Online Optimization in Dynamic Environments

Eric C. Hall and Rebecca M. Willett

July 7, 2014

Abstract

High-velocity streams of high-dimensional data pose significant “big data” analysis challenges across a range of applications and settings. Online learning and online convex programming play a significant role in the rapid recovery of important or anomalous information from these large data streams. While recent advances in online learning have led to novel and rapidly converging algorithms, these methods are unable to adapt to nonstationary environments arising in real-world problems. This paper describes a dynamic mirror descent framework which addresses this challenge, yielding low theoretical regrets bounds and accurate, adaptive, and computationally efficient algorithms which are applicable to broad classes of problems. The methods are capable of learning and adapting to an underlying and possibly time-varying dynamical model. Empirical results in the context of dynamic texture analysis, sequential compressed sensing of a dynamic scene, and tracking self-exciting point processes support the core theoretical findings.

1 Introduction

Modern sensors are collecting very high-dimensional data at unprecedented rates, often from platforms with limited processing power. These large datasets allow scientists and analysts to consider richer physical models with larger numbers of variables, and thereby have the potential to provide new insights into the underlying complex phenomena. For example, the Large Hadron Collider (LHC) at CERN “generates so much data that scientists must discard the overwhelming majority of it – hoping hard they’ve not thrown away anything useful.” [1] Typical NASA missions collect hundreds of terabytes of data every hour [2]: the Solar Data Observatory generates 1.5 terabytes of data daily [3], and the upcoming Square Kilometer Array (SKA, [4]) is projected to generate an exabyte of data daily, “more than twice the information sent around the internet on a daily basis and 100 times more information than the LHC produces” [5]. In these and a variety of other science and engineering settings, there is a pressing need to recover relevant or anomalous information accurately and efficiently from a high-dimensional, high-velocity data stream.
Rigorous analysis of such data poses major issues, however. First, we are faced with the notorious “curse of dimensionality”, which states that the number of observations required for accurate inference in a stationary environment grows exponentially with the dimensionality of each observation. This requirement is often unsatisfied even in so-called “big data” settings, as the underlying environment varies over time in many applications. Furthermore, any viable method for processing massive data must be able to scale well to high data dimensions with limited memory and computational resources. Finally, in a variety of large-scale streaming data problems, ranging from motion imagery formation to network analysis, the underlying environment is dynamic yet predictable, but many general-purpose and computationally efficient methods for processing streaming data lack a principled mechanism for incorporating dynamical models. Thus a fundamental mathematical and statistical challenge is accurate and efficient tracking of dynamic environments with high-dimensional streaming data.

Classical stochastic gradient descent methods, including the least mean squares (LMS) or recursive least squares (RLS) algorithms do not have a natural mechanism for incorporating dynamics. Classical stochastic filtering methods such as Kalman or particle filters or Bayesian updates [6] readily exploit dynamical models for effective prediction and tracking performance. However, these methods are also limited in their applicability because (a) they typically assume an accurate, fully known dynamical model and (b) they rely on strong assumptions regarding a generative model of the observations. Some techniques have been proposed to learn the dynamics [7,8], but the underlying model still places heavy restrictions on the nature of the data. Performance analysis of these methods usually does not address the impact of “model mismatch”, where the generative models are incorrectly specified.

A contrasting class of prediction methods, receiving widespread recent attention within the machine learning community, is based on an “individual sequence” or “universal prediction” [9] perspective; these strive to perform provably well on any individual observation sequence without assuming a generative model of the data.

Online convex programming provides a variety of tools for sequential universal prediction [10–13]. Here, a Forecaster measures its predictive performance according to a convex loss function, and with each new observation it computes the negative gradient of the loss and shifts its prediction in that direction. Stochastic gradient descent methods stem from similar principles and have been studied for decades, but recent technical breakthroughs allow these approaches to be understood without strong stochastic assumptions on the data, even in adversarial settings, leading to more efficient and rapidly converging algorithms in many settings.

This paper describes a novel framework for prediction in the individual sequence setting which incorporates dynamical models – effectively a novel combination of state updating from stochastic filter theory and online convex optimization from universal prediction. We establish tracking regret bounds for our proposed algorithm, Dynamic Mirror Descent (DMD), which characterize
how well we perform relative to some alternative approach (e.g., a computationally intractable batch algorithm) operating on the same data to generate its own predictions, called a “comparator sequence.” Our novel regret bounds scale with the deviation of this comparator sequence from a dynamical model. These bounds simplify to previously shown bounds when there are no dynamics. In addition, we describe methods based on DMD for adapting to the best dynamical model from either a finite or parametric class of candidate models. In these settings, we establish tracking regret bounds which scale with the deviation of a comparator sequence from the best sequence of dynamical models.

While our methods and theory apply in a broad range of settings, we are particularly interested in the setting where the dimensionality of the parameter to be estimated is very high. In this regime, the incorporation of both dynamical models and sparsity regularization plays a key role. With this in mind, we focus on a class of methods which incorporate regularization as well as dynamical modeling. The role of regularization, particularly sparsity regularization, is increasingly well understood in batch settings and has resulted in significant gains in ill-posed and data-starved settings [14–17]. More recent work has examined the role of sparsity in online methods such as recursive least squares (RLS) algorithms, but do not account for dynamic environments [18].

1.1 Organization of paper and main contributions

The remainder of this paper is structured as follows. In Section 2, we formulate the problem and introduce notation used throughout the paper, and Section 3 provides some background definitions for online convex optimization. Our Dynamic Mirror Descent (DMD) method, along with tracking regret bounds are presented in Section 4; this section also describes the application of data-dependent dynamical models and their connection to recent work on online learning with predictable sequences. DMD uses only a single series of dynamical models, but we can use it to choose among a family of candidate dynamical models. This is described for finite families in Section 5 using a fixed share algorithm, and for parametric families in Section 6. Section 7 shows experimental results of our methods in a variety of contexts ranging from imaging to self-exciting point processes. Finally, Section 8 makes concluding remarks while proofs are relegated to Section A.

2 Problem formulation

The problem of sequential prediction is posed as an iterative game between a Forecaster and the Environment. At every time point, $t$, the Forecaster generates a prediction $\hat{\theta}_t$ from a closed, convex set $\Theta \subset \mathbb{R}^d$. After the Forecaster makes a prediction, the Environment reveals the loss function $\ell_t(\cdot)$ where $\ell_t$ is a convex function which maps the space $\Theta$ to the real number line. We will assume that the loss function is the composition of a convex function $f_t : \Theta \to \mathbb{R}$ from the Environment and a convex regularization function $r : \Theta \to \mathbb{R}$ which does
not change over time. Frequently the loss function, \( f_t \) will measure the accuracy of a prediction compared to some new data point \( x_t \in X \) where \( X \) is the domain of possible observations. The regularization function promotes low-dimensional structure (such as sparsity) within the predictions. We additionally assume that we can compute a subgradient of \( \ell_t \) or \( f_t \) at any point \( \theta \in \Theta \), which we denote \( \nabla \ell_t(\hat{\theta}_t) \) and \( \nabla f_t(\hat{\theta}_t) \). Thus the Forecaster incurs the loss \( \ell_t(\hat{\theta}_t) = f_t(\hat{\theta}_t) + r(\hat{\theta}_t) \).

The goal of the Forecaster is to create a sequence of predictions \( \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_T \) that has a low cumulative loss \( \sum_{t=1}^{T} \ell_t(\hat{\theta}_t) \). Because the loss functions are being revealed sequentially, the prediction at each time can only be a function of all previously revealed losses to ensure causality. Thus, the task facing the Forecaster is to create a new prediction, \( \hat{\theta}_{t+1} \), based on the previous prediction and the new loss function \( \ell_t(\cdot) \), with the goal of minimizing loss at the next time step.

We characterize the efficacy of \( \hat{\theta}_T \) relative to a comparator \( \theta_T \) using a concept called regret, which measures the difference of the total accumulated loss of the Forecaster with the total accumulated loss of the comparator. We are particularly interested in comparators which are computationally intractable batch algorithms; in a sense, then, regret encapsulates how much one regrets working in an online setting as opposed to a batch setting with full knowledge of past and future observations:

**Definition 1 (Regret)** The regret of \( \hat{\theta}_T \) with respect to a comparator \( \theta_T \) is

\[
R_T(\theta_T) \doteq \sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{T} \ell_t(\theta_t).
\]

Our goal is to develop an online convex optimization algorithm with low (sublinear in \( T \)) regret relative to a broad family of comparator sequences. Previous work proposed algorithms which yielded regret of \( O(\sqrt{T}) \) for the relatively small family of static comparators, where \( \theta_t = \theta \) for some \( \theta \in \Theta \) and all \( t \). In contrast, our main result is an algorithm which incorporates a dynamical model, denoted \( \Phi_t : \Theta \rightarrow \Theta \), and admits a tracking regret bound of the form

\[
O(\sqrt{T}[1 + \sum_{t=1}^{T-1} \|\theta_{t+1} - \Phi_t(\theta_t)\|]).
\]

## 3 Online convex optimization preliminaries

One common approach to forming the predictions \( \hat{\theta}_t \), Mirror Descent (MD) [10, 11], consists of solving the following optimization problem:

\[
\hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \eta_t (\nabla \ell_t(\hat{\theta}_t), \theta) + D(\theta(\hat{\theta}_t)),
\]

where \( \nabla \ell_t(\theta) \) denotes an arbitrary subgradient of \( \ell_t \) at \( \theta \), \( D(\theta(\hat{\theta}_t)) \) is the Bregman divergence [19, 20] between \( \theta \) and \( \hat{\theta}_t \), and \( \eta_t > 0 \) is a step size parameter. Let \( \psi : \Theta \rightarrow \mathbb{R} \) denote a continuously differentiable function that is \( \sigma \)-strongly
convex for some parameter $\sigma > 0$ and some norm $\| \cdot \|$:  
\[
\psi(\theta_1) \geq \psi(\theta_2) + \langle \nabla \psi(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\sigma}{2} \| \theta_1 - \theta_2 \|^2
\]  
(2)

The Bregman divergence associated with $\psi$ is  
\[
D(\theta_1 \parallel \theta_2) \triangleq \psi(\theta_1) - \psi(\theta_2) - \langle \nabla \psi(\theta_2), \theta_1 - \theta_2 \rangle \geq \frac{\sigma}{2} \| \theta_1 - \theta_2 \|^2
\]  
(3)

An important consequence of this definition is the following generalization of the law of cosines: for all $\theta_1, \theta_2, \theta_3 \in \Theta$
\[
D(\theta_1 \parallel \theta_2) = D(\theta_3 \parallel \theta_2) + D(\theta_1 \parallel \theta_3) + \langle \nabla \psi(\theta_2) - \nabla \psi(\theta_3), \theta_3 - \theta_1 \rangle.
\]  
(4)

The MD approach is a generalization of online learning algorithms such as online gradient descent [12] and weighted majority [21]. Several recently proposed methods consider the data-fit term separately from the regularization term [22–24]. For instance, consider Composite Objective Mirror Descent (COMD) [22], where:
\[
\hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \eta_t \langle \nabla f_t(\hat{\theta}_t), \theta \rangle + \eta_t r(\theta) + D(\theta \parallel \hat{\theta}_t).
\]  
(5)

This formulation is helpful when the regularization function $r(\theta)$ promotes sparsity in $\theta$, and helps ensure that the individual $\hat{\theta}_t$ are indeed sparse, rather than approximately sparse as are the solutions to the MD formulation.

3.1 Static regret

In much of the online learning literature, the comparator sequence is constrained to be static or time-invariant. In this paper we refer to the regret with respect to a static comparator as static regret:

**Definition 2 (Static regret)** The static regret of $\tilde{\theta}_T$ is  
\[
R_T(\tilde{\theta}_T) \triangleq \sum_{t=1}^T \ell_t(\hat{\theta}_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta).
\]

Static regret bounds are useful in characterizing how well an online algorithm performs relative to, say, a loss-minimizing batch algorithm with access to all the data simultaneously. More generally, static regret bounds compare the performance of the algorithm against a static point which can be chosen with full knowledge of the data.

3.2 Tracking regret

Static regret fails to illuminate the performance of online algorithms in dynamic settings where the underlying parameters may be changing in time. Performance relative to a temporally-varying or dynamic comparator sequence has
been studied previously in the literature in the context of tracking regret (also
known as shifting regret) [25, 26], and the closely-related concept of adaptive
regret [21, 27].

In particular, tracking regret compares the output of the online algorithm
to a sequence of points $\theta_1, \theta_2, \ldots, \theta_T$ which can be chosen collectively with full
knowledge of the data. This is a fair comparison for a batch algorithm that
detects and fits to drift in the data, instead of fitting a single point. Frequently,
in order to bound tracking regret there needs to be a measure of the complexity
of the sequence $\theta_1, \theta_2, \ldots, \theta_T$, characterized via a measure of the temporal
variability of the sequence, such as

$$V(\theta_T) \triangleq \sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|.$$  

If this complexity is allowed to be very high, we could imagine that the com-
parator series would fit the datastream and associated series of losses closely
and hence generalize poorly. Conversely, if this complexity is restricted to be 0,
the tracking regret is equivalent to static regret. Generally, sublinear tracking
regret is only possible when the comparator sequence $\theta_T$ is piecewise constant
(where $\theta_{t+1} - \theta_t = 0$ for all but a few $t$) or varying quite slowly over time – that
is, for a small family of comparators.

4 Dynamical models in online convex programming

In contrast to previous tracking regret bounds, we develop methods and tracking
regret bounds which scale with $\sum_{t=1}^{T-1} \|\theta_{t+1} - \Phi_t(\theta_t)\|$, where $\{\Phi_t\}, t = 1, 2, \ldots$ is
a sequence of dynamical models, yielding small regret bounds for much broader
classes of dynamic comparator sequences. Specifically, we propose the alter-
native to (1) and (5) in Algorithm 1, which we call Dynamic Mirror Descent
(DMD).

By including $\Phi_t$ in the process, we effectively search for a predictor which (a)
attempts to minimize the loss and (b) which adheres to the dynamical model $\Phi_t$.
This is similar to a stochastic filter which alternates between using a dynamical
model to update the “state”, and then uses this state to perform the filtering
action. A key distinction of our approach and analysis, however, is that we make
no assumptions about $\Phi_t$’s relationship to the observed data. Our approach
effectively includes dynamics into the COMID approach.1 Indeed, for a case

1Rather than considering COMID, we might have used other online optimization algo-
rithms, such as the Regularized Dual Averaging (RDA) method [23], which has been shown
to achieve similar performance with more regularized solutions. However, to the best of our
knowledge, no tracking or shifting regret bounds have been derived for dual averaging meth-
ods (regularized or otherwise). Recent results on the equivalence of COMID and RDA [28]
suggest that the bounds derived here might also hold for a variant of RDA, but proving this
remains an open problem.
Algorithm 1 Dynamic mirror descent (DMD) with known dynamics

Given decreasing sequence of step sizes $\eta_t > 0$
Initialize $\tilde{\theta}_1 \in \Theta$
for $t = 1, \ldots, T$ do
  Observe $x_t$ and incur loss $\ell_t(\tilde{\theta}_t)$
  Receive dynamical model $\Phi_t$
  Set
  \[
  \tilde{\theta}_{t+1} = \arg\min_{\theta \in \Theta} \eta_t(\nabla f_t(\tilde{\theta}_t), \theta) + \eta_t r(\theta) + D(\|\tilde{\theta}_t\|)
  \] (6a)
  \[
  \tilde{\theta}_{t+1} = \Phi_t(\tilde{\theta}_{t+1})
  \] (6b)
end for

with no dynamics, so that $\Phi_t(\theta) \equiv \theta$ for all $\theta$ and $t$, our method is equivalent to COMID.

Our main result uses the following assumptions:

- For all $t = 1, \ldots, T$ the functions $\ell_t$ and $\psi$ are Lipschitz with constants $G$ and $M$ respectively, such that $\|\nabla \ell_t(\theta)\| \leq G$ and $\|\psi(\theta)\| \leq M$ for all $\theta \in \Theta$. The function $\| \cdot \|_*$ used in these assumptions is the dual to the norm in (2).
- There exists a constant $D_{\text{max}}$ such that $D(\theta_1 \| \theta_2) \leq D_{\text{max}}$ for all $\theta_1, \theta_2 \in \Theta$.
- For all $t = 1, \ldots, T$, the transformation $\Phi_t$ has a maximum distortion factor $\Delta_{\Phi}$ such that $D(\Phi_t(\theta_1) \| \Phi_t(\theta_2)) - D(\theta_1 \| \theta_2) \leq \Delta_{\Phi}$ for all $\theta_1, \theta_2 \in \Theta$. When $\Delta_{\Phi} \leq 0$ for all $t$, we say that $\Phi_t$ satisfies the contractive property.

4.1 Tracking regret bound

Theorem 3 (Tracking regret of dynamic mirror descent) Let $\Phi_t$ be a dynamical model such that $\Delta_{\Phi} \leq 0$ for $t = 1, 2, \ldots, T$ with respect to the Bregman used in 6. Let the sequence $\theta_T$ be generated using Alg. 1 using a non-increasing series $\eta_{t+1} \leq \eta_t$, with a convex, Lipschitz function $\ell_t$ on a closed, convex set $\Theta$, and let $\theta_T$ be an arbitrary sequence in $\Theta^T$. Then

$$R_T(\theta_T) \leq \frac{D_{\text{max}}}{\eta_{T+1}} + \frac{2M}{\eta_T} V_\Phi(\theta_T) + \frac{C^2}{2\sigma} \sum_{t=1}^T \eta_t$$

with $V_\Phi(\theta_T) \triangleq \sum_{t=1}^{T-1} \|\theta_{t+1} - \Phi_t(\theta_t)\|$.

where $V_\Phi(\theta_T)$ measures variations or deviations of the comparator sequence $\theta_T$ from the sequence of dynamical models $\Phi_1, \Phi_2, \ldots, \Phi_T$. If $\eta_t \propto \frac{1}{\sqrt{t}}$, then for
some $C > 0$ independent of $T$,

$$R(\theta_T) \leq C \sqrt{T} (1 + V_\Phi(\theta_T))$$

This bound scales with the comparator sequence’s deviation from the sequence of dynamical models $\{\Phi_t\}_{t>0}$—a stark contrast to previous tracking regret bounds which are only sublinear for comparators which change slowly with time or at a small number of distinct time instances. Note that when $\Phi_t$ corresponds to an identity operator, the bound in Theorem 3 corresponds to existing tracking or shifting regret bounds [13, 26].

It is intuitively satisfying that this measure of variation, $V_\Phi(\theta_T)$, appears in the tracking regret bound. First, if the comparator actually follows the dynamics, this variation term will be very small, leading to low tracking regret. This fact holds whether $\Phi_t$ is part of the generative model for the observations or not. Second, we can get a dynamic analog of static regret, where we enforce $V_\Phi(\theta_T) = 0$. This is equivalent to saying that the batch comparator is fitting the best single trajectory using $\Phi_t$ instead of the best single point. Using this, we would recover a bound analogous to a static regret bound in a stationary setting.

The condition that $\Delta_\Phi \leq 0$ is similar to requiring that $\Phi_t$ be a contractive mapping. This restriction is important; without it, any poor prediction made at one time step could be exacerbated by repeated application of the dynamics. For instance, linear dynamic models with all eigenvalues less than or equal to unity satisfy this condition with respect to the squared $\ell_2$ Bregman Divergence, similar in spirit to restrictions made in more classical adaptive filtering work such as [29].

Notice also that if $\Phi_t(\theta) = \theta$ in for all $t$, then Theorem 3 gives a novel, previously unknown tracking regret bound for COMID.

4.2 Data-dependent dynamics

An interesting example of dynamical models is the class of data-dependent dynamical models. In this regime the state of the system at a given time is not only a function of the previous state, but also the actual observations. One key example of this scenario arises in self-exciting point processes, where the state of the system is directly related to the previous, stochastic observations. Our algorithm can account for such models since the function $\Phi_t(\theta)$ is time varying, and therefore can implicitly depend on all data up to time $t$, i.e. $\Phi_t(\theta) = \Phi_t(\theta, x_1, x_2, \ldots, x_t)$. Our regret bounds therefore scale with how well the comparator series matches these data dependent dynamics:

$$R(\theta_T) \leq C \left( \sqrt{T} \left[ 1 + \sum_{t=1}^{T-1} \| \theta_{t+1} - \Phi_t(\theta, x_1, \ldots, x_t) \| \right] \right).$$

Notice now that the data plays a part in the regret bounds, whereas before we only measured the variation of the comparator. Data-dependent regret bounds
are not new. Concurrent related work considers online algorithms where the
data sequence is described by a “predictable process” [30]. The basic idea of
that paper is that if one has a sequence functions \( M_t \) which predict \( x_t \) based
on \( x_1, x_2, \ldots, x_{t-1} \), then the output of a standard online optimization routine
should be combined with the predictor generated by \( M_t \) to yield tighter regret
bounds that scale with \( (\sum_t \|x_t - M_t(x_1, \ldots, x_{t-1})\|^2)^{1/2} \). However, [30] only
works with static regret (i.e., regret with respect to a static comparator) and
their regret has a variation term that expresses the deviation of the input data
from the underlying process. In contrast, our tracking regret bounds scale with
the deviation of a comparator sequence from a prediction model.

5 Prediction with a finite family of dynamical
models

DMD in the previous section uses a single sequence of dynamical models. In
practice, however, we may not know the best dynamical model to use, or
the best model may change over time in nonstationary environments. To
address this challenge, we assume a finite set of candidate dynamical models
\( \{\Phi_{1,t}, \Phi_{2,t}, \ldots, \Phi_{N,t}\} \) at every time \( t \), and describe a procedure which uses this
collection to adapt to nonstationarities in the environment. In particular we
establish tracking regret bounds which scale not with the deviation of a com-
parator from a single dynamical model, but with how it deviates from a series of
different dynamical models on different time intervals with at most \( m \) switches.
These switches define \( m+1 \) different time segments \( [t_i, t_{i+1}-1] \) with time points
\( 1 = t_1 < \cdots < t_{m+2} = T \). We can bound the regret associated with the best
dynamical model on each time segment and then bound the overall regret using
a Prediction with Experts Advice algorithm.

Our dynamic fixed share (DFS) estimate is presented in Algorithm 2. Let
\( \hat{\theta}_{i,t} \) denote the output of Alg. 1 using dynamical models \( \Phi_{i,1}, \Phi_{i,2}, \ldots, \Phi_{i,t} \); we
choose \( \hat{\theta}_t \) by using the Fixed Share forecaster on these outputs.\(^2\) In Fixed Share,
each expert (here, each sequence of candidate dynamical models) is assigned a
weight that is inversely proportional to its cumulative loss at that point yet with
some weight shared amongst all the experts, so that an expert with very small
weight can quickly regain weight to become the leader [26,31].

In this update, \( \lambda \in (0, 1) \) is a parameter which controls how much of the

\(^2\)There are many algorithms from the Prediction with Expert Advice literature which can
be used to form a single prediction from the predictions created by the set of dynamical mod-
els. We use the Fixed Share algorithm [31] as a means to combine estimates with different
dynamics; however, other methods could be used with various tradeoffs. One of the primary
drawbacks of the Fixed Share algorithm is that an upper bound on the number of switches \( m \)
must be known a priori. However, this method has a simple implementation and tracking
regret bounds. One common alternative to Fixed Share allows the switching parameter (\( \lambda \) in
Alg. 2) to decrease to zero as the algorithm runs [32,33]. This has the benefit of not requiring
knowledge about the number of switches, but comes at the price of higher regret. Alterna-
tive expert advice algorithms exist which decrease the regret but increase the computational
complexity. For a thorough treatment of existing methods see [34].
Algorithm 2 Dynamic fixed share (DFS)

Given decreasing sequence of step sizes $\eta_t > 0$ and $\eta_r > 0$
Initialize $\hat{\theta}_1 \in \Theta$, $\hat{\theta}_{1,1} \in \Theta$ and $w_{i,1} = \frac{1}{N}$ for $i = 1, \ldots, N$, $\lambda \in (0, 1)$, and $\eta_t, \eta_r > 0$.
for $t = 1, \ldots, T$ do
  Observe $x_t$ and incur loss $\ell_t(\hat{\theta}_t)$
  Receive dynamical model $\Phi_{i,t}$ for $i = 1, \ldots, N$
  for $i = 1, \ldots, N$ do
    Set $\tilde{w}_{i,t+1} = \frac{w_{i,t} \exp \left( -\eta_t \ell_t(\tilde{\theta}_{i,t}) \right)}{\sum_{j=1}^{N} w_{j,t} \exp \left( -\eta_t \ell_t(\tilde{\theta}_{j,t}) \right)}$
    $w_{i,t+1} = \frac{\lambda}{N} + (1 - \lambda) \tilde{w}_{i,t}$
    $\tilde{\theta}_{i,t+1} = \arg \min_{\theta \in \Theta} \eta_t \langle \nabla f_t(\tilde{\theta}_t), \theta \rangle + \eta_t r(\theta) + D(\theta \| \tilde{\theta}_t)$
    $\hat{\theta}_{t+1} = \Phi_{i,t}(\tilde{\theta}_{i,t+1})$
  end for
Set $\hat{\theta}_{t+1} = \sum_{i=1}^{N} w_{i,t+1} \hat{\theta}_{i,t+1}$
end for

weight is shared amongst the experts. By sharing some weight, it allows experts with high loss, and therefore low weight, to quickly regain weight if they start performing well. This is the mechanism that allows fast switching between experts.

Theorem 4 (Tracking regret of DFS algorithm) Assume all the candidate dynamic sequences are contractive such that $\Delta_{\Phi} \leq 0$ for $\Phi_{i,t}$ for all $t = 1, \ldots, T$ and $i = 1, \ldots, N$ with respect to the Bregman Divergence in alg 1. Then for some $C > 0$, the dynamic fixed share algorithm in Algorithm 2 with parameter $\lambda$ set equal to $\frac{m}{T}$, $\eta_r = \sqrt{\frac{8((m+1) \log(N) + m \log(T+1))}{T}}$ and $\eta_t = \frac{1}{\sqrt{t}}$ with a convex, Lipschitz function $\ell_t$ on a closed, convex set $\Theta$, has tracking regret

$$R_T(\theta_T) \leq C \left( \sqrt{T} \left( \sqrt{(m+1) \log N + m \log T} + V^{(m+1)}(\theta_T) \right) \right),$$

where

$$V^{(m+1)}(\theta_T) \triangleq \min_{t_2, \ldots, t_m+1} \sum_{k=1}^{m+1} \min_{i_k \in \{1, \ldots, N\}} \sum_{t=t_k}^{t_{k+1}-1} \| \theta_{t+1} - \Phi_{i_k,t}(\theta_t) \|$$

10
measures the deviation of the sequence \( \theta_T \) from the best sequence of dynamical models with at most \( m \) switches (where \( m \) does not depend on \( T \)).

Note that the family of comparator sequences \( \theta_T \) for which \( R_T(\theta_T) \) scales sublinearly in \( T \) is significantly larger than the set of comparators yielding sublinear regret for MD.

If \( T \) is not known in advance the doubling trick \[13\] can be used, where temporary time horizons are set of increasing length. Note that \( V^{(m+1)}(\theta_T) \leq V_{\Phi_i}((\theta_T)) \) for any fixed \( i \in \{1, \ldots, N\} \), thus this approach yields a lower variation term than using a fixed dynamical model. However, we incur some loss by not knowing the optimal number of switches \( m \) or when the optimal switching times are.

6 Parametric dynamical models

Rather than having a finite family of dynamical models, as we did in Section 5, we may consider a parametric family of dynamical models, where the parameter \( \alpha \in \mathbb{R}^n \) of \( \Phi_t \) is allowed to vary across a closed, convex domain, denoted \( A \). In other words, we consider \( \Phi_t : \Theta \times A \mapsto \Theta \). In this context we would like to jointly predict both \( \alpha \) and \( \theta \). One might consider defining \( \zeta \equiv (\theta; \alpha) \) as the concatenation of \( \theta \) and \( \alpha \), and then generating a sequence of \( \hat{\zeta} \)'s using COMID (5) or DMD (6). However, the COMID regret would not capture deviations of a comparator sequence of predictions from a series of dynamical models as desired. To use DMD, we would need to define a dynamical model \( \Psi : \Theta \times A \mapsto \Theta \times A \), so that \( \Psi(\theta, \alpha) = (\Phi(\theta, \alpha), \alpha) \), and use this in place of \( \Phi_t \) in (6). However, \( \Psi \) is not contractive for most \( \Phi \) of interest, so the DMD regret bounds would not hold.

To address these challenges, we consider two approaches. First, in Section 6.1 we consider tracking only a finite subset of the possible model parameters, in a manner similar to when we had a finite collection of possible dynamical models, which provide a “covering” of the parameter space. In this case, the overall regret and computational complexity both depend on the resolution of the covering set. Second, in Section 6.2, we consider a special family of additive dynamical models; in this setting, we can efficiently learn the optimal dynamics.

6.1 Covering the set of dynamical models

In this section we show that by tracking a subset which appropriately covers the entire space of candidate models, we can bound the overall regret, as well as bound the number of parameter values we have to track, and the inherent tradeoff between the two. We propose to choose a finite collection of parameters from a closed, convex set \( A \) and perform DFS (Alg. 2) on this collection. We specifically consider the case where the true dynamical model \( \alpha^* \in A \) is unchanging in time and use DFS with \( m = 0 \). (Fixed share with \( m = 0 \) amounts to the Exponentially Weighted Averaging Forecaster \[13, 21, 35\].) In the below,
for any $\alpha \in \mathcal{A}$, let
\[ V_\Phi(\theta_T, \alpha) \triangleq \sum_{t=1}^{T-1} \| \theta_{t+1} - \Phi_t(\theta_t, \alpha) \|. \]

**Theorem 5 (Covering sets of dynamics parameter space)** Let $\varepsilon_N > 0$ and $\mathcal{A}_N$ denote a covering set for $\mathcal{A}$ with cardinality $N$, such that for every $\alpha \in \mathcal{A}$, there is some $\alpha' \in \mathcal{A}_N$ such that $\| \alpha - \alpha' \| \leq \varepsilon_N$. Define candidate dynamical models as $\Phi_t(\cdot, \alpha)$ for $\alpha \in \mathcal{A}_N$ and assume they are all contractive with respect to the Bregman Divergence used in Alg. 1. If $\| \Phi_t(\theta, \alpha) - \Phi_t(\theta, \beta) \| \leq L \| \alpha - \beta \|$ for some $L > 0$ for all $\alpha, \beta \in \mathcal{A}$, then for some constant $C > 0$, the DFS algorithm with $\eta_t = \frac{1}{\sqrt{t}}, \eta_r = \sqrt{\frac{2 \log(N)}{t}}, \lambda = 0$ yields a tracking regret bounded by
\[ C \left( \sqrt{T \left[ \sqrt{\log(N)} + \min_{\alpha \in \mathcal{A}} V_\Phi(\theta_T, \alpha) + T \varepsilon_N \right]} \right). \]

Intuitively, we know that if we set $\varepsilon_N$ to be very small we will have good performance because any possible parameter value $\alpha \in \mathcal{A}$ would have to be close to a candidate dynamic; however, we would need to choose many candidates. Conversely, if we run DFS on only a few candidate models, it will be computationally much more efficient but our total regret will grow due to parameter mismatch.

**Corollary 6** Assume $\mathcal{A} \subseteq [A_{\text{min}}, A_{\text{max}}]^n$, and let $\gamma > 0$ be given. Let $k = \lceil (A_{\text{max}} - A_{\text{min}}) n T^\gamma / 2 \rceil$ and $\partial = (A_{\text{max}} - A_{\text{min}}) / (2k)$; let $\mathcal{A}_N = \{ A_{\text{min}} + \partial, A_{\text{min}} + 3\partial, \ldots, A_{\text{min}} (2k-1) \partial \}^n$ correspond to an $n$-dimensional grid with $k^n$ grid points over $\mathcal{A}$. Then
\[ \max_{\alpha \in \mathcal{A}} \min_{\alpha' \in \mathcal{A}_N} \| \alpha - \alpha' \|_1 \leq T^{-\gamma}. \]

Additionally, the total number of grid points is upper bounded by
\[ N \leq \left( \frac{(A_{\text{max}} - A_{\text{min}}) n T^\gamma}{2} + 1 \right)^n = O(T^{n \gamma}). \]

Under the assumptions of Theorem 5, with this set $\mathcal{A}_N$ and using the fact that norms are equivalent on finite-dimensional vectors (i.e., there’s a finite $Z > 0$ such that $\| \alpha - \beta \|_1 \leq Z \| \alpha - \beta \|$ for any $\alpha, \beta \in \mathcal{A}$ for any norm), we get the following bound on regret for some constant $C > 0$.
\[ R_T(\theta_T) \leq C \left( \sqrt{T \left[ \sqrt{n \log(T)} + \min_{\alpha \in \mathcal{A}} V_\Phi(\theta_T, \alpha) \right]} + T^{1-\gamma n} \right). \]

Here we have an explicit tradeoff between regret and computational accuracy controlled by $\gamma$, since the computational computational complexity is linear in $N = O(T^{n \gamma})$.

We can further control the tradeoff between computation complexity and performance by allowing $\varepsilon_N$ to vary in time. This could be done by using the doubling trick, setting temporary time horizons, and then refining the grid once the temporary time horizon is reached using a slightly different experts algorithm which could account for the changing number of experts as in [36].
6.2 Additive dynamics in exponential families

The approach described above for generating a covering set of dynamical models may be effective when the dimension of parameters is small; however, in higher dimensions, this approach can require significant computational resources. In this section, we consider an alternative approach that only requires the computation of predictions for a single dynamical model. We will see that in some settings, the prediction produced by Dynamic Mirror Descent (DMD) and a certain set of parameters for the dynamic model can quickly be converted to the prediction for a different set of parameters. While the method described in this section is efficient and admits strong regrets bounds, it is applicable only for loss functions derived from exponential families.

The basics of exponential families are described in [37,38], and mirror descent in this setting is explored in [39, 40]. We assume some $\phi : X \rightarrow \mathbb{R}^d$ which is a measurable function of the data, and let $\phi_k$, $k = 1, 2, \ldots, d$, denote its components:

$$\phi(x) = (\phi_1(x), \ldots, \phi_d(x))^T.$$  

We use the specific loss function

$$\ell_t(\theta) = -\log p_\theta(x_t) \quad (7a)$$  

where

$$p_\theta(x) \triangleq \exp\{\langle \theta, \phi(x) \rangle - Z(\theta)\} \quad (7b)$$

for a sufficient statistic $\phi$ and $Z(\theta) \triangleq \log \int \exp\{\langle \theta, \phi(x) \rangle\} dx$, known as the log-partition function, ensures that $p_\theta(x)$ integrates to a constant independent of $\theta$.

Furthermore, as in [39,40], we use the Bregman divergence corresponding to the Kullback-Liebler divergence between two members of the exponential family:

$$D(\theta_1 \parallel \theta_2) = Z(\theta_1) - Z(\theta_2) - \langle \nabla Z(\theta_2), \theta_1 - \theta_2 \rangle.$$  

In our analysis we will be using the Legendre–Fenchel dual of $Z$ [41,42]:

$$Z^*(\mu) \triangleq \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - Z(\theta)\}.$$  

Let $\Theta^*$ denote the image of Int $\Theta$ under the gradient mapping $\nabla Z$ i.e. $\Theta^* = \nabla Z(\text{Int } \Theta)$. An important fact is that the gradient mappings $\nabla Z$ and $\nabla Z^*$ are inverses of one another [11,13,43]:

$$\begin{align*}
\nabla Z^*(\nabla Z(\theta)) &= \theta \\
\nabla Z(\nabla Z^*(\mu)) &= \mu 
\end{align*} \quad \forall \theta \in \text{Int } \Theta, \mu \in \text{Int } \Theta^*$$

Following [13], we may refer to the points in Int $\Theta$ as the primal points and to their images under $\nabla Z$ as the dual points. For simplicity of notation, in the sequel we will write $\mu = \nabla Z(\theta)$, $\theta = \nabla Z^*(\mu)$, $\bar{\mu}_t = \nabla Z(\bar{\theta}_t)$, etc.

Additionally, we will use a dynamical model that takes on a specific form:

$$\Phi_t(\theta, \alpha) = \nabla Z^*(A_t \nabla Z(\theta) + B_t \alpha + c_t) \quad (8)$$
for \( \theta \in \text{Int } \Theta, c_t \in \mathbb{R}^d, \alpha \in \mathbb{R}^n, A_t \in \mathbb{R}^{d \times d}, \text{ and } B_t \in \mathbb{R}^{d \times n}. \) \( A_t, B_t \) and \( c_t \) are considered known. Using these dynamics, we let \( \hat{\theta}_{\alpha,t} \) denote the output of DMD (Alg. 1) at time \( t \) and \( \hat{\mu}_{\alpha,t} \) be its dual. Under all these conditions, we have the following Lemma.

**Lemma 7** For any \( \alpha, \beta \in A \), let \( \hat{\mu}_{\alpha,1} = \hat{\mu}_{\beta,1} \) be the duals of the initial prediction for DMD and \( K_1 = 0 \in \mathbb{R}^{d \times n} \). Additionally assume that the minimizer of equation 6a is a point in \( \text{Int } \Theta \) for any parameter \( \alpha \in A \). Then the DMD prediction under a dynamical model parameterized by \( \alpha \) can be calculated directly from the DMD prediction under a dynamical model parameterized by \( \beta \) for \( t > 0 \) as

\[
\hat{\mu}_{\alpha,t} = \hat{\mu}_{\beta,t} + K_t (\alpha - \beta)
\]

where

\[
K_t = (1 - \eta_t)(A_{t-1}K_{t-1} + B_{t-1}).
\]

From Lemma 7, we see that the prediction for dynamical model \( \alpha \) can be computed simply from the prediction using parameters \( \beta \) and the value \( K_t \). This is a significant computational gain compared to DFS, where we had to keep track of predictions for each candidate dynamical model individually and therefore needed to bound the number of experts for tractability.

Algorithm 3 leverages Lemma 7 to simultaneously track both \( \hat{\theta}_t \) and the best dynamical model parameter \( \alpha \). In this algorithm, \( \ell_t \) is the function defined as

\[
\tilde{\ell}_t(\mu) \triangleq \ell_t(\nabla Z^*(\mu)) \equiv \ell_t(\hat{\theta}).
\]

The basic idea is the following: we use mirror descent to compute an estimate of the best dynamical model parameter, compute the DMD prediction associated with that parameter, and then use DMD to update that prediction for the next round.

**Algorithm 3** Dynamic mirror descent (DMD) with parametric additive dynamics

Given decreasing sequence of step sizes \( \rho_t, \eta_t > 0 \)

Initialize \( \hat{\alpha}_1 = 0, K_1 = 0, \hat{\theta}_1 \in \Theta, \hat{\mu}_1 = \nabla Z(\hat{\theta}_1) \)

for \( t = 1, ..., T \) do

Observe \( x_t \)

Incur loss \( \ell_t(\hat{\theta}_t) = -\langle \hat{\mu}_t, \phi(x_t) \rangle + Z(\hat{\theta}_t) \)

Set \( g_t(\alpha) = \ell_t(\hat{\mu}_{\alpha,t}) \equiv \ell_t(\hat{\theta}_{\alpha,t}) \)

Set \( \hat{\alpha}_{t+1} = \text{proj}_A(\hat{\alpha}_t - \rho_t g_t(\hat{\alpha}_t)) \)

Set \( \mu'_{t+1} = \hat{\mu}_t + K_t(\hat{\alpha}_{t+1} - \hat{\alpha}_t) \)

Set \( \hat{\mu}_{t+1} = (1 - \eta_t)\mu'_{t+1} + \eta_t \phi(x_t) \)

Set \( \hat{\theta}_{t+1} = \nabla Z^*(\hat{\mu}_{t+1}) \)

Set \( \hat{\theta}_{t+1} = \Phi_t(\hat{\theta}_{t+1}, \hat{\alpha}_{t+1}) \)

Set \( K_{t+1} = (1 - \eta_t)A_tK_t + B_t \)

end for
Theorem 8 Assume that the observation space $X$ is bounded. Let $\Theta \subset \mathbb{R}^d$ be a bounded, convex set satisfying the following properties for a given constant $H > 0$:

- For all $\theta \in \Theta$, \[ Z(\theta) \overset{\Delta}{=} \int_X \exp\left\{ \langle \theta, \phi(x) \rangle \right\} d\nu(x) < +\infty. \]
- For all $\theta \in \Theta$, $\nabla^2 Z(\theta) \succeq 2HI_{d \times d}$.
- Let $f_t$ denote the objective function in (6a). For every $x \in X$ and $t \in \{1, 2, 3, \ldots\}$, the solution to $\arg\min_{\theta \in \Theta} f_t(\theta)$ occurs where $\nabla f_t = 0$.

If the assumptions of Lemma 7 hold, and $\Phi_t(\theta, \alpha)$ is contractive for all $\alpha \in \mathcal{A}$ with respect to the Bregman Divergence induced by $Z(\theta)$, and loss function is of the form (7) and \[ \tilde{\ell}_t(\mu) \overset{\Delta}{=} \ell_t(\nabla Z^*(\mu)) \]

is convex in $\mu$, and $\eta_t, \rho_t \propto 1/\sqrt{t}$, then the tracking regret associated with Algorithm 3 for dynamical models of the form (8) is \[ R_T(\theta_T) \leq C\sqrt{T} \left( 1 + \min_{\alpha \in \mathcal{A}} V_{\Phi}(\theta_T, \alpha) \right) \]

for some constant $C > 0$.

Theorem 8 shows that Algorithm 3 allows us to simultaneous track predictions and dynamics, and we perform nearly as well as if we knew the best dynamical models for the entire sequence in hindsight. While this approach is only applicable for specific forms of the loss functions and dynamical models, those forms arise in a wide variety of practical problems.

7 Experiments and results

As mentioned in the introduction, many online learning problems can benefit from the incorporation of dynamical models. In the below, we describe how the ideas described and analyzed in this paper might be applied to anomaly detection from streaming dynamic textures, compressive video reconstruction, and analysis of neuron firing rates within networks.

7.1 DMD experiment: dynamic textures with missing data

As mentioned in the introduction, sensors such as the Solar Data Observatory are generating data at unprecedented rates. Heliophysicists have physical models of solar dynamics, and often wish to identify portions of the incoming data which are inconsistent with their models. This “data thinning” process is an
essential element of many big data analysis problems. We simulate an analogous situation in this section.

In particular, we consider a datastream corresponding to a dynamic texture, where spatio-temporal dynamics within motion imagery are modeled using an autoregressive process. In this experiment, we consider a setting where “normal” autoregressive parameters are known, and we use these within DMD to track a scene from noisy image sequences with missing elements. (Missing elements arise in our motivating solar astronomy application, for instance, when cosmic rays interfere with the imaging detector array.) As suggested by our theory, the tracking will fail and generate very large losses when the posited dynamical model is inaccurate.

More specifically, the idea of dynamic textures is that a low dimensional, autoregressive model can be used to simulate a video which replicates a moving texture such as flowing water, swirling fog, solar plasma flows, or rising smoke. This process is modeled in the following way:

\[
\begin{align*}
\theta_t &= A\theta_{t-1} + Bu_t \\
x_t &= C_0 + C\theta_t + Dv_t \\
v_t, u_t &\sim \mathcal{N}(0, I).
\end{align*}
\]

In the above, \(\theta_t\) denotes the true underlying parameters of the system, and \(x_t\) the observations. The matrix \(A\) is the autoregressive parameters of the system, which will be unique for the type of texture desired, \(C_0\) the average background intensity, \(C\) is the sensing matrix which is usually a tall matrix, and \(B\) and \(D\) encode the strength of the driving and observation noises respectively. Using the toolbox developed in [46] and samples of a 220 by 320 pixel ocean scene [45], we learned two sets of parameters \(A, A' \in \mathbb{R}^{50 \times 50}\), one representing the water flowing when the data is played forward, and the other when played backwards, as well as corresponding parameters \(C_0 \in \mathbb{R}^{70400}, C, C' \in \mathbb{R}^{70400}, B, B' \in \mathbb{R}^{50}\) and \(D, D' \in \mathbb{R}^{70400}\). Parameters \(\theta_t \in [-500, 500]^{50}\) and data \(x_t \in [-500, 500]^{70400}\) were then generated using these parameters, with the parameters \(A', B', C', D'\) and \(C_0\) on \(t = 100, ..., 120\) and \(t = 300, ..., 320\) and the parameters \(A, B, C, D, C_0\) on the rest of \(t = 1, ..., 550\) according to the above equations. Finally, every observation is corrupted by 50% missing values, chosen uniformly at random at every time point. The parameters \(A, C_0\), and \(C\) were then used to define our (imperfect) dynamical model for DMD, \(\Phi_t(\theta) = A\theta\), and a loss function \(\ell_t(\theta) = ||P_t(C\theta - C_0 - x_t)||^2_2\), where \(P_t\) is a linear operator accounting for the missing data. Note that \(B\) and \(D\) are not reflected in these choices despite playing a role in generating the data; our theoretical results hold regardless. We use \(\psi(\cdot) = \frac{1}{2} \cdot ||\cdot||^2_2\) so the Bregman Divergence \(D(x\|y) = \frac{1}{2} \cdot ||x - y||^2_2\) is the usual squared Euclidean distance, and we perform no regularization (\(r(\theta) = 0\)). We set \(\eta_t = \frac{1}{2\sqrt{t}}\), and ran 100 different trials comparing the DMD method to regular Mirror Descent (MD) to see the advantage of accounting for underlying dynamics. The results are shown in Figure 1.

There are a few important observations about this procedure. The first is that by incorporating the dynamic model, we produce an estimate which
Figure 1: Simulation results for the experiment in Section 7.1. The vertical dashed lines indicate the intervals where the posited dynamical model was not reflected by the underlying data; note the sharp increases in the losses associated with DMD over those intervals, particularly in contrast with the losses associated with MD. Standard online learning methods like MD do not facilitate the detection of subsets of data which do not fit hypothesized physical models.

visually looks like the dynamic texture of interest, instead of the Mirror Descent prediction, which looks like a single snapshot of the water. Second, we can recover a good representation of the scene with a large amount of missing data, due to the autoregressive parameters being of a much lower dimension than the data itself. Finally, because we are using the dynamics of forward moving water, when the true data starts moving backward, a change that is imperceptible visually, the loss spikes, alerting us of the abnormal behavior.

7.2 DFS experiment: compressive video reconstruction

There is increasing interest in using “big data” analysis techniques in applications like high-throughput microscopy, where scientists wish to image large collections of specimens. This work is facilitated by the development of novel microscopes, such as the recent fluorescence microscope based on structured illumination and compressed sensing principles [47]. However, measurements in such systems are acquired sequentially, posing significant challenges when imaging live specimens.

Knowledge of underlying motion in compressed sensing image sequences can allow for faster, more accurate reconstruction [48–50]. By accounting for the underlying motion in the image sequence, we can have an accurate prediction of the scene before receiving compressed measurements, and when the measurements are noisy and the number of observations is far less than the number of
pixels of the scene, these predictions allow both fast and accurate reconstructions. If the dynamics are not accounted for, and previous observations are used as prior knowledge, the reconstruction could end up creating artifacts such as motion blur or overfitting to noise. There has been significant recent interest in using models of temporal structure to improve time series estimation from compressed sensing observations [51, 52]; the associated algorithms, however, are typically batch methods poorly suited to large quantities of streaming data. In this section we demonstrate that DMD helps bridge this gap.

In this section, we simulate fluorescence microscopy data generated by the system in [47] while imaging a paramecium moving in a 2-dimensional plane; the $t^{th}$ frame is denoted $\theta_t$ (a 120×120 image stored as a length-14400 vector) which takes values between 0 and 1. The corresponding observation is $x_t = A_t \theta_t + n_t$, where $A_t$ is a 50 × 14400 matrix with each element drawn iid from $\mathcal{N}(0,1)$ and $n_t$ corresponds to measurement noise with $n_t \sim \mathcal{N}(0,\sigma^2)$ with $\sigma^2 = 0.1$. This model coincides with several compressed sensing architectures [47,53].

Our loss function uses $f_t(\theta) = \frac{1}{2\sigma^2} \|x_t - A_t \theta\|_2^2$ and $r(\theta) = \tau \|\theta\|_1$, where $\tau > 0$ is a tuning parameter. We construct a family of $N = 9$ dynamical models, where $\Phi_t(\theta)$ shifts the (unvectorized) frame, $\theta$, one pixel in a direction corresponding to an angle of $2\pi i/(N-1)$ as well as a “dynamic” corresponding to no motion. (With the zero motion model, DMD reduces to COMID.) The true video sequence uses different dynamical models over $t = \{1,\ldots,550\}$ (upward motion) and $t = \{551,\ldots,1000\}$ (motion to the right). Finally, we use $\psi(\cdot) = \frac{1}{2} \|\cdot\|_2^2$ so the Bregman Divergence $D(x\|y) = \frac{1}{2} \|x - y\|_2^2$ is the usual squared Euclidean distance. The DMD sub-algorithms use $\eta_t = \frac{1}{\sqrt{t}}$, $\tau = .002$ and the DFS forecaster uses $\lambda = \frac{m}{m-1} = \frac{1}{999}$ and $\eta_r$ is set as in Theorem 4. The experiment was then run 100 times.

Figures 2 and 3 show the impact of using DFS. We see that DFS switches between dynamical models rapidly and outperforms all of the individual predictions, including COMID, used as a baseline, to show the advantages of incorporating knowledge of the dynamics.

### 7.3 DMD with parametric additive dynamics

Finally, we look at self-exciting point processes on connected networks [54,55]. Here we assume there is an underlying rate for nodes in a network which dictate how likely each node is to participate in an action. Then, based on which nodes act, it will increase other nodes likelihood to act in a dynamic fashion. For example, in a social network a node could correspond to a person and an action could correspond to crime [56]. In a biological neural network, a node could correspond to a neuron and an action could correspond to a neural spike [57].

We simulate observations of a such a self-exciting point process in the following way:

$$\begin{align*}
\mu_{t+1} &= \Phi_t(\mu_t, W) = \tau \mu_t + W x_t + (1 - \tau) \bar{\mu} \\
x_t &\sim \text{Poisson}(\mu_t)
\end{align*}$$
Figure 2: Tracking dynamics using DFS and comparing individual models for directional motion for the experiment in Section 7.2. Only shown are DMD losses for motions which are true before or after $t = 550$ for clarity. Before $t = 550$ the upward motion dynamic model incurs small loss, where as after $t = 550$ the motion to the right does well, and DFS successfully tracks this change.

For our experiments $\mu_t \in [0, 5]^{100}$ represents the average number of actions each of 100 nodes will make during time interval $t$, and $W \in [0, 5]^{100 \times 100}$ reflects the unknown underlying network structure which encodes how much an event by a one node will increase the likelihood of an event by another node in future time intervals. Here we assume $\tau$ is a known parameter between zero and one, $\bar{\mu} \in \mathbb{R}^{100}$ is a underlying base event rate.

Our goal is to track the event rates $\mu_t$ and the network model $W$ simultaneously; Algorithm 3 is applied with

$$\ell_t(\theta) = \langle 1, \exp(\theta) \rangle - \langle x_t, \theta \rangle,$$
$$\tilde{\ell}_t(\mu) = \langle 1, \mu \rangle - \langle x_t, \log \mu \rangle,$$
$$Z(\theta) = \langle 1, \exp(\theta) \rangle,$$
$$\mu = \nabla Z(\theta) = \exp(\theta).$$

We generated data according to this model for $t = 1, \ldots, 50000$ for 1000 different trials, using $\tau = 0.5$, $\bar{\mu} = 0.1$ and $W$ generated such that it is all zeros except on each distinct $10 \times 10$ block along the diagonal, elements are chosen to be $uu^T$ for a vector $u \in [0, 1, 1.1]^{10}$ with elements chosen uniformly at random. The matrix $W$ is then normalized so that its spectral norm is 0.25 for stability. Using this generated data we ran DMD with known $W$ (Alg. 1), MD, and DMD with additive dynamics (Alg. 3) to learn the dynamic rates. The step size parameters were set as $\eta_t = .9/\sqrt{t}$ and $\rho_t = .005/\sqrt{t}$. The results are shown for DMD with the matrix $W$ known in advance, MD and Alg. 3 in Figure 4.

We again see several important characteristics in these plots. The first is that by incorporating knowledge of the dynamics, we incur significantly less loss than standard Mirror Descent. Secondly, we see that even without knowing
Figure 3: Zoomed in instantaneous predictions at $t = 1000$ for the experiment in Section 7.2. Top Left: $\theta_t$. Top Right: $\hat{\theta}_{\text{Right},t}$. Bottom Left: $\hat{\theta}_{\text{COMID},t}$. Bottom Right: $\hat{\theta}_t$. The prediction made with the prevailing motion is an accurate representation of the ground truth, while the prediction with the wrong dynamic is an unclear picture. The DFS algorithm correctly picks out the cleaner picture.

what the values of the matrix $W$, we can learn it simultaneously with the rate vectors $\mu_t$ from streaming data, and the resulting accurate estimate leads to low loss in the estimates of the rates.

8 Conclusions and future directions

Processing high-velocity streams of high-dimensional data is a central challenge to big data analysis. Scientists and engineers continue to develop sensors capable of generating large quantities of data, but often only a small fraction of that data is carefully examined or analyzed. Fast algorithms for sifting through such data can help analysts track dynamic environments and identify important subsets of the data which are inconsistent with past observations.

In this paper we have proposed a novel online optimization method, called Dynamic Mirror Descent (DMD), which incorporates dynamical models into the prediction process and yields low regret bounds for broad classes of comparator sequences. The proposed methods are applicable for a wide variety of observation models, noise distributions, and dynamical models. There is no assumption within our analysis that there is a “true” known underlying dynamical model, or that the best dynamical model is unchanging with time. The proposed
Figure 4: Experimental results tracking a self-exciting point process on a network, described in Section 7.3. Notice how the loss curve for Alg. 3 approaches the DMD curve (associated with clairvoyant knowledge of the underlying network matrix $W$) as the estimate of $W$ improves, and significantly outperforms conventional mirror descent.

Dynamic Fixed Share (DFS) algorithm adaptively selects the most promising dynamical model from a family of candidates at each time step. Additionally we show methods which learn in parametric families of dynamical models. In experiments DMD shows strong tracking behavior even when underlying dynamical models are switching, in such applications as dynamic texture analysis, compressive video, and self-exciting point process analysis.
A Proofs

A.1 Proof of Theorem 3

The proof of Theorem 3 shares some ideas with the tracking regret bounds of [12], but uses properties of the Bregman Divergence to eliminate some terms, while additionally incorporating dynamics. We employ the following lemma.

Lemma 9 Let the sequence \( \hat{\theta}_t \) be as in Alg. 1, and let \( \theta_T \) be an arbitrary sequence in \( \Theta^T \); then

\[
\ell_t(\hat{\theta}_t) - \ell_t(\theta_t) \leq \frac{1}{\eta_t} \left[ D(\theta_t \| \hat{\theta}_t) - D(\theta_{t+1} \| \hat{\theta}_{t+1}) \right]
+ \frac{\Delta_f}{\eta_t} + \frac{2M}{\eta_t} \| \theta_{t+1} - \Phi_t(\theta_t) \| + \frac{\eta_t}{2\sigma} G^2.
\]

Proof of Lemma 9: The optimality condition of (6a) implies

\[
\langle \nabla f_t(\hat{\theta}_t) + \nabla r(\hat{\theta}_{t+1}), \tilde{\theta}_{t+1} - \theta_t \rangle \leq \frac{1}{\eta_t} \langle \nabla \psi(\hat{\theta}_t) - \nabla \psi(\hat{\theta}_{t+1}), \tilde{\theta}_{t+1} - \theta_t \rangle.
\]

The proof has a similar structure to that in [22]

\[
f_t(\hat{\theta}_t) - f_t(\theta_t) + r(\hat{\theta}_t) - r(\theta_t)
= f_t(\hat{\theta}_t) - f_t(\theta_t) + r(\hat{\theta}_t) - r(\hat{\theta}_{t+1}) + r(\hat{\theta}_{t+1}) - r(\theta_t) \quad (10a)
\leq \langle \nabla f_t(\hat{\theta}_t), \hat{\theta}_t - \theta_t \rangle + \langle \nabla r(\hat{\theta}_t), \hat{\theta}_t - \hat{\theta}_{t+1} \rangle + \langle \nabla r(\hat{\theta}_{t+1}), \tilde{\theta}_{t+1} - \theta_t \rangle \quad (10b)
\leq \langle \nabla f_t(\hat{\theta}_t) + \nabla r(\hat{\theta}_{t+1}), \tilde{\theta}_{t+1} - \theta_t \rangle + \langle \nabla f(\hat{\theta}_t) + \nabla r(\hat{\theta}_t), \hat{\theta}_t - \tilde{\theta}_{t+1} \rangle \quad (10c)
\leq \frac{1}{\eta_t} \langle \nabla \psi(\hat{\theta}_t) - \nabla \psi(\hat{\theta}_{t+1}), \tilde{\theta}_{t+1} - \theta_t \rangle + \langle \nabla f_t(\hat{\theta}_t) + \nabla r(\hat{\theta}_t), \hat{\theta}_t - \tilde{\theta}_{t+1} \rangle \quad (10d)
= \frac{1}{\eta_t} \left[ D(\theta_t \| \hat{\theta}_t) - D(\theta_{t+1} \| \tilde{\theta}_{t+1}) - D(\tilde{\theta}_{t+1} \| \hat{\theta}_t) \right] + \langle \nabla f_t(\hat{\theta}_t) + \nabla r(\hat{\theta}_t), \hat{\theta}_t - \tilde{\theta}_{t+1} \rangle \quad (10e)
= \frac{1}{\eta_t} \left[ D(\theta_t \| \hat{\theta}_t) - D(\theta_{t+1} \| \tilde{\theta}_{t+1}) \right] + \frac{1}{\eta_t} \left[ D(\theta_{t+1} \| \tilde{\theta}_{t+1}) - D(\Phi_t(\theta_t) \| \tilde{\theta}_{t+1}) \right] \quad (10f)
+ \frac{1}{\eta_t} \left[ D(\Phi_t(\theta_t) \| \hat{\theta}_t) - D(\theta_t \| \tilde{\theta}_{t+1}) \right] - \frac{1}{\eta_t} D(\tilde{\theta}_{t+1} \| \hat{\theta}_t) + \langle \nabla f_t(\hat{\theta}_t) + \nabla r(\hat{\theta}_t), \hat{\theta}_t - \tilde{\theta}_{t+1} \rangle
\]

where (10c) follows from the convexity of \( f_t \) and \( r_t \), (10d) follows from the optimality condition in (6a), and (10e) follows from (4). Each of these terms can be bounded individually, and then recombined to complete the proof.

\[
T_1 = \psi(\theta_{t+1}) - \psi(\Phi_t(\theta_t)) - \langle \nabla \psi(\hat{\theta}_{t+1}), \theta_{t+1} - \Phi_t(\theta_t) \rangle \quad (11a)
\leq \langle \nabla \psi(\hat{\theta}_{t+1}), \theta_{t+1} - \Phi_t(\theta_t) \rangle \quad (11b)
\]
\[ \leq \| \nabla \psi (\theta_{t+1}) - \nabla \psi (\tilde{\theta}_{t+1}) \| \| \theta_{t+1} - \Phi_t (\theta_t) \| \]  

(11c)

\[ \leq 2M \| \theta_{t+1} - \Phi_{t+1} (\theta_t) \| \]  

(11d)

\[ T_2 = D (\Phi_t (\theta_t) \| \Phi_t (\tilde{\theta}_{t+1})) - D (\theta_t \| \tilde{\theta}_{t+1}) \leq \Delta \Phi \]  

(11e)

\[ T_3 \leq - \frac{\sigma}{2\eta_t} \| \tilde{\theta}_{t+1} - \tilde{\theta}_t \|^2 + \| \nabla f_t (\tilde{\theta}_t) + \nabla r (\tilde{\theta}_t) \| \| \tilde{\theta}_t - \tilde{\theta}_{t+1} \| \]  

(11f)

\[ \leq - \frac{\sigma}{2\eta_t} \| \tilde{\theta}_{t+1} - \tilde{\theta}_t \|^2 + \frac{\sigma}{2\eta_t} \| \tilde{\theta}_{t+1} - \tilde{\theta}_t \|^2 + \frac{\eta_t}{2\sigma} G^2 = \frac{\eta_t}{2\sigma} G^2 \]  

(11g)

Here, (11b) is due to the convexity of $\psi$ and (11c) is from the Cauchy-Schwarz inequality. Additionally, (11f) is due to (3) and (11g) uses Young’s Inequality. Combining these inequalities with (11a) gives the Lemma as it is stated. \( \square \)

The proof of the theorem is a matter of summing the bounds of Lemma 9 over time. For simplicity denote $D_t \equiv D (\theta_t \| \tilde{\theta}_t)$ and $V_t \equiv \| \theta_{t+1} - \Phi_t (\theta_t) \|$. Remember, we have assumed that $\Delta \Phi \leq 0$.

\[
R_T (\theta_T) \leq \sum_{t=1}^T \left( \frac{D_t}{\eta_t} - \frac{D_{t+1}}{\eta_{t+1}} \right) + \frac{G^2}{2\sigma} \sum_{t=1}^T \eta_t + D_{\max} \sum_{t=1}^T \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \sum_{t=1}^T 2M V_t \\
\leq \frac{D_1}{\eta_1} - \frac{D_{T+1}}{\eta_{T+1}} + D_{\max} \left( \frac{1}{\eta_{T+1}} - \frac{1}{\eta_T} \right) + \frac{2M}{\eta_T} V_T (\theta_T) + \frac{G^2}{2\sigma} \sum_{t=1}^T \eta_t \\
\leq \frac{D_{\max}}{\eta_{T+1}} + \frac{2M}{\eta_T} V_T (\theta_T) + \frac{G^2}{2\sigma} \sum_{t=1}^T \eta_t. \quad \square
\]

### A.2 Proof of Theorem 4

The tracking regret can be decomposed as:

\[
R_T (\theta) = \sum_{t=1}^T \ell_t (\tilde{\theta}_t) - \min_{i_1, \ldots, i_T} \sum_{t=1}^T \ell_t \left( \hat{\theta}_{i_t, t} \right) - \min_{\sum_{t=1}^T \mathbf{1}_{[i_t \neq i_{t+1}]} \leq m} \sum_{t=1}^T \ell_t (\tilde{\theta}_{i_t}) - \min_{\sum_{t=1}^T \mathbf{1}_{[i_t \neq i_{t+1}]} \leq m} \sum_{t=1}^T \ell_t (\theta_t) \quad (12)
\]

where the minimization in the second term of $T_4$ and first term of $T_5$ is with respect to sequences of dynamical models with at most $m$ switches, such that $\sum_{t=1}^T \mathbf{1}_{[i_t \neq i_{t+1}]} \leq m$. In (12), $T_1$ corresponds to the tracking regret of our algorithm relative to the best sequence of dynamical models within the DMD framework, and $T_2$ is the regret of that sequence relative to the best comparator in the class $\Theta_m$.  

23
We use Corollary 3 of [26] to bound $T_1$. A slight modification to their proof needs to be considered, because their losses are bounded between $[0, 1]$. In our case, we assume our loss function $\ell_t$ is convex, and Lipschitz on a closed, convex set $\Theta$. Therefore, we can say there exists a value $L$ such that $L \triangleq \max_{t \in [1, T], \theta \in \Theta} \ell_t(\theta) - \min_{t \in [1, T], \theta \in \Theta} \ell_t(\theta)$. This value can be easily incorporated into the proofs and bounds of [26] to give

$$T_4 \leq L^2 \sqrt{\frac{T}{2} \left( (m + 1) \ln N + (T - 1) b \left( \frac{m}{T - 1} \right) \right)}$$

$$\leq L^2 \sqrt{\frac{T}{2} \left( (m + 1) \log N + m \log T + 1 \right)}$$

where $b(x) = -x \log(x) - (1-x) \log(1-x)$ with respect to the natural logarithm. $T_2$ can be bounded using the method described in Section 4 on each time interval $[t_k, t_{k+1} - 1]$ and summing over the $m + 1$ intervals, yielding

$$T_5 \leq \frac{(m + 1)D_{\max}}{\eta_{T+1}} + \frac{2M}{\eta_T} V^{(m+1)}(\theta_T) + \frac{G^2}{2\sigma} \sum_{t=1}^{T} \eta_t. \quad \square$$

### A.3 Proof of Theorem 5

Let $\alpha_*$ to be the dynamical parameter in our candidate covering set, $A_N$ which minimizes the cumulative loss, $\alpha^*$ as the dynamical parameter in the entire space, $A$ which minimizes total loss, and $\tilde{\alpha}$ to be the parameter in $A_N$ closest to $\alpha^*$. Formally, we use the following definitions:

$$\alpha_* \triangleq \min_{\alpha \in A_N} \sum_{t=1}^{T} \ell_t(\hat{\theta}_{\alpha,t}),$$

$$\alpha^* \triangleq \min_{\alpha \in A} \sum_{t=1}^{T} \ell_t(\hat{\theta}_{\alpha,t}),$$

$$\tilde{\alpha} \triangleq \min_{\alpha \in A_N} \| \alpha - \alpha^* \|.$$

We decompose the regret in the following way:

$$R_T(\theta_T) = \sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{T} \ell_t(\hat{\theta}_{\alpha^*,t})$$

$$+ \sum_{t=1}^{T} \ell_t(\hat{\theta}_{\alpha^*,t}) - \sum_{t=1}^{T} \ell_t(\tilde{\theta}_{\alpha,t})$$

$$+ \sum_{t=1}^{T} \ell_t(\hat{\theta}_{\tilde{\alpha},t}) - \sum_{t=1}^{T} \ell_t(\theta)$$
To bound \( T_6 \) we use the bounds of [13] for the Exponentially Weighted Average Forecaster. Similarly to the proof of Theorem 4, the bound needs to be adjusted to account for the fact that our loss function, instead of being bounded by \([0,1]\) instead has arbitrary bounds \([\ell_{\min}, \ell_{\max}]\). Because, \(\ell_t\) is a convex function defined on a closed, convex set, these bounds are finite, and incorporating them into the proof of [13] is not difficult, yielding

\[
T_6 \leq (\ell_{\max} - \ell_{\min})\sqrt{2T \log(N)}.
\]

The term \( T_7 \) is upper bounded by 0 by the definitions of \( \alpha^* \) and \( \tilde{\alpha} \). Finally, the bound on \( T_8 \) is just the DMD regret bound (Theorem 3) with respect to \( \tilde{\alpha} \):

\[
T_8 \leq D_{\max} \eta_{T+1} + \frac{2M}{\eta} \sum_{t=1}^{T-1} \| \theta_{t+1} - \Phi_t(\theta_t, \tilde{\alpha}) \| + \frac{G^2}{2\sigma^2} \sum_{t=1}^{T} \eta_t.
\]

We now use the Lipschitz assumption on \( \Phi_t \) to show the bound with respect to \( \alpha^* \). Notice the variation term can be separated.

\[
V_{\Phi}(\theta_T) = \sum_{t=1}^{T-1} \| \Phi_t(\theta_t, \tilde{\alpha}) - \Phi_t(\theta_t, \alpha^*) + \Phi_t(\theta_t, \alpha^*) - \theta_{t+1} \|
\]

\[
\leq \sum_{t=1}^{T-1} \| \Phi_t(\theta_t, \alpha^*) - \theta_{t+1} \| + TL\varepsilon_N
\]

This shows we get a regret bound which scales like the variation from the best possible parameter \( \alpha^* \), in the set \( A \). Setting \( \eta_t \propto \frac{1}{\sqrt{t}} \) gives the result.

A.4 Proof of Lemma 7

The proof is by induction, and we assume without loss of generality that \( \beta = 0 \). Assume that \( \tilde{\mu}_{\alpha,t} = \tilde{\mu}_{0,t} + K_t \alpha \); this is trivially true for \( t = 1 \) since \( K_1 = 0 \) and \( \tilde{\mu}_{\alpha,1} = \tilde{\mu}_{\beta,1} \) by construction. For \( t \geq 1 \), applying (6) in this setting yields

\[
\begin{align*}
\tilde{\mu}_{0,t+1} &= A_t[(1 - \eta_t)\tilde{\mu}_{0,t} + \eta_t \phi(x_t)] + c_t \\
\tilde{\mu}_{\alpha,t+1} &= A_t[(1 - \eta_t)\tilde{\mu}_{\alpha,t} + \eta_t \phi(x_t)] + B_t \alpha + c_t \\
&= A_t[(1 - \eta_t)(\tilde{\mu}_{0,t} + K_t \alpha) + \eta_t \phi(x_t)] + B_t \alpha + c_t \\
&= A_t[(1 - \eta_t)\tilde{\mu}_{0,t} + \eta_t \phi(x_t)] + c_t + (1 - \eta_t)A_t K_t \alpha + B_t \alpha \\
&= \tilde{\mu}_{0,t+1} + (1 - \eta_t)A_t K_t \alpha + B_t \alpha \\
&= \tilde{\mu}_{0,t+1} + K_{t+1} \alpha. \quad \square
\end{align*}
\]

Notice that we must assume that \( \tilde{\theta}_t \) must lie on the interior of the set \( \Theta \) such that we can set the gradient to 0 to find the minimizer of equation 6a without projecting back onto the set.
A.5 Proof of Theorem 8

Since $\hat{\mu}_{\alpha,t}$ is an affine function of $\alpha$ given $\hat{\mu}_{\beta,t}$ for any $\beta$, and $\tilde{\ell}_t(\mu)$ is a convex function of $\mu$, we have that $g_t(\alpha)$ is a convex function of $\alpha$. For any $\alpha \in \mathbb{R}^n$, we have

$$
\sum_{t=1}^T \ell_t(\hat{\theta}_t) - \sum_{t=1}^T \ell_t(\theta_t) = \sum_{t=1}^T \ell_t(\hat{\theta}_t) - \sum_{t=1}^T \ell_t(\hat{\theta}_{\alpha,t}) + \sum_{t=1}^T \ell_t(\hat{\theta}_{\alpha,t}) - \sum_{t=1}^T \ell_t(\theta_t).
$$

To bound $T_9$, note that $\hat{\theta}_t = \hat{\theta}_{\alpha,t}$ and

$$
\sum_{t=1}^T \ell_t(\hat{\theta}_t) - \sum_{t=1}^T \ell_t(\hat{\theta}_{\alpha,t}) = \sum_{t=1}^T \ell_t(\hat{\mu}_t) - \sum_{t=1}^T \ell_t(\hat{\mu}_{\alpha,t}).
$$

Furthermore,

$$
g_t(\alpha) = \tilde{\ell}_t(\hat{\mu}_{\alpha,t}) = \tilde{\ell}_t(\hat{\mu}_{\alpha,t} + K_t(\alpha - \hat{\alpha}_t))
$$

is convex in $\alpha$, where we have used Lemma 7. Thus

$$
T_9 \leq \sum_{t=1}^T [\tilde{\ell}_t(\hat{\mu}_t) - \tilde{\ell}_t(\hat{\mu}_{\alpha,t})] = \sum_{t=1}^T [g_t(\hat{\alpha}_t) - g_t(\alpha)] \leq C_1 \sqrt{T}
$$

via [12]. The term $T_{10}$ can be bounded directly using Theorem 3, yielding for any $\alpha$

$$
T_{10} \leq C_2 \sqrt{T} \left(1 + \sum_{t=1}^T \|\theta_{t+1} - \Phi_t(\theta_t, \alpha)\|\right). \quad \square
$$

References

[1] E. Dumbill, “What is big data? An introduction to the big data landscape,” 2012. http://strata.oreilly.com/2012/01/what-is-big-data.html.

[2] W. Clavin, “Managing the deluge of ‘big data’ from space,” 2013. http://www.jpl.nasa.gov/news/news.php?release=2013-299.

[3] J. Matson, “Nasa readies a satellite to probe the sun–inside and out,” Scientific American, 2010. http://www.sciencemag.org/content/325/5939/724.

[4] P. E. Dewdney, P. J. Hall, R. T. Schilizzi, and T. J. L. W. Lazio, “The square kilometre array,” Proceedings of the IEEE, vol. 97, no. 8, 2009.

[5] J. Marlow, “What to do with 1,000,000,000,000,000 bytes of astronomical data per day,” 2012. Wired, http://www.wired.com/2012/04/what-to-do-with-10000000000000000-bytes-of-astronomical-data-per-day/.
[6] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*. Springer, 2009.

[7] L. Xie, Y. C. Soh, and C. E. de Souza, “Robust Kalman filtering for uncertain discrete-time systems,” *IEEE Trans. Autom. Control*, vol. 39, pp. 1310–1314, 1994.

[8] Y. Theodor and U. Shaked, “Robust discrete-time minimum-variance filtering,” *IEEE Trans. Sig. Proc.*, vol. 44(2), pp. 181–189, 1996.

[9] N. Merhav and M. Feder, “Universal prediction,” *IEEE Trans. Info. Th.*, vol. 44, pp. 2124–2147, October 1998.

[10] A. S. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization*. New York: John Wiley & Sons, 1983.

[11] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex programming,” *Operations Research Letters*, vol. 31, pp. 167–175, 2003.

[12] M. Zinkevich, “Online convex programming and generalized infinitesimal gradient descent,” in *Proc. Int. Conf. on Machine Learning (ICML)*, pp. 928–936, 2003.

[13] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games*. New York: Cambridge University Press, 2006.

[14] O. Banerjee, L. El Ghaoui, and A. d’Aspremont, “Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data,” *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.

[15] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, “High-dimensional Ising model selection using $\ell_1$-regularized logistic regression,” *Annals of Statistics*, vol. 38, pp. 1287–1319, 2010.

[16] E. Candès, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[17] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.*, vol. 15, pp. 1373–1396, June 2003.

[18] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, “Online adaptive estimation of sparse signals: Where rls meets the 11-norm,” *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3436–3447, 2010.

[19] L. M. Bregman, “The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming,” *Comput. Mathematics and Math. Phys.*, vol. 7, pp. 200–217, 1967.
[20] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications*. Oxford, UK: Oxford Univ. Press, 1997.

[21] N. Littlestone and M. K. Warmuth, “The weighted majority algorithm,” *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[22] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, “Composite objective mirror descent,” in *Conf. on Learning Theory (COLT)*, 2010.

[23] L. Xiao, “Dual averaging methods for regularized stochastic learning and online optimization,” *J. Mach. Learn. Res.*, vol. 11, pp. 2543–2596, 2010.

[24] J. Langford, L. Li, and T. Zhang, “Sparse online learning via truncated gradient,” *J. Mach. Learn. Res.*, vol. 10, pp. 777–801, 2009.

[25] M. Herbster and M. K. Warmuth, “Tracking the best linear predictor,” *Journal of Machine Learning Research*, vol. 35, no. 3, pp. 281–309, 2001.

[26] N. Cesa-Bianchi, P. Gaillard, G. Lugosi, and G. Stoltz, “A new look at shifting regret.” arXiv:1202.3323, 2012.

[27] E. Hazan and C. Seshadhri, “Efficient learning algorithms for changing environments,” in *Proc. Int. Conf on Machine Learning (ICML)*, pp. 393–400, 2009.

[28] B. McMahan, “A unified view of regularized dual averaging and mirror descent with implicit updates.” arXiv:1009.3240v2, 2011.

[29] S. Haykin, *Adaptive Filter Theory*. New Jersey: Prentice-Hall, 2002.

[30] A. Rakhlin and K. Sridharan, “Online learning with predictable sequences.” arXiv:1208.3728, 2012.

[31] M. Herbster and M. K. Warmuth, “Tracking the best expert,” *Machine Learning*, vol. 32, pp. 151–178, 1998.

[32] W. Koolen and S. de Rooij, “Combining expert advice efficiently,” in *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pp. 275–286, 2008.

[33] D. Adamskiy, W. M. Koolen, A. Chernov, and V. Vovk, “A closer look at adaptive regret,” in *Proceedings of the 23rd international conference on Algorithmic Learning Theory*, ALT’12, pp. 290–304, 2012.

[34] A. Gyorgy, T. Linder, and G. Lugosi, “Efficient tracking of large classes of experts,” *IEEE Transaction on Information Theory*, vol. 58, pp. 6709–6725, November 2012.

[35] V. Vovk, “Aggregating algorithms,” *Conf. on Learning Theory (COLT)*, 1990.
[36] C. Shalizi, A. Jacobs, K. Klikner, and A. Clauset, “Adapting to non-
stationarity with growing expert ensembles.” arXiv:1103:0949, 2011.

[37] S. Amari and H. Nagaoka, Methods of Information Geometry. Providence: American Mathematical Society, 2000.

[38] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” Foundations and Trends in Machine Learning, vol. 1, pp. 1–305, December 2008.

[39] K. S. Azoury and M. K. Warmuth, “Relative loss bounds for on-line density estimation with the exponential family of distributions,” Machine Learning, vol. 43, pp. 211–246, 2001.

[40] M. Raginsky, R. Willett, C. Horn, J. Silva, and R. Marcia, “Sequential anomaly detection in the presence of noise and limited feedback,” IEEE Transactions on Information Theory, vol. 58, no. 8, pp. 5544–5562, 2012.

[41] J.-B. Hiriart-Urruty and C. Lemaréchal, Fundamentals of Convex Analysis. Berlin: Springer, 2001.

[42] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, UK: Cambrdige Univ. Press, 2004.

[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” SIAM J. Optim., vol. 19, no. 4, pp. 1574–1609, 2009.

[44] M. Szummer and R. W. Picard, “Temporal texture modeling,” in Proceedings of International Conference on Image Processing (ICIP), 1996.

[45] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, “Dynamic textures,” International Journal of Computer Vision, vol. 51, no. 2, pp. 91–109, 2003.

[46] A. Ravichandran, R. Chaudhry, and R. Vidal, “Dynamic texture toolbox,” 2011. http://www.vision.jhu.edu.

[47] V. Studer, J. Bobin, M. Chahid, H. S. Mousavi, E. Candes, and M. Dahan, “Compressive fluorescence microscopy for biological and hyperspectral imaging,” Proceedings of the National Academy of Sciences, vol. 109, no. 26, pp. E1679–E1687, 2012.

[48] R. Marcia and R. Willett, “Compressive coded aperture video reconstruction,” in Proc. European Signal Processing Conference EUSIPCO, 2008.

[49] J. Y. Park and M. B. Wakin, “A multiscale framework for compressive sensing of video,” in Picture Coding Symposium (PCS), (Chicago, IL), May 2009.
[50] A. C. Sankaranarayanan, C. Studer, and R. G. Baraniuk, “CS-MUVI: video compressive sensing for spatial-multiplexing cameras,” in 2012 IEEE International Conference on Computational Photography (ICCP), pp. 1–10, Apr. 2012.

[51] D. Angelosante, G. B. Giannakis, and E. Grossi, “Compressed sensing of time-varying signals,” in Intl Conf. on Dig. Sig. Proc., 2009.

[52] N. Vaswani and W. Lu, “Modified-CS: Modifying compressive sensing for problems with partially known support,” IEEE Trans. Sig. Proc., vol. 58, pp. 4595–4607, 2010.

[53] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, “Single pixel imaging via compressive sampling,” IEEE Sig. Proc. Mag., vol. 25, no. 2, pp. 83–91, 2008.

[54] C. Blundell, K. A. Heller, and J. M. Beck, “Modelling reciprocating relationships with hawkes processes,” in Proc. NIPS, 2012.

[55] S. W. Linderman and R. P. Adams, “Discovering latent network structure in point process data.” arXiv:1402.0914, 2014.

[56] A. Stomakhin, M. B. Short, and A. Bertozzi, “Reconstruction of missing data in social networks based on temporal patterns of interactions,” Inverse Problems, vol. 27, no. 11, 2011.

[57] E. N. Brown, R. E. Kass, and P. P. Mitra, “Multiple neural spike train data analysis: state-of-the-art and future challenges,” Nature neuroscience, vol. 7, no. 5, pp. 456–461, 2004.