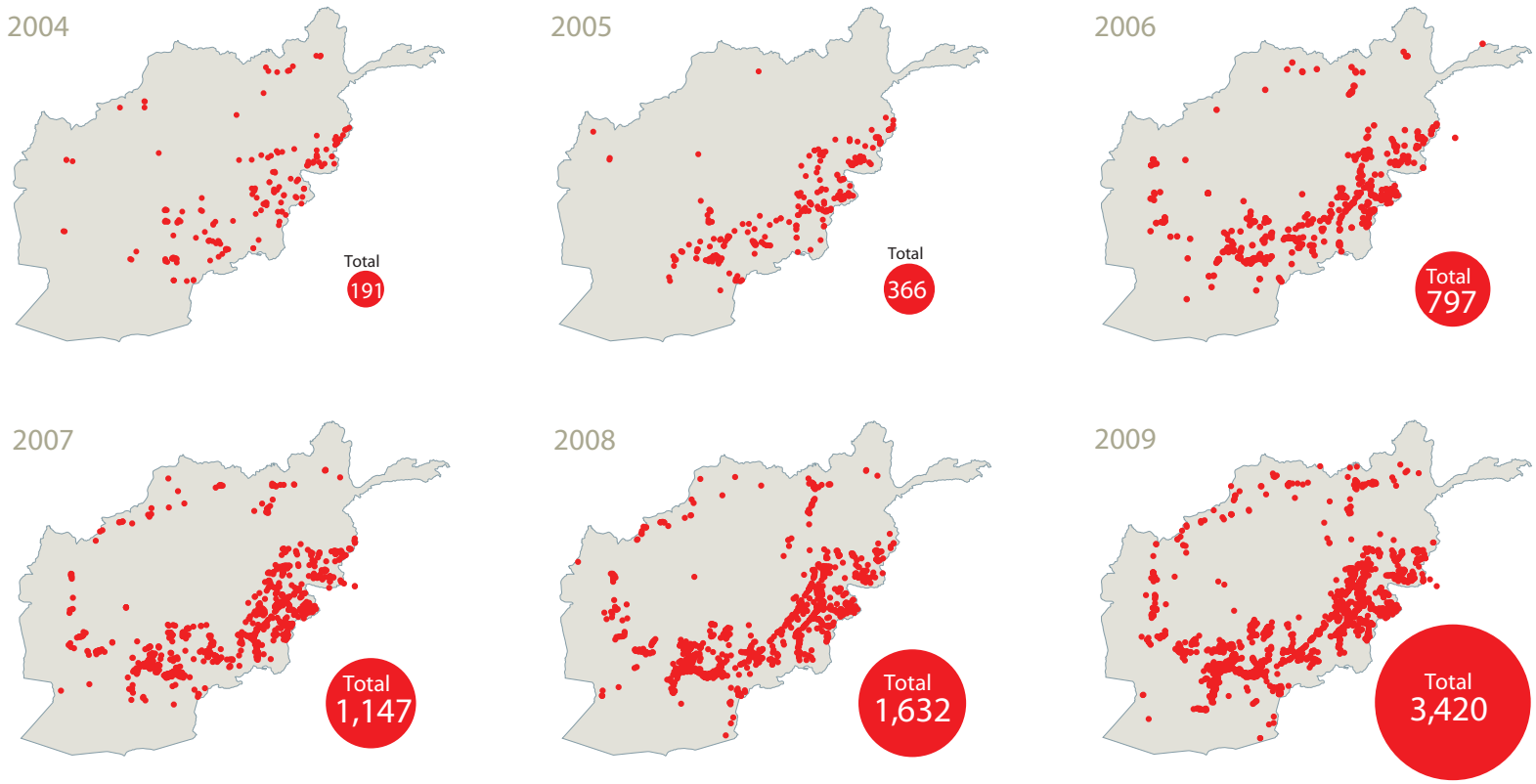
Sequential compressed sensing



We collect compressed sensing projections sequentially.

Can we quickly reconstruct a dynamic scene?

IED tracking and prediction



We sequentially observe IED locations and salient features.

Can we predict likely future locations?

Online social network inference

We monitor meetings and communications in a social network.

Can we track the dynamic network structure?

Big challenges processing big data

- Limited processing power, memory, and bandwidth: data is of such **large volume** that it cannot be stored and processed with batch algorithms
- **High velocity**: data arrives sequentially
- **No prior** notion of what is "significant"
- "Typical" behavior can change over time: we must **adapt quickly** to sharp changes in the environment
- Environmental **dynamics** are unknown
- Observations may have unknown **dependencies** or may not be **stochastic**
- Data may have **corrupted or missing elements**

Stochastic filters

Stochastic filters allow for (sometimes) fast predictions in dynamic environments, but...

- Their **application** relies on having a known dynamical state-space model
- Their **analysis** relies upon strong modeling assumptions (*e.g.* Gaussian processes) and does is not robust to model mismatch

Problem Formulation

Sequence of events: set initial "prediction" $\widehat{\theta}_1$. At time t:

1. Observe datum x_t
2. Incur loss
$$\ell_t(\widehat{\theta}_t) \triangleq \underbrace{f_t(\widehat{\theta}_t; x_t)}_{\text{data fit}} + \underbrace{r(\widehat{\theta}_t)}_{\text{regularizer}}$$
3. Make a prediction, $\widehat{\theta}_{t+1}$

Definition: The **Regret** of $\widehat{\boldsymbol{\theta}}_T = (\widehat{\theta}_1, \ldots, \widehat{\theta}_T)$ with respect to a comparator $\boldsymbol{\theta}_T = (\theta_1, \ldots, \theta_T)$ is

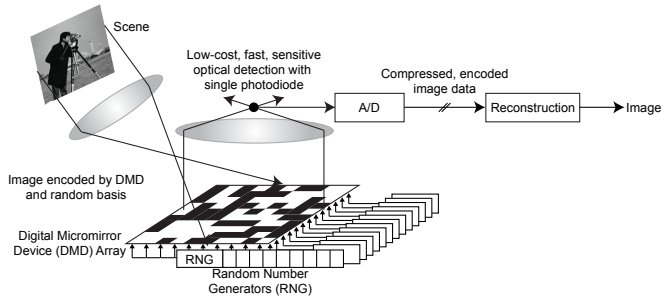
$$R_T(\boldsymbol{\theta}_T) \triangleq \sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) - \sum_{t=1}^{T} \ell_t(\theta_t).$$

Goal: Generate losses comparable to what a batch algorithm might achieve; *i.e.*, **sublinear regret**:

$$\frac{1}{T} R_T(\boldsymbol{\theta}_T) \to 0 \text{ as } T \to \infty$$

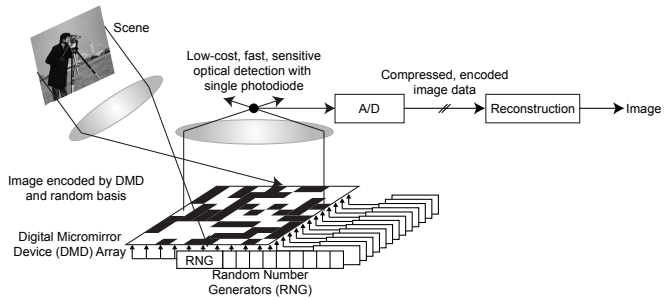
Examples

- Online compressed sensing:
  - $x_t =$ light intensity measurement
  - $\theta_t =$ scene at time t
  - $\ell_t(\theta) = \|x_t - \langle a_t, \theta \rangle\|_2^2 + \tau \|\theta\|_1$



Examples

- Online compressed sensing:
  - x_t = light intensity measurement
  - θ_t = scene at time t
  - $\ell_t(\theta) = \|x_t - \langle a_t, \theta \rangle\|_2^2 + \tau\|\theta\|_1$
- Social networks:
  - x_t = agent actions (e.g. communications or votes)
  - θ_t = amount of influence agents have on one another
  - $\ell_t(\theta) = -\log p(x_t|\theta) + \tau\|\theta\|_1$ where p is a pseudolikelihood



Salient features of approach

- **causality** — for each t, $\widehat{\theta}_t$ may depend only on $x_{t-1} = (x_1, \ldots, x_{t-1})$
- **(generalized) sparsity** — while $\widehat{\theta}_t$ may be very high-dimensional, it should exhibit low-dimensional structure or a parsimonious representation.
- **dynamic** — incorporates or learns environmental dynamics

Good

Bad *and* Ugly

Bregman divergence and strong convexity

If ℓ is a convex function, then $\forall\ \theta, \theta'$

$$\ell(\theta) - \ell(\theta') - \left\langle \nabla \ell(\theta'), \theta - \theta' \right\rangle \geq 0$$

If ℓ is a σ-strongly convex with respect to $\|\cdot\|$, then $\forall\ \theta, \theta'$

$$\ell(\theta) - \ell(\theta') - \left\langle \nabla \ell(\theta'), \theta - \theta' \right\rangle \geq \frac{\sigma}{2}\|\theta - \theta'\|^2$$

Let $\psi(\theta)$ be a σ-strongly convex function with respect to $\|\cdot\|$. The Bregman divergence associated with ψ is

$$D(\theta, \theta') \equiv D_\psi(\theta, \theta') \triangleq \psi(\theta) - \psi(\theta') - \left\langle \nabla \psi(\theta'), \theta - \theta' \right\rangle$$

Mirror Descent[1]

$$\widehat{\theta}_{t+1} = \arg\min_{\theta} \eta_t \left\langle \nabla \ell_t(\widehat{\theta}_t), \theta \right\rangle + D(\theta, \widehat{\theta}_t)$$

- $\nabla \ell_t$ is an arbitrary subgradient of ℓ_t
- η_t is the step size
- Special case where $D(\theta, \theta') = \|\theta - \theta'\|^2$:

$$\widehat{\theta}_{t+1} \equiv \widehat{\theta}_t - \frac{1}{\eta_t} \nabla \ell_t(\widehat{\theta}_t)$$

[1] Nemirovski & Yudin 1983; Beck & Teboulle 2003; Zinkevich 2003

Mirror Descent

$$\begin{aligned}\widehat{\theta}_{t+1} &= \arg\min_{\theta} \eta_t \left\langle \nabla \ell_t(\widehat{\theta}_t), \theta \right\rangle + D(\theta, \widehat{\theta}_t) \\ &= \arg\min_{\theta} \eta_t \left\langle \nabla f_t(\widehat{\theta}_t), \theta \right\rangle + \left\langle \nabla r(\widehat{\theta}_t), \theta \right\rangle + D(\theta, \widehat{\theta}_t)\end{aligned}$$

Problem: when $r(\theta)$ corresponds to a sparsity penalty, the estimate $\widehat{\theta}_{t+1}$ is *close to* sparse, but not actually sparse

- increased computational burden
- more storage requirements
- less interpretable results

Composite Objective Mirror Descent (COMD)[2]

$$\widehat{\theta}_{t+1} = \arg\min_{\theta} \eta_t \left\langle \nabla f_t(\widehat{\theta}_t), \theta \right\rangle + \eta_t r(\theta) + D(\theta, \widehat{\theta}_t)$$

or equivalently (in unconstrained cases)

(1) move in direction of negative gradient of f_t

$$\widetilde{\theta}_{t+1} = \arg\min_{\theta} \eta_t \left\langle \nabla f_t(\widehat{\theta}_t), \theta \right\rangle + D(\theta, \widehat{\theta}_t)$$

(2) regularize

$$\widehat{\theta}_{t+1} = \arg\min_{\theta} \eta_t r(\theta) + D(\theta, \widetilde{\theta}_{t+1})$$

[2] Duchi, Shalev-Shwartz, Singer and Tewari, 2010

Static regret bounds

Theorem[3]**:** Assume $\boldsymbol{\theta}_T$ is static, so that $\theta \triangleq \theta_1 = \theta_2 = \ldots = \theta_T$. If $\eta_t \propto 1/\sqrt{T}$, then

$$R_T(\boldsymbol{\theta}_T) = O\left(\sqrt{T}\right).$$

Furthermore, if r is strongly convex and $\eta_t \propto 1/t$, then

$$R_T(\boldsymbol{\theta}_T) = O\left(\log T\right).$$

So what's missing?

- Comparing against a static model is weak; how do we do relative to a dynamic comparator?
- What about dynamic environments?
- What about unknown environmental dynamics?

[3] Duchi *et al.* 2010

Tracking, Shifting, and Adaptive Regret

- Adaptive Regret[4] looks at accumulated loss over an arbitrary time interval of length τ relative to a static comparator:

$$R_\tau \triangleq \max_{\substack{[r,s] \subset [1,T]: \\ s-r \leq \tau}} \sum_{t=r}^{s} \ell_t(\widehat{\theta}_t) - \min_\theta \sum_{t=r}^{s} \ell_t(\theta)$$

- Shifting and Tracking regret[5]
  - Compare output of algorithm to sequence $(\theta_1, \ldots, \theta_T)$ which can be chosen collectively with full knowledge of all data
  - Typical bounds vary with complexity of comparator sequence

[4] Littlestone & Warmuth 2001, Hazan & Seshadhri 2009

[5] Herbster & Warmuth 2001, Cesa-Bianchi & Lugosi 2006, Cesa-Bianchi *et al.* 2012

COMD Regret against time-varying reference models

Theorem: If $\eta_t = 1/\sqrt{t}$, then

$$R_T(\boldsymbol{\theta}_T) = O\left((V_T(\boldsymbol{\theta}_T) + 1)\sqrt{T}\right),$$

where

$$V_T(\boldsymbol{\theta}_T) \triangleq \sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|$$

measures the temporal variation in $\boldsymbol{\theta}_T$.

In other words, the algorithm can track a dynamically changing environment, provided the **changes are sufficiently infrequent and/or smooth (restrictive!)**

Dynamic Mirror Descent (DMD)

Our approach: Let $\Phi_t : \Theta \mapsto \Theta$ be a series of predetermined dynamical models; set

$$\widetilde{\theta}_{t+1} = \arg\min_{\theta} \eta_t \langle \nabla f_t(\Phi_t \widetilde{\theta}_t), \theta \rangle + \eta_t r(\theta) + D(\theta \| \Phi_t \widetilde{\theta}_t)$$

$$\widehat{\theta}_{t+1} = \Phi_{t+1} \widetilde{\theta}_{t+1}$$

Theorem: Assume each Φ_t is contractive, so that

$$D(\Phi_t \theta \| \Phi_t \theta') \leq D(\theta \| \theta') \ \ \forall \theta, \theta'.$$

Then if $\eta_t \propto \frac{1}{\sqrt{t}}$ we have $R_T(\boldsymbol{\theta}_T) \leq O([1 + V_\Phi(\boldsymbol{\theta}_T)]\sqrt{T})$ where

$$V_\Phi(\boldsymbol{\theta}_T) \triangleq \sum_{t=1}^{T} \|\theta_{t+1} - \Phi_{t+1}\theta_t\|$$

measures **the deviation of the comparator from the dynamic models (Φ_ts)**.

Sketch proof

Consider no regularization, $\Phi_t = \Phi \ \ \forall t$.

Recall

$$\widetilde{\theta}_{t+1} = \arg\min_{\theta} \eta_t \langle \nabla f_t(\Phi_t \widetilde{\theta}_t), \theta \rangle + D(\theta \| \Phi_t \widetilde{\theta}_t).$$

By the first-order optimality condition,

$$\langle \eta_t \nabla f_t(\Phi \widetilde{\theta}_t), \widetilde{\theta}_{t+1} - \theta \rangle \leq \langle \nabla \psi(\Phi \widetilde{\theta}_t) - \nabla \psi(\widetilde{\theta}_{t+1}), \widetilde{\theta}_{t+1} - \theta \rangle \ \ \forall \theta. \ \ (1)$$

Now

$$\begin{aligned}
f_t(\widehat{\theta}_t) - f_t(\theta_t) &= f_t(\Phi \widetilde{\theta}_t) - f_t(\theta_t) \\
&\leq \langle \nabla f_t(\Phi \widetilde{\theta}_t), \Phi \widetilde{\theta}_t - \theta_t \rangle && \text{(convexity of } f_t\text{)} \\
&\leq (1/\eta_t) \langle \nabla \psi(\Phi \widetilde{\theta}_t) - \nabla \psi(\widetilde{\theta}_{t+1}), \widetilde{\theta}_{t+1} - \theta_t \rangle \\
&\quad + \langle \nabla f_t(\Phi \widetilde{\theta}_t), \Phi \widetilde{\theta}_t - \theta_t \rangle && \text{by (1)}
\end{aligned}$$

Sketch proof (cont.)

$$
\begin{aligned}
& f_t(\widehat{\theta}_t) - f_t(\theta_t) \\
& \quad \leq \frac{1}{\eta_t} \langle \nabla\psi(\Phi\widetilde{\theta}_t) - \nabla\psi(\widetilde{\theta}_{t+1}), \widetilde{\theta}_{t+1} - \theta_t \rangle && \text{(from before)} \\
& \qquad + \langle \nabla f_t(\Phi\widetilde{\theta}_t), \Phi\widetilde{\theta}_t - \theta_t \rangle \\
& \quad = \frac{1}{\eta_t} \Big[D(\theta_t \| \Phi\widetilde{\theta}_t) - D(\theta_{t+1} \| \Phi\widetilde{\theta}_{t+1}) && \text{(telescopes)} \\
& \qquad + D(\theta_{t+1} \| \Phi\widetilde{\theta}_{t+1}) - D(\Phi\theta_t \| \Phi\widetilde{\theta}_{t+1}) && \text{(bounded by } \|\theta_{t+1} - \Phi\theta_t\|\text{)} \\
& \qquad + D(\Phi\theta \| \Phi\widetilde{\theta}_{t+1}) - D(\theta_t \| \widetilde{\theta}_{t+1}) && (\leq 0 \text{ because } \Phi \text{ contractive}) \\
& \qquad - D(\widetilde{\theta}_{t+1} \| \Phi\widetilde{\theta}_t) \Big] && (= O(\eta_t)) \\
& \qquad + \langle \nabla f_t(\Phi\widetilde{\theta}_t), \Phi\widetilde{\theta}_t - \widetilde{\theta}_{t+1} \rangle
\end{aligned}
$$

Collection of Dynamical Models

Consider a collection of dynamics $\{\Phi_1, \Phi_2, \ldots \Phi_N\}$ and a comparator sequence that can change dynamics at unknown times t_i, $1 = t_1 < \cdots < t_{m+1} = T + 1$

$$\widehat{\theta}_t^{(i)} = \text{DMD prediction using } \Phi_i$$

$$\widetilde{w}_{i,t} = w_{i,t-1} \exp(-\eta_r \ell_t(\widehat{\theta}_t^{(i)}))$$

$$w_{i,t} = (\lambda/N) \sum_{j=1}^{N} \widetilde{w}_{j,t} + (1-\lambda)\widetilde{w}_{i,t} \quad \text{(weight on } i^{\text{th}} \text{ dynamics)}$$

$$\widehat{\theta}_t = \sum_{i=1}^{N} w_{i,t}\widehat{\theta}_t^{(i)} / \sum_{i=1}^{N} w_{i,t} \quad \text{(weighted combination of individual dynamics' predictions)}$$

DFS Regret Bound

$$R_T(\boldsymbol{\theta}_T) \leq O\Big(\sqrt{T}\Big[4V^{(m)}(\boldsymbol{\theta}_T) + m\log N - \log[\lambda^m(1-\lambda)^{(T-m-1)}]\Big]\Big)$$

where

$$V^{(m)}(\boldsymbol{\theta}_T) \triangleq \min_{t_1,\ldots,t_{m+1}} \sum_{k=1}^{m} \min_{i_k \in \{1,\ldots,N\}} \sum_{t=t_k}^{t_{k+1}-1} \|\Phi_{i_k}\theta_t - \theta_{t+1}\|$$

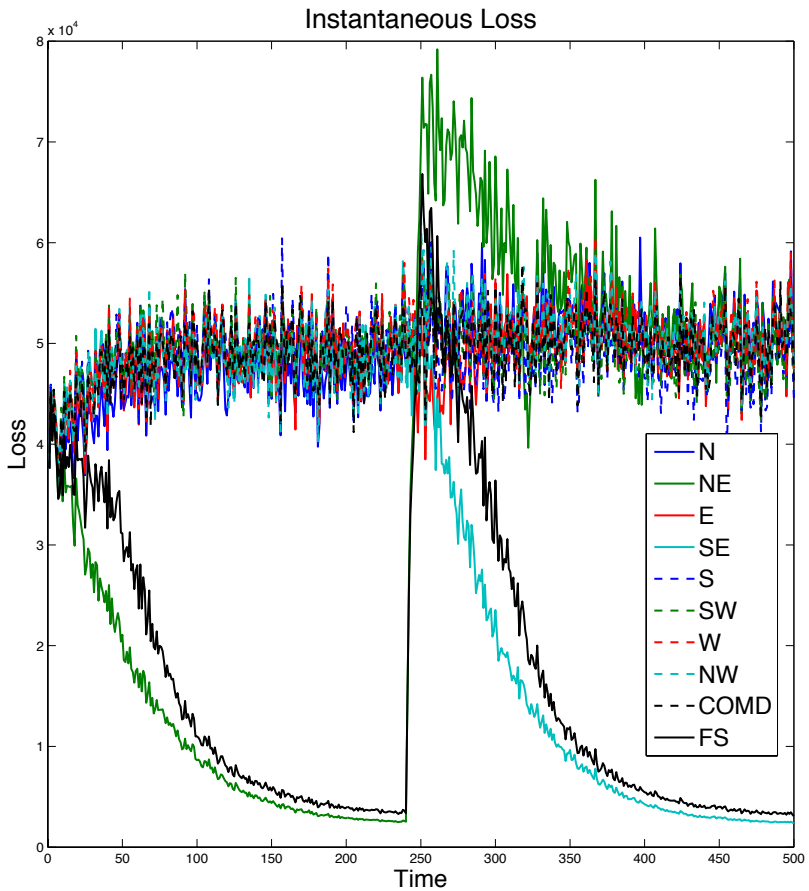
Analogous to $V_\Phi(\theta_T)$ with different dynamics on different intervals.

In other words, regret scales with how much best comparator sequence of predictions (that someone with infinite computational power and non-sequential data might compute) deviates from the **best series of dynamical models** one might choose to fit that comparator.

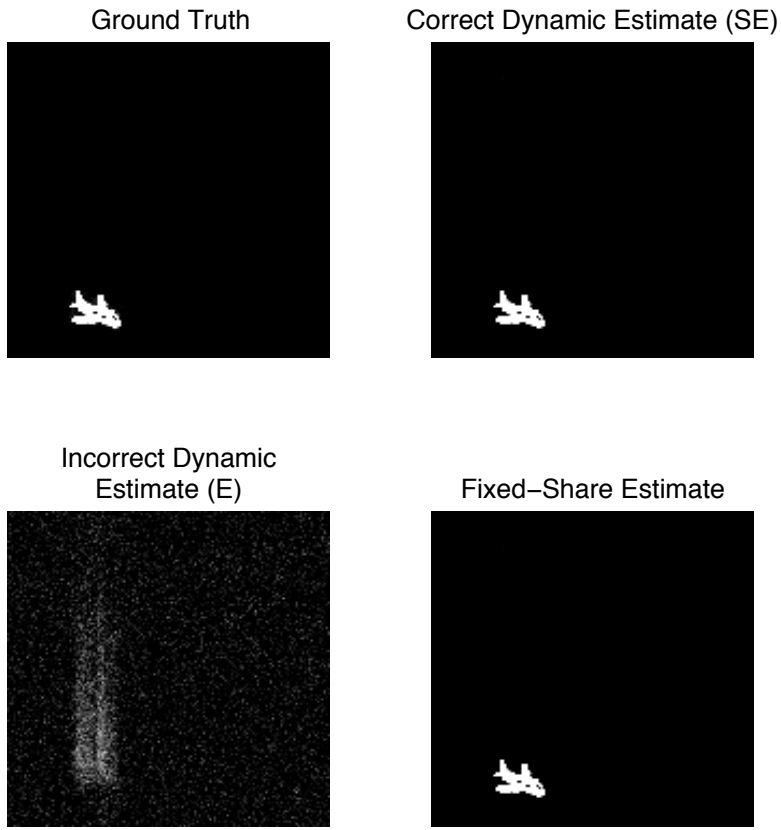
Online video compressed sensing

Observe $x_t = A_t\theta_t + n_t$, $A_t \in \mathbb{R}^{500\times 22500}$

Online video compressed sensing



(a)



(b)

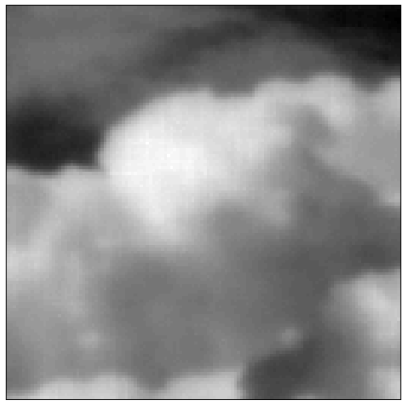
Left Plot: Instantaneous loss of a collection of DMD predictions.

Right Plots: Example estimates at $t = 480$.

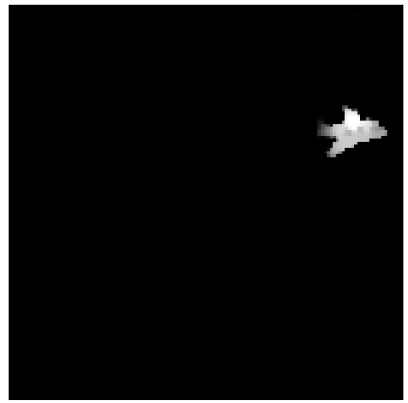
Poisson video foreground and background separation

Truth

Observation



Foreground Est.



Background Est.

Online solar data analysis, full data

Process all pixels from every frame, look for anomalies in peaks in losses

Online solar data analysis, 25% missing

Process 75% of pixels from each frame, look for anomalies in peaks in losses

Online solar data analysis, 50% missing

Process 50% of pixels from each frame, look for anomalies in peaks in losses

Online solar data analysis, 75% missing

Process 25% of pixels from each frame, look for anomalies in peaks in losses

Example: online social network inference

Consider the following problem:

> Given sequential observations of agents' behavior in a social network, perform online inference of network structure

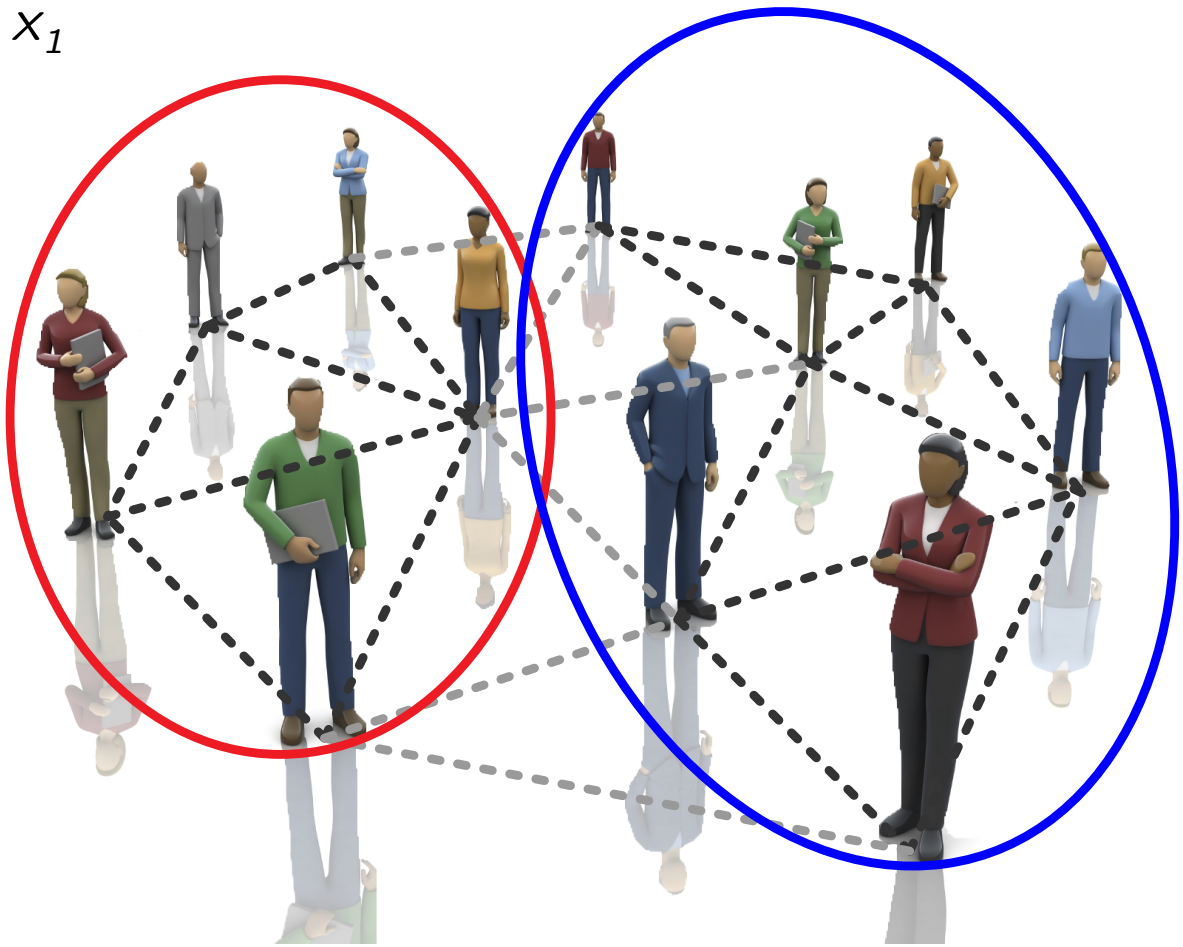
Individual sequence (universal prediction) setting:

- No probabilistic generative model of agents' behavior
- Instead, pick a suitable reference class of models and seek to minimize regret — difference between cumulative performance of our online scheme and that of the best model selected in hindsight given the entire observation sequence

Online social network inference

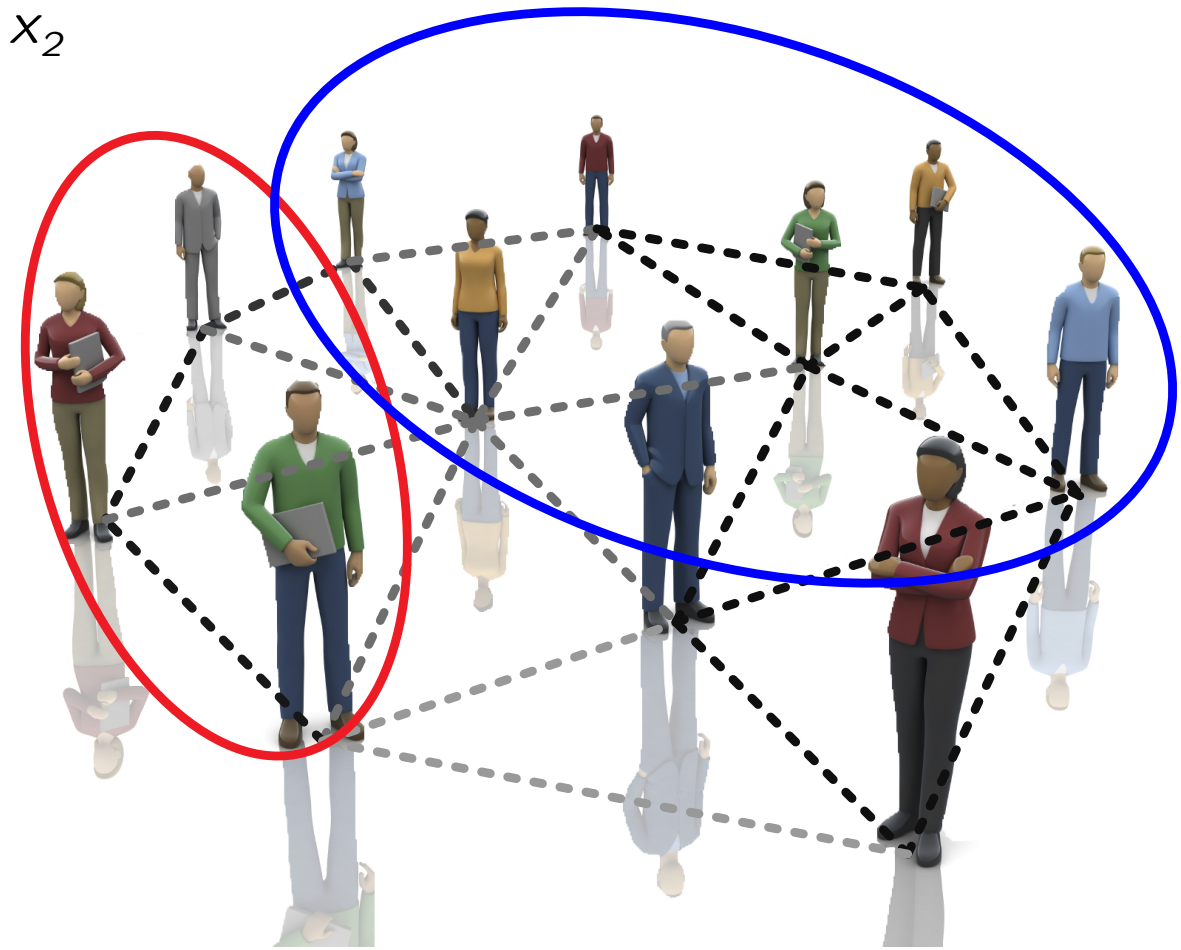
Start with initial estimate of network structure.

Online social network inference



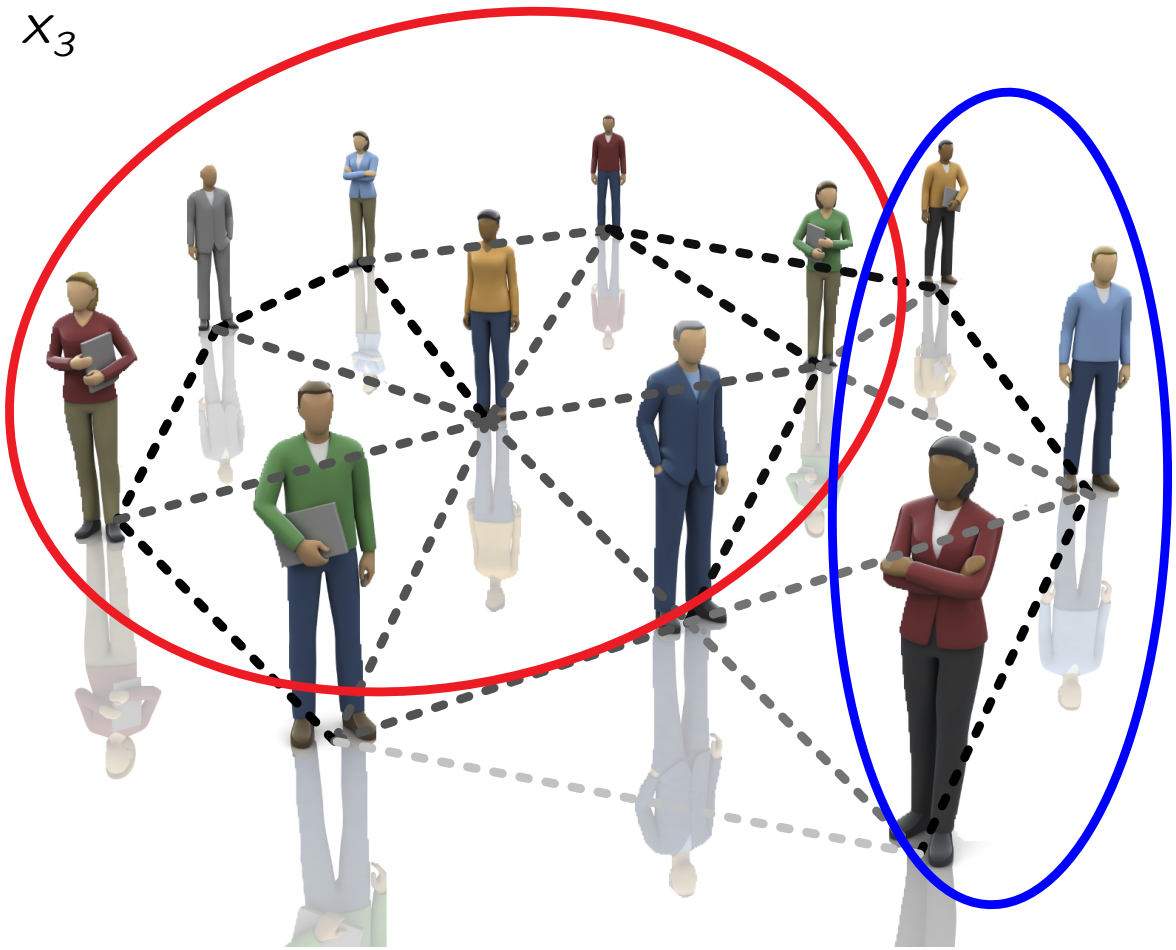
Observe first meeting or communication, see how well it fits initial prediction, and update network estimate.

Online social network inference



Observe second meeting or communication, see how well it fits previous prediction, and update network estimate.

Online social network inference



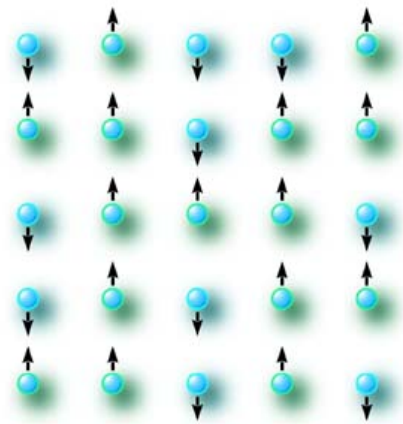
Observe third meeting or communication, see how well it fits previous prediction, and update network estimate.

Ising models

- $\mathcal{V} = \{1, \ldots, p\}$ is set of p agents making up network
- For each $\alpha \in \mathcal{V}$, let $x_\alpha \in \{-1, +1\}$ denote the action of that agent; $x = (x_\alpha : \alpha \in \mathcal{V})$
- Influence (precision) matrix $\theta = (\theta_{\alpha\beta})_{\alpha,\beta\in\mathcal{V}}$
- Ising model distribution is

$$\mathbb{P}_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{\alpha,\beta\in\mathcal{V}} \theta_{\alpha\beta} x_\alpha x_\beta \right\}$$

- $Z(\theta)$ is normalization factor known as the *partition function*.



Pseudolikelihood[6]

- Ising model $\mathbb{P}_\theta(x)$ problematic because $Z(\theta)$ not efficiently computable
- Consider instead distribution of x_α given the rest of the network:

$$\mathbb{P}_\theta(x_\alpha | x_{\setminus\alpha}) = \frac{\exp\left[2x_\alpha\left(\theta_{\alpha\alpha} + \sum_{\beta\in\mathcal{V}\setminus\alpha}\theta_{\alpha\beta}x_\beta\right)\right]}{\exp\left[2x_\alpha\left(\theta_{\alpha\alpha} + \sum_{\beta\in\mathcal{V}\setminus\alpha}\theta_{\alpha\beta}x_\beta\right)\right] + 1}$$

[6] Ravikumar, Wainwright & Lafferty 2009; Höfling & Tibshirani 2009

Online Ising loss function

Data fit: Let

$$f_t^{(\alpha)}(\theta) \triangleq -\log\left(\mathbb{P}_\theta(x_{t,\alpha}|x_{t,\setminus\alpha})\right)$$

and

$$f_t(\theta) \triangleq \sum_{\alpha\in\mathcal{V}} f_t^{(\alpha)}(\theta);$$

- convex
- computable loss and computable gradient

Regularization: Let

$$r(\theta) = \tau\|\theta\|_1$$

where τ is a tuning parameter.

Senate Roll Call

Senate Data Set: US Senate voting records from 1795 to 2011.

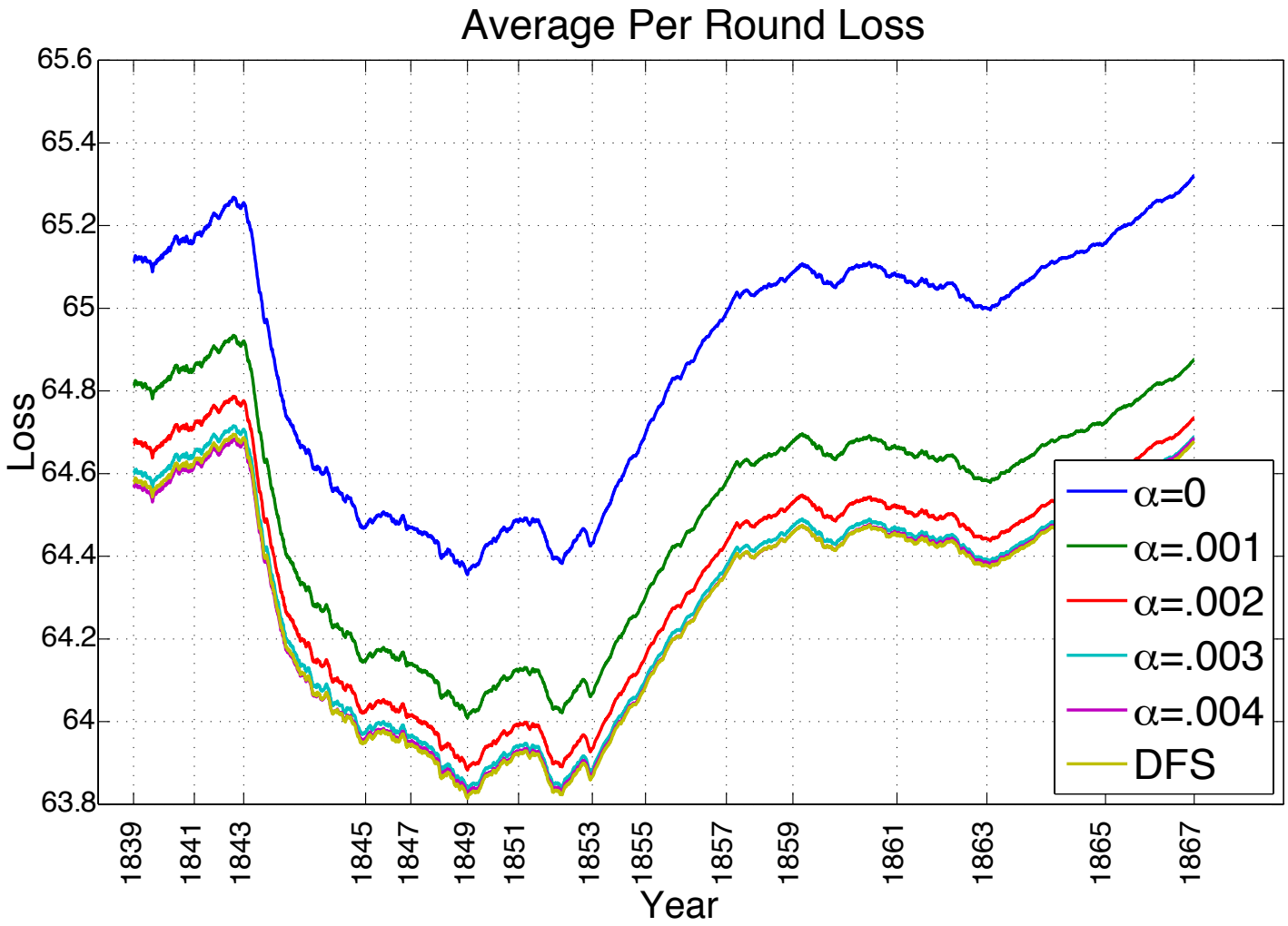
At time t, observe the "yea" (+1), "nay" (-1) or absent (0) vote of $p = 100$ Senators in vector x_t.

$\theta_t \in [-1, 1]^{p \times p}$, is the influence matrix where $(\theta_t)_{ab}$ corresponds to voting influence between agents a and b at time t.

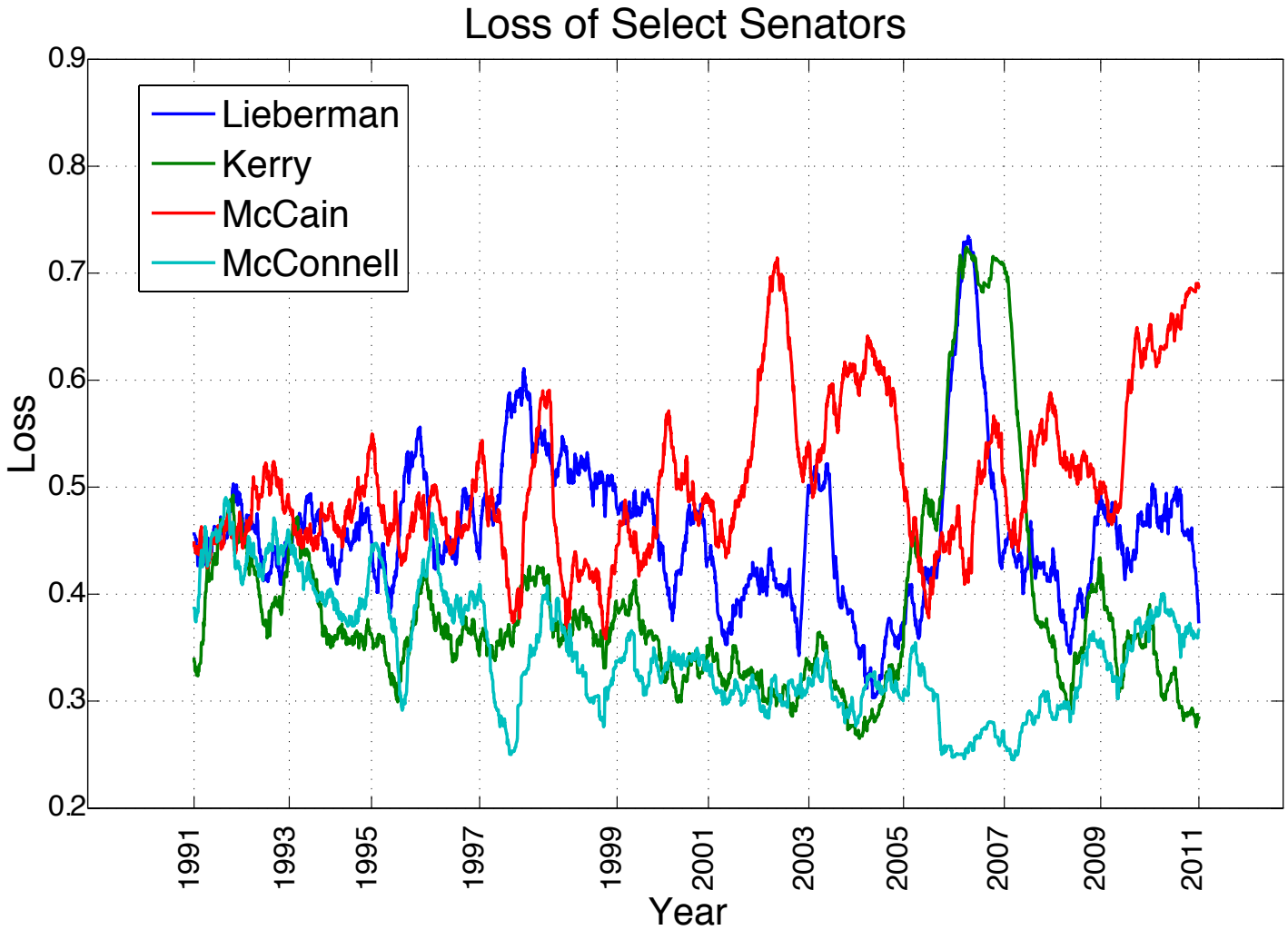
Dynamics: let $c^* = \arg\max_c |\theta_{ac}\theta_{bc}|$

$$(\Phi_i\theta)_{ab} = \begin{cases} (1-\alpha_i)\theta_{ab} + \alpha_i\theta_{ac^*}\theta_{bc^*} & \text{if } |\theta_{ac^*}\theta_{bc^*}| > |\theta_{ab}| \\ \theta_{ab} & \text{otherwise} \end{cases},$$

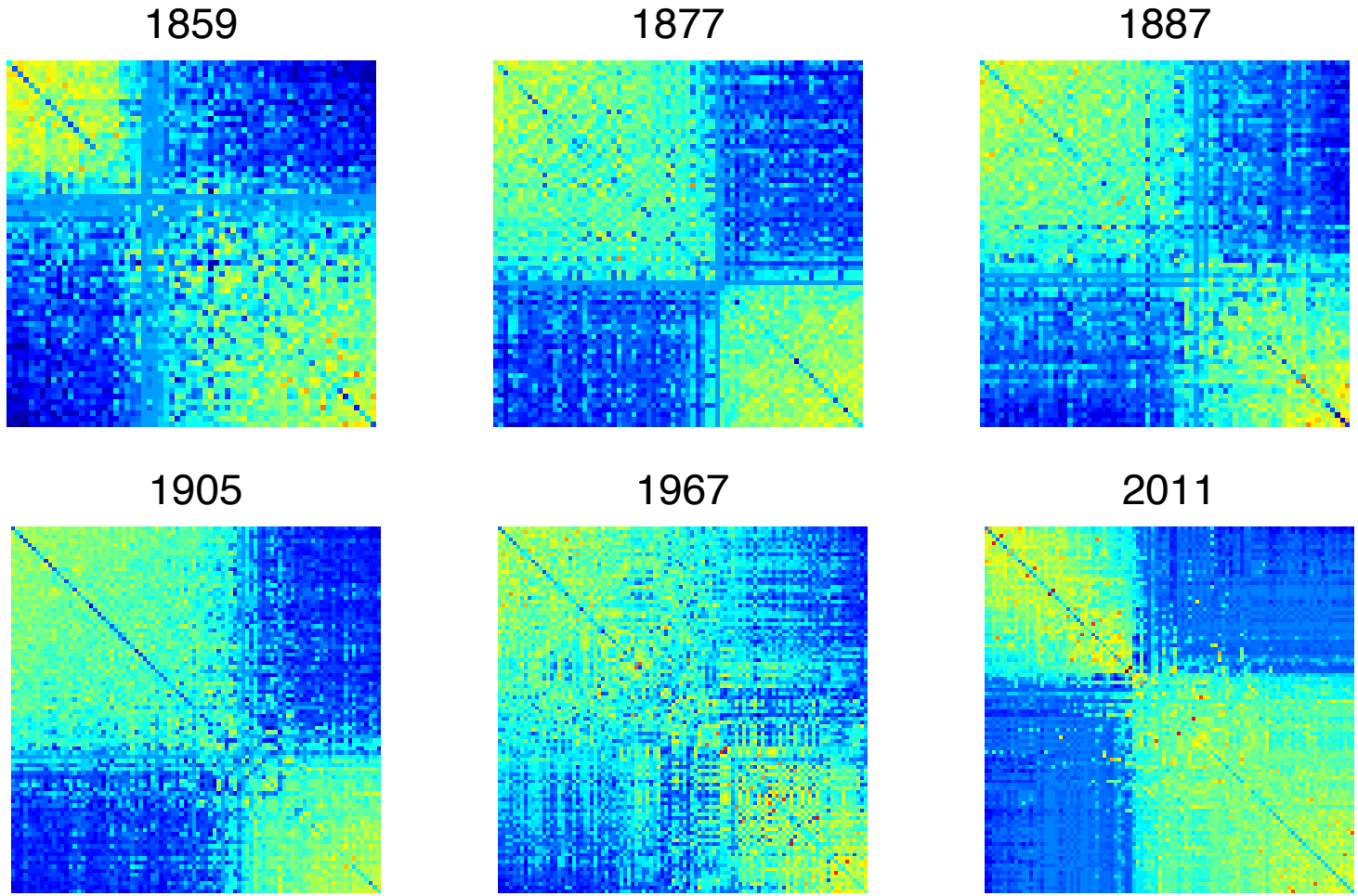
(says that two agents' correspondence will grow if they have a strong common bond)



Average per round loss of each model, and the DFS estimator. Applying the dynamical model improves performance relative to COMD ($\alpha_i = 0$). DFS estimator aggregates the predictions successfully.



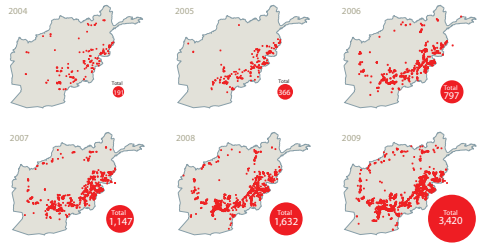
Moving average losses for select individual senators. Low losses correspond to consistent voting behavior. Notice, for instance, that John Kerry (D-MA) has generally very low loss, but spikes around 2006, and drops again before a reelection campaign in 2008.

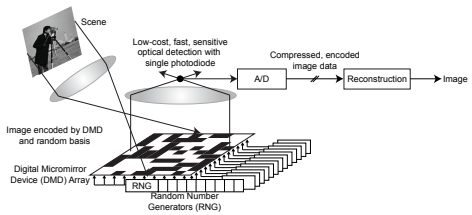


Influence matrices for select years spanning Civil War and Civil Rights Movement to present. We see tight factions forming in the mid- to late-1800s (post Civil War), and less so in the mid-1900s during the Civil Rights Movement and upheaval among southern Democrats.

Conclusions

- Our techniques offer principled mechanisms for using *streaming, high-dimensional data* to track *dynamic, uncertain environments*
- Flexible tools are applicable across a broad range of applications:
  - video surveillance
  - social network analysis
  - data thinning
  - monitoring financial transactions
- Computation scales well with data dimension and sparsity – applicable to *big data* problems
- Theoretical performance bounds are *robust* to model mismatch, missing data, and changing dynamics





Thank you.


NSF


SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951


AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE