Learning from Censored and Dependent Data: The case of Linear Dynamics

Orestis Plevrakis ORESTISP@PRINCETON.EDU

Princeton University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Observations from dynamical systems often exhibit irregularities, such as *censoring*, where values are recorded only if they fall within a certain range. Censoring is ubiquitous in practice, due to saturating sensors, limit-of-detection effects, image frame effects, and combined with temporal dependencies within the data, makes the task of system identification particularly challenging.

In light of recent developments on learning linear dynamical systems (LDSs), and on censored statistics with *independent* data, we revisit the decades-old problem of learning an LDS, from censored observations (Lee and Maddala (1985); Zeger and Brookmeyer (1986)). Here, the learner observes the state $x_t \in \mathbb{R}^d$ if and only if x_t belongs to some set $\mathcal{S}_t \subseteq \mathbb{R}^d$. We develop the first computationally and statistically efficient algorithm for learning the system, assuming only oracle access to the sets $\mathcal{S}_t$. Our algorithm, *Stochastic Online Newton with Switching Gradients*, is a novel second-order method that builds on the Online Newton Step (ONS) of Hazan et al. (2007). Our Switching-Gradient scheme does not always use (stochastic) gradients of the function we want to optimize, which we call *censor-aware* function. Instead, in each iteration, it performs a simple test to decide whether to use the censor-aware, or another *censor-oblivious* function, for getting a stochastic gradient.

In our analysis, we consider a "generic" Online Newton method, which uses arbitrary vectors instead of gradients, and we prove an error-bound for it. This can be used to appropriately design these vectors, leading to our Switching-Gradient scheme. This framework significantly deviates from the recent long line of works on censored statistics (e.g, Daskalakis et al. (2018); Kontonis et al. (2019); Daskalakis et al. (2019)), which apply Stochastic Gradient Descent (SGD), and their analysis reduces to establishing conditions for off-the-shelf SGD-bounds. Our approach enables to relax these conditions, and gives rise to phenomena that might appear counterintuitive, given the previous works. Specifically, our method makes progress even when the current "survival probability" is exponentially small. We believe that our analysis framework will have applications in more settings where the data are subject to censoring.

Keywords: Truncated/Censored Data, Linear Dynamical Systems, Time-Series, Online Learning

1. Introduction

System identification is the problem of learning the evolution equations of a dynamical system from data. Mathematically, we have a sequence of system states $(x_t)_t$, and observations $(y_t)_t$ evolving as

$$x_{t+1} = f(x_t, u_t, w_t), \;\; y_t = g(x_t, v_t),$$

where u_t's are inputs to the system, w_t is process noise, and v_t is sensor noise. At each step t, the learner observes the input u_t, and the resulting output y_{t+1}. The goal is to learn the functions f

and g, from an observed trajectory.[1] In this paper, we consider system identification with *censored* observations, where the learner observes y_t if and only if y_t belongs to some set $\mathcal{S}_t$, which we call the *observable* set.

Dynamical data with missing observations are ubiquitous in practice (Honaker and King, 2010). This is often due to censoring, which frequently manifests in the fields of signal processing and control theory (Yang and Li, 2009), time series analysis (Lee and Maddala, 1985), business (Hausman and Wise, 1977), economics (Johannsen and Mertens, 2018), in medicine, and in physical sciences. For example, consider learning the dynamics of some target dynamical system, using observations from cameras. Here, censoring naturally arises due to occlusions blocking visibility, or from the system exiting the camera frame. Despite the numerous applications, the problem is not statistically and computationally understood even for linear, fully-observable dynamics:

$$x_{t+1} = A_* x_t + B_* u_t + w_t, \;\; y_t = x_t,$$

where $x_t \in \mathbb{R}^d$, $u_t \in \mathbb{R}^m$, and $w_t \overset{\text{i.i.d}}{\sim} N(0, I)$. Actually, even for one-dimensional linear systems $(d = 1)$, with no inputs $(u_t = 0)$, and observable sets $\mathcal{S}_t$ being half-lines, no known efficient algorithm learns (the scalar) A_*, not even asymptotically. Specifically, the existing methods fall in two categories: 1) iterative methods trying to maximize the non-concave log-likelihood, as in Lee and Maddala (1985) and Zeger and Brookmeyer (1986), and 2) EM-based imputation methods (e.g., Park et al. (2007)). Both of them are not guaranteed to recover the underlying system. On top of that, for large dimension d, the likelihood-based approach is inefficient to implement because it requires computing high-dimensional integrals over $\mathcal{S}_t$'s.

In this work, we study the multidimensional linear case with no inputs: $x_{t+1} = A_* x_t + w_t$. We allow *arbitrary* observable sets $\mathcal{S}_t$, with only requirement being that we have oracle access to each $\mathcal{S}_t$, namely, given a point x, the oracle efficiently computes $\mathbb{1}\{x \in \mathcal{S}_t\}$. For this model, we obtain the first computationally and statistically efficient algorithm for learning A_*, under the following assumptions (stated informally here, and in detail in Section 3):

Assumption 1: The system is *stable*, i.e., the spectral radius $\rho(A_*)$ is less than one. Stability is a classical assumption in linear dynamical systems (LDSs). If $\rho(A_*) > 1$, then the state explodes exponentially, with high probability.

Assumption 2: The number of times we observe the state is at least a constant fraction (say 1%) of the trajectory-length.

Assumption 3: For most of the timesteps t, given that we observed x_t, the probability of observing x_{t+1} is at least a constant (say 1%).

To motivate the last two assumptions, we note that for the simpler problems of censored Gaussian estimation and censored linear regression (with independent data), the only known computationally and statistically efficient algorithms assume that the probability of observing a sample is at least a constant (see Daskalakis et al. (2018), Daskalakis et al. (2019)). Assumptions 2 and 3 are the adaption of this, to fit our dynamical setting. Under these assumptions, our estimation error bound matches (up to logarithmic factors) the best known bound for learning uncensored LDSs (Simchowitz et al. (2018)), with respect to the dimension d, the trajectory-length T, and the spectrum

[1] Some works consider access to multiple trajectories.

of A_*. It is also the first, to the best of our knowledge, estimator that can accommodate arbitrary observable sets, since previous works considered intervals, half-lines, and products of these (e.g., Zeger and Brookmeyer (1986), Yang and Li (2009)). Key for tackling this decades-old problem are recent advances in 1) (uncensored) linear system identification theory, and 2) censored/truncated statistics (CTS) for independent data.

Learning Linear Dynamics. The difficulty in learning an LDS, compared to standard linear regression, is that the observations are dependent. A classical approach that avoids this issue is based on the system's mixing time, which is roughly $\tau_{\text{mix}} = \frac{1}{1-\rho(A_*)}$ steps. Thus, the learner can use one x_t every τ_{mix} steps, and reduce the analysis to standard linear regression (e.g., Yu (1994)). However, the resulting bounds get worse with larger τ_{mix}, and as pointed in Simchowitz et al. (2018), this behavior is qualitatively incorrect. Intuitively, larger A_* gives larger states, which implies larger signal-to-noise ratio, i.e., easier estimation. The authors provided bounds that express this intuition, and that was the first sharp analysis for stable systems.[2] This progress initiated a large line of work on learning LDSs (see Section 1.1).

Censored/Truncated Statistics (CTS). Consider the problem of learning a Gaussian distribution, having access only to samples from some set $\mathcal{S}$. These are called *truncated* samples. Censoring is when we also know the number of unobserved samples, as in our case, where this number can be inferred from the lengths of time-intervals during which we do not observe anything. Truncation and censoring go back to at least Galton (1898) and Pearson (1902), and there has been a large volume of research devoted to them (see Cohen (1991)). Nevertheless, the first provably computationally and statistically efficient algorithm, for learning a truncated Gaussian, was only recently discovered by Daskalakis et al. (2018). The authors developed a general algorithmic framework, based on Stochastic Gradient Descent (SGD), which bypasses the computation of high-dimensional integrals over $\mathcal{S}$. The result and the generality of the approach created a lot of excitement, and a large number of subsequent works applied the SGD framework to other problems in CTS (for independent data), e.g., linear regression by Daskalakis et al. (2019).

Our Contributions

We build on the above advances, by introducing new algorithmic and technical ideas, which we now overview.

Our algorithm. Our first observation is that non-convexity of the negative log-likelihood can be bypassed by focusing on "paired observations", i.e., (x_t, x_{t+1}) such that $x_t \in \mathcal{S}_t$ and $x_t \in \mathcal{S}_{t+1}$. By ignoring the other terms in the objective, we get a convex function. The second observation is that for LDSs, if an SGD-based algorithm relies on the Markovian property $\mathbb{E}[x_{t+1}|x_t, x_{t-1}, \ldots, x_1] = A_* x_t$ to produce *unbiased* gradient estimates, then the algorithm should process the data in *temporal order*. The reason is that if we "see" x_t and x_{t+s} (for some $s \geq 2$), then the expectation of x_{t+1} is not $A_* x_t$, i.e., $\mathbb{E}[x_{t+1}|x_t, x_{t+s}] \neq A_* x_t$. In other words, we need an *online* algorithm. Unfortunately, for censored linear regression, the SGD framework of Daskalakis et al. (2019) processes the data in *random* order, and also does multiple passes over them. This is not a technicality; making that algorithm online will lead to slow statistical rates, as we explain later. For this reason, we design a new stochastic *second-order* method, building on the Online Newton Step method (ONS) of Hazan

[2] The authors also considered $\rho(A_*) = 1$, known as the *marginally stable* regime.

et al. (2007). The crucial difference with ONS, and all recent works on CTS, is that we do not always use (stochastic) gradients of the function we want to optimize, which we call *censor-aware* function. Instead, in each iteration, we perform a simple test, based on which we decide whether to use the censor-aware function, or another *censor-oblivious* function, for getting a stochastic gradient. We call our method *Stochastic Online Newton with Switching Gradients* (SON-SG, given in Algorithm 1), and we show that it can be combined with a least-squares-based warmup procedure, to learn censored LDSs (Algorithm 4). We also show that SON-SG can be applied to an even broader setting, i.e., general linear-response time-series (Section 4).

Algorithm-design framework. We came up with this switching-gradient scheme, by considering a "generic" ONS method, where instead of gradients we have an arbitrary vector sequence $g_1, g_2, \ldots$. We proved a general estimation error bound for this method, which serves as a guideline for designing the g_i's. A similar framework for designing g_i's has been proposed in the context of non-convex optimization (for first-order methods), by Arora et al. (2015). This approach gives a lot of freedom for algorithm design, compared to previous works on CTS which apply off-the-shelf SGD-bounds, and we believe it will have more applications in censored and truncated statistics.

Technical contributions and insights. In all recent CTS papers (see Section 1.1 for an extensive list), a crucial step for proving parameter recovery by SGD is establishing *anti-concentration* of truncated Gaussians $\mathcal{N}(\mu, \Sigma, \mathcal{S})$, where $\mathcal{N}(\mu, \Sigma, \mathcal{S})$ is the Gaussian $\mathcal{N}(\mu, \Sigma)$, given that the sample is in $\mathcal{S}$. Daskalakis et al. (2018) reduced this task to showing that the "survival probability" is large, i.e., $\mathbb{P}_{x \sim \mathcal{N}(\mu,\Sigma)}[x \in \mathcal{S}] \geq \Omega(1)$. Hence, this and all follow-up works have focused on lower bounding survival probabilities. In our case, this methodology does not apply, due to temporal dependencies. However, our approach (generic ONS bound) enables to significantly relax the high survival probability condition, and gives rise to phenomena that might appear counterintuitive, given the intuition built in the recent literature. Specifically, in high dimensions ($d \to \infty$) our analysis deals with cases where the survival probability is exponentially small ($e^{-\Omega(d)}$), while at the same time anti-concentration tends to infinity. Our other technical contribution is a lower-bound on the covariance matrix of the *observed* states. Simchowitz et al. (2018) proved such a bound when all states are observed. Here, due to censoring, we observe only a subset of the whole trajectory. For independent data, Daskalakis et al. (2018) address this issue using union-bound over all possible subsets.[3] Unfortunately, for LDSs union-bound gives vacuous guarantees, because for a fixed subset, the concentration degrades with larger mixing times. We resolve this difficulty, by generalizing the "small-ball" technique of Simchowitz et al. (2018), and we lower-bound the covariance matrices of *all subsets* of size $\Omega(T)$ *simultaneously*, where T is the trajectory-length.

1.1. Further Related Work

Censored and Truncated Statistics. As we mentioned, there has been a long line of recent works on several settings within CTS: Gaussian parameter estimation (Daskalakis et al., 2018; Kontonis et al., 2019), linear, logistic and probit regression (Daskalakis et al., 2019; Ilyas et al., 2020), compressed sensing (Daskalakis et al., 2020), sparse graphical models (Bhattacharyya et al., 2020), estimation of boolean product distributions (Fotakis et al., 2020), mixtures of Gaussians (Nagarajan

[3] This is implicit in their analysis. They use a result of Diakonikolas et al. (2019), which applies union bound over subsets.

and Panageas, 2020). All these works consider independent data and apply the SGD framework[4] introduced in Daskalakis et al. (2018).

Linear System Identification. Even though linear system identification is a decades-old field (Ljung, 1999), a sharp non-asymptotic theory was only recently developed (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019; Tsiamis and Pappas, 2019; Oymak and Ozay, 2019; Simchowitz et al., 2019).

Online Convex Optimization. To design our algorithm, we build on ideas from online convex optimization (OCO). In the recent years, OCO has been extensively used for learning and controlling LDSs (e.g., Agarwal et al. (2019); Hazan et al. (2020); Simchowitz et al. (2020); Ghai et al. (2020); Simchowitz (2020)). For a general overview of OCO see Hazan (2019).

2. Notation

For every vector x, we use $\|x\|$ to denote the ℓ_2 norm $\|x\|_2$. Also, we use $\langle A, B\rangle = \operatorname{tr}(A^\top B)$ to denote the matrix inner product. For a matrix A and a $\Sigma \succ 0$, we define the norm $\|A\|_\Sigma = \sqrt{\langle A^\top A, \Sigma\rangle}$ [5]. The covariance matrix between two random vectors x, y is $\mathrm{Cov}[x, y]$. For a sequence $(x_t)_{t=1}^T$, we denote by $x_{\leq\tau}$, $x_{<\tau}$ and $x_{-\tau}$ the subsequences $(x_t)_{t\leq\tau}, (x_t)_{t<\tau}, (x_t)_{t\neq\tau}$ respectively. For a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^d$, and a measurable $\mathcal{S} \subseteq \mathbb{R}^d$, we define the survival probability $\mathcal{N}(\mu, \Sigma; \mathcal{S}) = \mathbb{P}_{x\sim\mathcal{N}(\mu,\Sigma)}[x \in \mathcal{S}]$. We also define the truncated Gaussian $\mathcal{N}(\mu, \Sigma, \mathcal{S})$ to be $\mathcal{N}(\mu, \Sigma)$ conditioned on taking values in $\mathcal{S}$. Finally, whenever we say that a set $\mathcal{S}$ is "revealed" to the learner, we mean that she has access to a membership oracle $M_\mathcal{S}$, i.e., an efficient procedure that computes the $\mathbb{1}\{x \in \mathcal{S}\}$, for any point x.

3. Censored Linear Dynamics: Model, and Main Theorem

We study the system $x_{t+1} = A_* x_t + w_t$, where $x_t \in \mathbb{R}^d$, $A_* \in \mathbb{R}^{d\times d}$ and $w_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$. Starting from $x_0 = 0$,[6] consider the trajectory $x_1, x_2, \ldots, x_{T+1}$. The learner has access to censored observations, i.e., there is a process of observable sets $(\mathcal{S}_t)_t$, and she observes x_t if and only if $x_t \in \mathcal{S}_t$. Also, at time t, the set $\mathcal{S}_t$ is revealed to her. Now, $\mathcal{S}_t$'s may depend on the state-trajectory, but we assume that given x_t, the state x_{t+1} and the set $\mathcal{S}_{t+1}$ are statistically independent. To see why this is a natural assumption, consider the camera-based example (Section 1), and think of $x_t \in \mathbb{R}^3$ as the position of some object, and $\mathcal{S}_t$ as the visible part of the space at time t. Having observed x_t, the camera could adapt its frame (affecting $\mathcal{S}_{t+1}$), to improve the chances for observing x_{t+1}, but without knowing the next "excitation" w_t. We now formally state our assumptions, sketched in Section 1.

Assumption 1 *A_* is diagonalizable and stable, i.e., $A_* = UDU^{-1}$, where D is diagonal and $\rho(A_*) = \max_i |D_{ii}| < 1$.*[7]

Let $\mathcal{O}$ be the set of observation times: $\mathcal{O} = \{t \in [T] : x_t \in \mathcal{S}_t\}$. We assume that we observe "enough" data:

[4] Exception is Nagarajan and Panageas (2020) who consider the EM-algorithm.

[5] Usually, $\|A\|_\Sigma$ is used to denote $\sqrt{\langle AA^T, \Sigma\rangle}$. However, this definition is more appropriate for our setting.

[6] We assume $x_0 = 0$ to simplify the exposition. Our proofs generalize for any x_0, by paying a $\log\|x_0\|$ factor in the bound.

[7] Note that S and D can have complex entries.

Assumption 2 *For a known constant $\beta \in (0,1)$, with probability $1 - o(1)$, we have $|\mathcal{O}| \geq \beta T$, where $o(1)$ denotes a $\delta_T \to 0$, as $T \to \infty$.*

Let $\mathcal{B}(a)$ be the set of timesteps t, at which 1) we observe x_t, and 2) given x_t and $\mathcal{S}_{t+1}$, the probability of observing x_{t+1} is less than a, i.e., $\mathcal{B}(a) = \{t \in \mathcal{O} : \mathcal{N}(A_* x_t, I; \mathcal{S}_{t+1}) < a\}$.

Assumption 3 *For a known constant $\alpha \in (0,1)$, and some bound $L > 0$, with probability $1 - o(1)$, we have $|\mathcal{B}(\alpha)| \leq L$. Also, $\mathbb{E}\big[|\mathcal{B}(\alpha)|\big] \leq L$.*

Our bounds will depend on L, and will match the uncensored case if $L \leq \widetilde{O}(d)$.[8] As we mentioned in the introduction, Assumptions 2 and 3 are the adaptation of the $\Omega(1)$-survival-probability assumption in Daskalakis et al. (2018). We further motivate Assumption 3 with a natural one-dimensional example.

Example 1 *Let $x_{t+1} = a_* x_t + w_t$, where $a_* \in [0,1)$. The observable set is a static half-line: $\mathcal{S}_t = \mathcal{S} = \{x \in \mathbb{R} : x \geq \lambda\}$, $\lambda > 0$. We claim that here, Assumption 2 implies Assumption 3 with $\alpha = \Omega(1)$ and $L = 0$. This is clear for $a_* \approx 1$, since if $x_t \geq \lambda$, then $a_* x_t + w_t \geq x_t \geq \lambda$ with probability almost $1/2$.[9] For general $a_* \in [0,1)$, the implication is less obvious (see Appendix B).*

As in Simchowitz et al. (2018), our bounds depend on the controllability Gramian Γ_T, defined as $\Gamma_T := \sum_{s=0}^{T-1} A_*^s (A_*^s)^\top$. This matrix quantifies how much the noise process excites the system. In the theorem that follows, we use $\widetilde{\Theta}(1)$ to denote polylogarithmic factors in T and in cond(U), where cond(U) is the condition number of the eigenvector-matrix U.

Theorem 1 *Under Assumptions 1, 2, 3, there exist $C_{\alpha,\beta}, C'_{\alpha,\beta} = \widetilde{\Theta}(1) \cdot \text{poly}\left(\frac{1}{\alpha\beta}\right)$, such that if*

$$T \geq C'_{\alpha,\beta} \cdot \left(d^2 + \frac{d}{1 - \rho(A_*)} + dL\right),$$

then with probability at least 99%, Algorithm 4 runs in polynomial time, and outputs an $\widehat{A}$ such that

$$\left\|\widehat{A} - A_*\right\|_{\Gamma_T} \leq C_{\alpha,\beta} \sqrt{\frac{d^2 + dL}{T}}. \tag{1}$$

Remark 2 *For uncensored LDSs, and stable-diagonalizable A_*, the best known (spectral-norm) bound is $\left\|(\widehat{A} - A_*)\Gamma_T^{1/2}\right\|_2 \leq \widetilde{O}\left(\sqrt{d/T}\right)$. So, the Frobenius version is $\|\widehat{A} - A_*\|_{\Gamma_T} \leq \widetilde{O}\left(\sqrt{d^2/T}\right)$, which matches (1), if $\alpha, \beta = \Omega(1)$, and $L = \widetilde{O}(d)$. We leave the spectral-norm bound for censored LDSs for future work.*

Remark 3 *It is possible to drop the assumption that A_* is diagonalizable by paying an exponential dependence in the size of the largest Jordan block. This dependence (for a bound on $\|\cdot\|_{\Gamma_T}$) appears even in the uncensored case (see Ghai et al. (2020) for a discussion on this).*

As we mentioned, our algorithm is SON-SG, preceded by a least-squares warmup procedure. Let $\mathcal{P}$ be the set of of pairs (x_t, x_{t+1}), such that both $t, t+1 \in \mathcal{O}$. We will ignore all "isolated" observations, i.e., the ones not participating in any pair of $\mathcal{P}$. As we will see, by ignoring them This decision is justified by the following proposition:

[8] $\widetilde{O}(\cdot)$ hides logarithmic factors.

[9] Note that for $a_* \approx 1$, we did not need Assumption 2, to show Assumption 3. However, we need it for general a_*.

Proposition 4 *Let $M := |\mathcal{P}|$. With probability $1 - o(1)$, we have $M \geq \alpha\beta T/2$.*

We prove Proposition 4 in Appendix C, using Assumptions 2,3, and that $T >> L$. Let $\mathcal{P}_0$ be the first (in time-order) $\lfloor M/2 \rfloor$ pairs of $\mathcal{P}$, and $\mathcal{P}_1 = \mathcal{P} \setminus \mathcal{P}_0$. Our algorithm first uses $\mathcal{P}_0$ to create a (rough) confidence ellipsoid $\mathcal{K}$ that includes A_*, with high probability. Then, it uses $\mathcal{K}$ as constraint-set for SON-SG, which will operate on $\mathcal{P}_1$. In the next section, we present and analyze SON-SG in a more general setting, which reveals the key structure that our method exploits. Then, in Section 6, we give the full-algorithm and conclude the proof of Theorem 1.

4. Truncated Time Series with Linear Responses

For timesteps 1 to T, consider a covariate-response process $(x_t, y_t)_t$, where $x_t \in \mathbb{R}^d$, $y_t \in \mathbb{R}^n$. The learner only observes a subset of the data, based on a Bernoulli process $(o_t)_t$, i.e., if $o_t = 1$, then she observes (x_t, y_t), otherwise the pair is hidden. In untruncated linear-response time series, we have $y_t = A_* x_t + w_t$, where $w_t \sim N(0, I)$, and $A_* \in \mathbb{R}^{n \times d}$. Here, when the learner gets to see a datapoint, the noise will be biased due to truncation. Formally, consider a process of observable sets $(S_t)_t$, where $S_t \subseteq \mathbb{R}^n$. Let $\mathcal{F}_t$ be the σ-algebra generated by $x_{\leq t}$, $y_{<t}$, $o_{\leq t}$ and $S_{\leq t}$ (note that y_t is not in $\mathcal{F}_t$). We assume that given $\mathcal{F}_t$ and $o_t = 1$, the set S_t is revealed to the learner, and $y_t \sim \mathcal{N}(A_* x_t, I, S_t)$. This model has three notable special cases:

1. **Untruncated, linear-response time series.** This model was studied by Simchowitz et al. (2018), and corresponds to $o_t = 1$ and $S_t = \mathbb{R}^n$, for all t.
2. **Truncated linear regression.** First considered by Tobin (1958), this model was revisited by Daskalakis et al. (2019), and corresponds to $o_t = 1$ for all t, and it requires independent data, i.e., given x_t, the response y_t is independent of x_{-t}.[10]
3. **LDS with censored observations.** Here, $y_t = x_{t+1}$, $o_t = \mathbb{1}\{x_t \in \mathcal{S}_t \ \wedge \ x_{t+1} \in \mathcal{S}_{t+1}\}$, and $S_t = \mathcal{S}_{t+1}$. Note that we pretend we do not observe the "isolated" observations, which is aligned with what our algorithm does.

Initial confidence ellipsoid. SON-SG receives as input a rough initial estimate $A_0 \in \mathbb{R}^{n \times d}$, and a $\Sigma_0 \succ 0$ that represents a confidence ellipsoid $\mathcal{K} = \{A \in \mathbb{R}^{n \times d} : \ \|A - A_0\|_{\Sigma_0} \leq 1\}$. For LDSs, we will later show how to use $\mathcal{P}_0$ to construct a $\mathcal{K}$ with the following property:

Definition 5 *A confidence ellipsoid $\mathcal{K} = \{A \in \mathbb{R}^{n \times d} : \ \|A - A_0\|_{\Sigma_0} \leq 1\}$ is (R, ω)-accurate, for some $R, \omega > 0$, if (a) $A_* \in \mathcal{K}$, (b) $\Sigma_0 \succcurlyeq \omega \cdot I$, and (c) for all t, $\left\|\Sigma_0^{-1/2} x_t\right\|$ is R^2-subgaussian.*[11]

Theorem 6 *Fix $R, R_w, R_x, L, \alpha > 0$, and let $B(\alpha) := \{t \in [T] : \ o_t = 1, \ \mathcal{N}(A_* x_t, I; S_t) < \alpha\}$. Suppose that (a) we are given an (R, ω)-accurate confidence ellipsoid $\mathcal{K}$, (b) $\mathbb{E}[|B(\alpha)|] \leq L$, (c) for all t, the noise $w_t = y_t - A_* x_t$ has norm $\|w_t\|$ that is R_w^2-subgaussian, and (d) $\mathbb{E}[\|x_t\|^2] \leq R_x^2$. Let $t_1 < t_2 < \cdots < t_N$ be the observation times ($o_{t_i} = 1$). If the total number of steps $T \geq \mathrm{poly}(1/\alpha)$, then SON-SG (Algorithm 1) outputs an $\widehat{A}$ such that*

$$\mathbb{E}\left[\left\|\widehat{A} - A_*\right\|_{\Sigma}^2\right] \leq \frac{D + LD'}{N}, \tag{2}$$

[10] To be precise, Daskalakis et al. (2019) consider one-dimensional responses and fixed truncation set. However, the extension to multidimensional y_t's and time-varying truncation sets is relatively straightfoward.

[11] A random variable X is called σ^2-subgaussian if $\mathbb{P}[|X| \geq \delta \cdot \sigma] \leq \exp\left(-\delta^2/C\right)$, where $C = O(1)$.

where $\Sigma = \frac{1}{N}\sum_{i=1}^{N} x_{t_i} x_{t_i}^T$*, and* $D = \widetilde{O}(1) \cdot dD'$*,* $D' = \widetilde{O}(1) \cdot \mathrm{poly}(1/\alpha) \cdot (d + R^2 + R_w^2)$*. Here,* $\widetilde{O}(1)$ *denotes a polylogarithmic factor in the parameters defined in this section.*

Remark 7 *Note that in the LDS case,* $R_w = O(\sqrt{d})$ *by standard concentration of Gaussian norm. However, for slowly mixing systems (*$\rho(A_*) \to 1$*),* R_x *can grow polynomially with* T*. Our bounds only degrade in the logarithm of* R_x*, which is absorbed in* $\widetilde{O}(\cdot)$*, and the same happens with* $1/\omega$*.*

Before presenting SON-SG and its analysis, we give some background on existing techniques for CTS.

4.1. Existing Techniques and their Limitations

Even though Daskalakis et al. (2019) consider truncated linear regression with no temporal dependencies, we will use the same likelihood-based objective. Specifically, let

$$\ell_S(\mu; y) := -\frac{1}{2}\|y-\mu\|^2 - \log\left(\int_S \exp\left(-\frac{1}{2}\|z-\mu\|^2\right) dz\right),$$

and observe that given $\mathcal{F}_t$ and that $o_t = 1$, we have that for a candidate matrix A, the log-likelihood for y_t is $\ell_{S_t}(Ax_t; y_t)$. Also, let $f_t(A)$ be the (negative) population log-likelihood:

$$f_t(A) := -\mathbb{E}_y\left[\ell_{S_t}(Ax_t; y) \,\middle|\, \mathcal{F}_t, o_t = 1\right],$$

and observe that $f_t(A) = -\mathbb{E}_{y\sim\mathcal{N}(A_* x_t, I, S_t)}\left[\ell_{S_t}(Ax_t; y)\right]$. Given the observed data, our goal will be to minimize $f(A) := \frac{1}{N}\sum_{i=1}^{N} f_{t_i}(A)$. To see why $f(A)$ is a “good” objective, we take the first and second derivatives of $f_t(A)$:[12]

$$\nabla f_t(A) = \mathbb{E}_{z\sim\mathcal{N}(Ax_t, I, S_t)}\left[z\right] x_t^\top - \mathbb{E}_{y\sim\mathcal{N}(A_* x_t, I, S_t)}\left[y\right] x_t^\top$$

$$\nabla^2 f_t(A) = \mathrm{Cov}_{z\sim\mathcal{N}(Ax_t, I, S_t)}\left[z, z\right] \otimes \left(x_t x_t^\top\right), \tag{3}$$

where $\nabla^2 f_t(A)$ is used to denote $\nabla^2 f_t\left(\mathrm{vec}\left(A^\top\right)\right)$, and $\mathrm{vec}(\cdot)$ is the standard vectorization. Now, note that $\nabla f_t(A_*) = 0$, and so $\nabla f(A_*) = 0$. Also, since the Kronecker product of positive semidefinite matrices (PSD) is PSD, $f_t(A)$ is convex, and so $f(A)$ is also convex.[13] Now, if f was *strongly-convex* ($\nabla^2 f(A) \succ 0$), then A_* would be the unique optimal solution, justifying the use of the objective. Even though strong-convexity does not hold, Daskalakis et al. (2019) show how to address this for independent data, by restricting A in some set that contains A_*, and $\nabla^2 f(A) \succeq \Omega(1) \cdot I$, inside the set. Let's assume (for now) that here, $\nabla^2 f(A) \succeq \Omega(1) \cdot I$ holds. Now, note that a priori, it is not clear how to optimize $f(A)$, since we do not have a closed-form expression. An important conceptual contribution of Daskalakis et al. (2019) is the observation that by sampling $z_t \sim \mathcal{N}(Ax_t, I, S_t)$,[14] and computing $v_t = (y_t - z_t)x_t^\top$, we have $\mathbb{E}\left[v_t \mid \mathcal{F}_t, o_t = 1\right] = \nabla f_t(A)$, i.e,

[12] $\otimes$ denotes Kronecker product.

[13] As we mentioned in the introduction, for censored LDSs, the negative log-likelihood is non-convex. However, here we have convexity, because $f(A)$ corresponds to a part of the overall log-likelihood.

[14] This is done via rejection sampling, using the membership oracle for S_t.

we get an unbiased gradient estimate. Based on this observation, they employ a variant of SGD to minimize $f(A)$, but their algorithm is tailored to independent data. The reason is that it processes the data in random order, and also does multiple passes over them. Thus, by the time it computes v_t, it is very likely that before that, it had processed (x'_t, y'_t), for $t' > t$. Because of this, if the data are temporally dependent, v_t can be a *biased* estimate of $\nabla f_t(A)$.

4.2. Our Approach

To avoid the above issue, we need to process the data in temporal order. This is exactly the case for Online Convex Optimization (OCO). In OCO though, the goal is not to recover some parameter, but to minimize *regret*. However, regret bounds can often be transformed to statistical recovery rates via "online-to-batch" conversions (e.g., Cesa-Bianchi et al. (2004)). In our setting, this conversion can be done, but is trickier than usual, and we will deal with it later. Now, since we are aiming for a fast $\widetilde{O}(1/N)$-rate, the first natural attempt is online SGD, which has only logarithmic (in N) regret, provided all f_{t_i}'s are strongly-convex (Hazan (2019))[15]. Unfortunately, this in not true for any f_{t_i}, due to the rank-one component $x_{t_i}x_{t_i}^T$ in the Hessian $\nabla^2 f_{t_i}(A)$ (3). However, there is still structure we can exploit. Notice that every row of $\nabla f_{t_i}(A)$ has the same direction as x_{t_i}, which corresponds exactly to that "problematic" rank-one component in the Hessian. This structure is reminiscent of the *exp-concavity* property (Hazan et al. (2007)), which essentially is strong-convexity, *in the direction of the gradient*:

Definition 8 *A function f is called λ-exp-concave, if for all x, $\nabla^2 f(x) \succcurlyeq \lambda \cdot \nabla f(x) \nabla f(x)^\top$.*[16]

For exp-concave functions, the Online Newton Step (ONS) algorithm, introduced in Hazan et al. (2007), has regret that depends logarithmically in N. Unfortunately, this result does not apply here:

Obstacles for ONS. First, f_{t_i}'s are not necessarily exp-concave (unless λ is exponentially small, which is not useful). This is because of the covariance term in 3. To the best of our knowledge, the idea used in Daskalakis et al. (2019) to restrict A is some appropriate set, does not resolve this issue, due to temporal dependencies. The second obstacle, is that the regret bound of ONS (Hazan et al. (2007)) will have linear dependence in R_x, which as we said can grow as $\mathrm{poly}(T)$.

4.3. Stochastic Online Newton with Switching Gradients

We now describe SON-SG. First, we use as projection-set the ellipsoid $\mathcal{K}$. Second, we use preconditioning as in ONS, but while in ONS the preconditioner has outer-products of the gradients, here we use outer-products of the covariates. This is done to simplify the analysis.

Switching Gradients. The crucial difference with ONS is the choice of the g_i's, by the "SwitchGrad" function. To ease notation, suppose we are at time t, and $t = t_i$ for some i. We define $A(t) = A_i$, and $g(t) = g_i$. In ONS, $g(t)$ would simply be $\nabla f_t(A(t))$. Of course, we do not have access to this gradient, but as we said we can get a stochastic gradient by sampling $z_t \sim \mathcal{N}(A(t)x_t, I, S_t)$, and setting $g(t) = (z_t - y_t)x_t^\top$. Sampling from $\mathcal{N}(A(t)x_t, I, S_t)$ can be done via rejection sampling, using the membership oracle. However, to be efficient, the mass $\gamma_t := \mathcal{N}(A(t)x_t, I; S_t)$ should be sufficiently large. Unfortunately, here γ_t can be exponentially small.

[15] In online-to-batch conversions, a regret-bound $R_N > 0$ often translates to R_N/N statistical recovery rate.

[16] These functions are called exp-concave, because this property is equivalent to $e^{-\lambda f(x)}$ being concave.

Algorithm 1 Stochastic Online Newton with Switching Gradients

```
Input: $A_0 \in \mathbb{R}^{n \times d}$, PSD matrix $\Sigma_0 \in \mathbb{R}^{d \times d}$, data $(x_{t_i}, y_{t_i})_{i=1}^{N}$.
$\eta = (2/\alpha)^{c_\eta}$                                  $\triangleright c_\eta \geq 0$ is a large constant.
$\mathcal{K} = \{A \in \mathbb{R}^{n \times d} : \ \|A - A_0\|_{\Sigma_0} \leq 1\}$
 $A_1 = A_0$
 for $i = 1$ to $N$ do
    $g_i =$ SwitchGrad$(A_i x_{t_i}, x_{t_i}, y_{t_i}, S_{t_i})$
    $\Sigma_i = \Sigma_{i-1} + x_{t_i} x_{t_i}^\top$
    $\widetilde{A}_{i+1} = A_i - \eta \cdot g_i \Sigma_i^{-1}$
    $A_{i+1} = \arg\min_{A \in \mathcal{K}} \|A - \widetilde{A}_{i+1}\|_{\Sigma_i}$
end
return $\widehat{A} = A_{N+1}$
```

Algorithm 2 SwitchGrad

```
Input: $\mu, x, y, S$
$z = \mu$
if Test$(\mu, S)$ then
    Sample $z' \sim \mathcal{N}(\mu, I, S)$ via rejection sampling using the membership oracle $M_S$.
    $z = z'$
end
return $g = (z - y)x^T$
```

Algorithm 3 Test

```
Input: $\mu, S$.
$\gamma = (\alpha/2)^{c_\gamma}$, $k = \frac{4}{\gamma} \log T$.                $\triangleright c_\gamma \geq 0$ is a large constant.
Sample $\xi_1, \ldots, \xi_k \overset{\text{i.i.d}}{\sim} N(\mu, I)$
$p = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}\{\xi_j \in S\}$
return $(p \ \geq \ 2\gamma)$
```

However, the case of small γ_t is easily recognizable, by estimating γ_t via sampling from the normal $\mathcal{N}(A(t)x_t, I)$, and counting how many times we hit S_t, using the membership oracle. This is done by the "Test" function. If the Test returns "True", then with high probability (w.h.p), $\gamma_t \geq \alpha^{O(1)}$, and so we can efficiently sample a stochastic gradient, and assign it to $g(t)$. If it returns "False", then w.h.p, $\gamma_t \leq \alpha^{\Omega(1)}$. They key idea here is that, as we will show, γ_t being small is actually an "easy" case, and simply choosing $g(t) = (A(t)x_t - y_t)x_t^\top$ suffices to make progress towards A_*. Observe that here $g(t)$ is a stochastic gradient of $\widetilde{f}_t(A) = -\mathbb{E}_{y \sim \mathcal{N}(A_* x_t, I, S_t)} \left[\ell_{\mathbb{R}^n}(Ax_t; y)\right]$. In other words, the Test is a "switch" between ∇f_t and $\nabla \widetilde{f}_t$.

5. Proof of Theorem 6

In this section, we give an overview of the proof of Theorem 6, with an emphasis on the novel technical components. Our goal is to convey the key ideas, and so at some steps we are slightly informal. We provide the formal and detailed proof in Appendix D.

The Generic Bound. We first prove a bound on $\left\|\widehat{A} - A_*\right\|_\Sigma^2$ (remember that $\Sigma = \Sigma_N$ in Algorithm 1), which holds for any sequence $g_1, g_2, \ldots, g_N$. This bound will serve as a guideline for choosing the g_i's.

Lemma 9 *Independently of how g_i's are chosen,*

$$\left\|\widehat{A} - A_*\right\|_\Sigma^2 \le 1 - \sum_{i=1}^{N} \left(2\eta\langle g_i, A_i - A_*\rangle - \|(A_i - A_*)x_{t_i}\|^2\right) + \eta^2 \sum_{i=1}^{N} \operatorname{tr}\left(g_i \Sigma_i^{-1} g_i^\top\right). \tag{4}$$

The proof of the lemma is along the lines of the analysis of ONS in Hazan et al. (2007), and we provide it in Appendix D.1. Let E_1 be the first sum in 4, and E_2 the second. The rest of the proof is about showing that for our choice of g_i's, E_1 is (almost) non-negative, and E_2 is not too large (both in expectation). In the main text we present the crux of the proof, which is the bound on E_1. Furthermore, in the main text we assume that $L = 0$, since the extension for general L is straightforward.

Bound on E_1.

Fix a time t, condition on $\mathcal{F}_t$, and suppose that $t = t_i$, for some i. To ease notation, we define $A(t) = A_i$, $g(t) = g_i$, $\mu_t^* = A_* x_t$, $\mu_t = A(t)x_t$, and $V_t = 2\eta\langle g(t), A(t) - A_*\rangle - \|(A(t) - A_*)x_t\|^2$. To bound $\mathbb{E}[E_1]$, we show (roughly) that $\mathbb{E}\left[V_t \mid \mathcal{F}_t, t = t_i\right] \ge 0$. Observe that this immediately implies $\mathbb{E}[E_1] \ge 0$. We consider two cases, based on what the Test function returns.

Large Survival Probability. Suppose at time t, Test returns "True". We call this event $\mathcal{T}_t$, and here we assume that Test returning "True" implies $\mathcal{N}(\mu_t, I; S_t) \ge \gamma$ (this is correct w.h.p.).[17] Furthermore, $z_t \sim \mathcal{N}(\mu_t, I, S_t)$, and $g(t) = (z_t - y_t)x_t^T$. Also, since $y_t \sim \mathcal{N}(\mu_t^*, I, S_t)$, we have

$$\mathbb{E}\left[V_t \mid \mathcal{F}_t, t = t_i, \mathcal{T}_t\right] = 2\eta\langle \nu_t - \nu_t^*, \mu_t - \mu_t^*\rangle - \|\mu_t - \mu_t^*\|^2, \tag{5}$$

where $\nu_t = \mathbb{E}_{z\sim\mathcal{N}(\mu_t, I, S_t)}[z]$ and $\nu_t^* = \mathbb{E}_{y\sim\mathcal{N}(\mu_t^*, I, S_t)}[y]$. Now, note that since $t = t_i$ and $L = 0$, we have $\mathcal{N}(\mu_t^*, I; S_t) \ge \alpha \ge \gamma$. The key component of our proof is the following lemma, which we prove here in detail.

Lemma 10 *Let $\mathcal{L}_S(\mu; \mu^*) = -\mathbb{E}_{y\sim\mathcal{N}(\mu^*, I, S)}\left[\ell_S(\mu; y)\right]$, where $\mu, \mu^* \in \mathbb{R}^n$ and $S \subseteq \mathbb{R}^n$. Suppose that $\mathcal{N}(\mu, I; S)$, $\mathcal{N}(\mu^*, I; S) \ge \gamma$. Then, there exists an absolute constant $c > 0$ such that*

$$\left\langle \nabla_\mu \mathcal{L}_S(\mu; \mu^*), \mu - \mu^* \right\rangle \ge (\gamma/2)^c \cdot \|\mu - \mu^*\|^2. \tag{6}$$

[17] The parameter γ is defined in Algorithm 3. Also, in Appendix D, where we give the details for handling the low-probability events, we show that $\mathbb{E}[E_1] \ge \Delta$, for some $\Delta > 0$, but small.

Lemma 10 implies that, given large enough c_η,[18] the expectation in (5) is non-negative. Indeed, note that $\nabla_\mu L_S(\mu;\mu^*) = \nu - \nu^*$, where $\nu = \mathbb{E}_{z\sim\mathcal{N}(\mu,I,S)}[z]$ and $\nu^* = \mathbb{E}_{y\sim\mathcal{N}(\mu^*,I,S)}[z]$, so

$$\begin{aligned}\mathbb{E}\big[V_t \bigm| \mathcal{F}_t, t = t_i, \mathcal{T}_t\big] \geq (2\eta(\gamma/2)^c - 1)\cdot\|\mu_t - \mu_t^*\|^2 &= \left(2\,(2/\alpha)^{c_\eta}\,(\alpha/4)^{c\cdot c_\gamma} - 1\right)\cdot\|\mu_t - \mu_t^*\|^2 \\ &= \left(\frac{2^{c_\eta+1-2c\cdot c_\gamma}}{a^{c_\eta - c\cdot c_\gamma}} - 1\right)\cdot\|\mu_t - \mu_t^*\|^2 \geq 0,\end{aligned}$$

for any constant $c_\eta \geq 2c\cdot c_\gamma$. We now prove Lemma 10.

Proof Let $s(\gamma) := \sqrt{2\log(1/\gamma)} + 1$. We will need two technical claims, which hold for any $\mu,\mu^*\in\mathbb{R}^n$ and $S\subseteq\mathbb{R}^n$.

Claim 1 *Let $\nu = \mathbb{E}_{z\sim\mathcal{N}(\mu,I,S)}[z]$. If $\mathcal{N}(\mu,I;S)\geq\gamma$, then $\|\nu-\mu\|\leq s(\gamma)$.*

Claim 2 *Suppose $\mathcal{N}(\mu^*,I;S)\geq\gamma$, and for some $\widetilde{\mu}$ we have $\|\widetilde{\mu}-\mu^*\|\leq c\cdot s(\gamma)$. Then,*

$$\mathit{Cov}_{z\sim\mathcal{N}(\widetilde{\mu},I,S)}[z,z] \succcurlyeq (\gamma/2)^{\mathrm{poly}(c)}\cdot I.$$

Both Claims are mainly from Daskalakis et al. (2018). Claim 1 is Lemma 6 in that paper. Claim 2 is not explicitly stated, but can be derived by slightly adapting their proof (see Appendix H.1). We consider two cases:

Case 1: $\|\mu-\mu^*\|\geq 4s(\gamma)$. Then,

$$\begin{aligned}\langle\nu-\nu^*,\mu-\mu^*\rangle &= \|\mu-\mu^*\|^2 + \langle\nu-\mu,\mu-\mu^*\rangle + \langle\mu^*-\nu^*,\mu-\mu^*\rangle \\ &\geq\ \|\mu-\mu^*\|^2 - \|\nu-\mu\|\cdot\|\mu-\mu^*\| - \|\nu^*-\mu^*\|\cdot\|\mu-\mu^*\|\end{aligned}$$

Since $\mathcal{N}(\mu,I;S)$, $\mathcal{N}(\mu^*,I;S)\geq\gamma$, Claim 1 implies $\|\nu-\mu\|,\|\nu^*-\mu^*\|\leq s(\gamma)$. Being in Case 1,

$$\langle\nu-\nu^*,\mu-\mu^*\rangle\ \geq\ \|\mu-\mu^*\|\cdot\big(\|\mu-\mu^*\|-2s(\gamma)\big)\ \geq\ \|\mu-\mu^*\|^2/2.$$

Case 2: $\|\mu-\mu^*\| < 4s(\gamma)$. To ease notation, fix μ^*, S, and let $\mathcal{L}(\mu) = \mathcal{L}_S(\mu;\mu^*)$. From fundamental theorem of calculus,

$$\nabla\mathcal{L}(\mu) - \nabla\mathcal{L}(\mu^*) = \int_0^1 \nabla^2\mathcal{L}\left(\mu(\theta)\right)d\theta\cdot(\mu-\mu^*),$$

where $\mu(\theta) := \mu^* + \theta(\mu-\mu^*)$. Since $\nabla L(\mu^*) = 0$,

$$\big\langle\nabla\mathcal{L}(\mu),\mu-\mu^*\big\rangle = \int_0^1\Big\langle\nabla^2\mathcal{L}\left(\mu(\theta)\right)(\mu-\mu^*),\mu-\mu^*\Big\rangle d\theta. \tag{7}$$

For $\theta\in(0,1)$, $\|\mu(\theta)-\mu^*\| = \theta\|\mu-\mu^*\|\leq O\left(s(\gamma)\right)$. Using $\mathcal{N}(\mu^*,I,S)\geq\gamma$ and Claim 2,

$$\nabla^2\mathcal{L}\left(\mu(\theta)\right) = \mathrm{Cov}_{z\sim\mathcal{N}(\mu(\theta),I,S)}[z,z]\succcurlyeq(\gamma/2)^{O(1)}\cdot I.$$

Using 7, we finish the proof. ■

[18] Remember that $\eta = (2/\alpha)^{c_\eta}$.

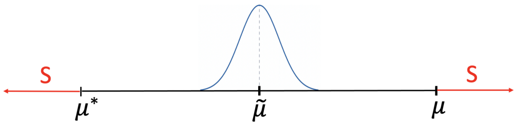

Figure 1: $\Omega(d)$ variance

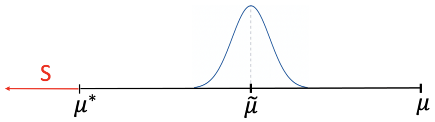

Figure 2: $O(1/d)$ variance

Remark. We want to highlight an important qualitative difference between Lemma 10 and all recent works in CTS. Suppose $\gamma = \Omega(1)$ and let $v = \frac{\mu-\mu^*}{\|\mu-\mu^*\|}$. From the argument in Case 2, 6 can be rephrased as

$$\mathbb{E}_{\widetilde{\mu}\sim[\mu^*,\mu]}\Big[\big\langle \mathrm{Cov}_{z\sim\mathcal{N}(\widetilde{\mu},I,S)}[z,z]\cdot v, v\big\rangle\Big] \geq \Omega(1), \tag{8}$$

where $\widetilde{\mu}$ is distributed uniformly on the segment $[\mu,\mu^*]$. In other words, the average variance in the v-direction is $\Omega(1)$. This is an anti-concentration bound. As we mentioned, proving anti-concentration bounds for truncated normals $\mathcal{N}(\widetilde{\mu}, I, S)$ is a core component in all recent works in CTS, where this task is reduced to lower bounding $\mathcal{N}(\widetilde{\mu}, I; S)$. However, inequality (8) cannot be proven with this methodology. We illustrate this with an insightful example (Figure 1). In the example, we consider $\mu,\mu^* \in \mathbb{R}^d$ and $\|\mu-\mu^*\| = \sqrt{d}$, because the guaranteed bound in the LDS-case for $\|\mu_t - \mu_t^*\|$ will be $\Theta(\sqrt{d})$. Let $\mu^* = 0$, $\mu = \sqrt{d}\cdot e_1$ (e_1 is the standard-basis vector), $S = \{x \in \mathbb{R}^d : x_1 \in (-\infty, 0] \cup [\sqrt{d}, +\infty)\}$. Observe that the conditions of Lemma 10 are satisfied with $\gamma = 1/2$. However, for most $\widetilde{\mu} \in [\mu^*, \mu]$, the mass $\mathcal{N}(\widetilde{\mu}, I; S)$ is exponentially small (in d). Consider now $\widetilde{\mu}$ exactly in the middle of μ^*, μ. Even though $\mathcal{N}(\widetilde{\mu}, I; S) = \exp(-\Omega(d))$, the variance in the e_1 direction is $\Omega(d)$, due to symmetry. It can actually be shown that precisely this $\widetilde{\mu}$ and small perturbations of it make the average variance $\Omega(1)$ in (8). Note that it is necessary for both corners (μ and μ^*) to have high survival probability, e.g., in Figure 2 (for the same distance-scales) it can be shown that the variance is $O(1/d)$.

Small Survival Probability. Suppose that at time t, Test returns "False" ($\neg\mathcal{T}_t$). Suppose also that this implies $\mathcal{N}(\mu_t, I; S_t) \leq 4\gamma$ (again this is w.h.p). We will show that $\mathbb{E}\big[V_t \bigm| \mathcal{F}_t, t = t_i, \neg\mathcal{T}_t\big] \geq 0$. Observe that given $\neg\mathcal{T}_t$, we have $g(t) = (\mu_t - y_t)x_t^\top$, and so

$$\mathbb{E}\big[V_t \bigm| \mathcal{F}_t, t = t_i, \neg\mathcal{T}_t\big] = 2\eta\langle \mu_t - \nu_t^*, \mu_t - \mu_t^*\rangle - \|\mu_t - \mu_t^*\|^2. \tag{9}$$

Again, since $t = t_i$ and $L = 0$, we have $\mathcal{N}(\mu_t^*, I; S_t) \geq \alpha$, so by Claim 1, $\|\nu_t^* - \mu_t^*\| \leq s(\alpha)$. We need one more technical claim.

Claim 3 *If $\mathcal{N}(\mu^*, I; S) \geq \alpha$, and $\|\mu - \mu^*\| \leq c \cdot s(\alpha)$, then $\mathcal{N}(\mu, I; S) \geq (\alpha/2)^{\mathrm{poly}(c)}$.*

Again, Claim 3 is proven in Appendix H.2, by slightly adapting the analysis of Daskalakis et al. (2018). Now, we claim that $\|\mu_t - \mu_t^*\| > 2s(\alpha)$, provided that $c_\gamma = O(1)$ is sufficiently large. Indeed, suppose $\|\mu_t - \mu_t^*\| \leq 2s(\alpha)$. Combining with $\mathcal{N}(\mu_t^*, I; S_t) \geq \alpha$ and Claim 3, we get $4(\alpha/2)^{c_\gamma} = 4\gamma \geq \mathcal{N}(\mu_t, I; S_t) \geq (\alpha/2)^{O(1)}$, which is a contradiction for large enough $c_\gamma = O(1)$. Using that $\|\nu_t^* - \mu_t^*\| \leq s(\alpha)$, we get

$$\begin{aligned}\langle \mu_t - \nu_t^*, \mu_t - \mu_t^*\rangle &= \|\mu_t - \mu_t^*\|^2 + \langle \mu_t^* - \nu_t^*, \mu_t - \mu_t^*\rangle \geq \|\mu_t - \mu_t^*\|^2 - \|\mu_t^* - \nu_t^*\| \cdot \|\mu_t - \mu_t^*\| \\ &\geq \|\mu_t - \mu_t^*\| \cdot \big(\|\mu_t - \mu_t^*\| - s(\alpha)\big) \geq \|\mu_t - \mu_t^*\|^2/2.\end{aligned}$$

Since $\eta \geq 1$, Equation 23 implies $\mathbb{E}\left[V_t \mid \mathcal{F}_t, t = t_i, \neg\mathcal{T}_t\right] \geq 0$. Thus, $\mathbb{E}[V_t \mid \mathcal{F}_t, t = t_i] \geq 0$, and $\mathbb{E}[E_1] \geq 0$.

6. The Initial Ellipsoid and Proof of Theorem 1

Here, we provide the overall algorithm (Algorithm 4), and an outline of the proof of Theorem 1. Let $\mathcal{I}_0 = \{t \in [T] : (x_t, x_{t+1}) \in \mathcal{P}_0\}$, and $\mathcal{I}_1 = \{t \in [T] : (x_t, x_{t+1}) \in \mathcal{P}_1\}$.

Algorithm 4 SON-SG for censored linear dynamics

```
$A_0 = \arg\min_{A \in \mathbb{R}^{d \times d}} \sum_{t \in \mathcal{I}_0} \|x_{t+1} - Ax_t\|^2$
$\Sigma_0 = \frac{1}{s \cdot |\mathcal{I}_0|} \cdot \sum_{t \in \mathcal{I}_0} x_t x_t^\top$,  where $s = c_s \left(\sqrt{\log(1/\alpha)} + 1\right)$.        ▷$c_s \geq 0$ is a large constant.
Get $\widehat{A}$ by running SON-SG (Algorithm 1) with $A_0, \Sigma_0$, and dataset $\mathcal{P}_1$.
return $\widehat{A}$
```

The least-squares method when all x_t's are observed was analyzed in Simchowitz et al. (2018). Here, we observe only a subset the x_t's. In Appendix G, we generalize the "small-ball" technique of Simchowitz et al. (2018), to prove the following theorem.

Theorem 11 *There exist $c_1, c_2, c_3 = \mathrm{poly}\left(\frac{1}{\alpha\beta}\right)$, such that if $T \geq \widetilde{\Theta}(c_1) \cdot \left(d^2 + \frac{d}{1-\rho(A_*)} + dL\right)$, then with probability $1 - o(1)$, the ellipsoid $\mathcal{K} = \{A \in \mathbb{R}^{d \times d} : \|A - A_0\|_{\Sigma_0} \leq 1\}$ is (R, ω)-accurate with $R = c_2\sqrt{d}$, $\omega = 1/c_2$, and also $\frac{1}{|\mathcal{I}_1|} \sum_{t \in \mathcal{I}_1} x_t x_t^T \succcurlyeq \frac{1}{c_3}\Gamma_T$.*

We now prove Theorem 1 using Theorems 6 and 11.

Proof We use the time-series model with $y_t = x_{t+1}$, $S_t = \mathcal{S}_{t+1}$, $o_t = \mathbb{1}(t \in \mathcal{I}_1)$. Thus, $B(\alpha) \subseteq \mathcal{B}(\alpha)$, so $\mathbb{E}[|B(\alpha)|] \leq L$. Let $\mathcal{E}_0$ be the event that all guarantees in Theorem 11 hold, so $\mathbb{P}[\neg\mathcal{E}_0] \leq o(1)$. As we mentioned, $\|w_t\|$ is $O(d)$-subgaussian. Also, in Appendix H.3, we show that $\mathbb{E}[\|x_t\|^2]$ is polynomial in all parameters. Let $\Sigma = \frac{1}{|\mathcal{I}_1|} \sum_{t \in \mathcal{I}_1} x_t x_t^T$. By conditioning on $\mathcal{E}_0$ and applying Theorem 6, using Markov's inequality in (2), we have that with probability at least $1 - \delta$, [19]

$$\left\|\widehat{A} - A_*\right\|_\Sigma^2 \leq \frac{\widetilde{O}(C)}{\delta N} \cdot \left(d(d + R^2 + R_w^2) + L(d + R^2 + R_w^2)\right) \leq \widetilde{O}(C) \cdot \frac{d^2 + dL}{\delta N}. \tag{10}$$

where $C = \mathrm{poly}\left(\frac{1}{\alpha\beta}\right)$. Now, on $\mathcal{E}_0$, $\Sigma \succcurlyeq \Gamma_T / \mathrm{poly}\left(\frac{1}{\alpha\beta}\right)$. From Proposition 4, $\mathbb{P}\left[N < \frac{\alpha\beta T}{4}\right] \leq o(1)$. Thus, by choosing small enough $\delta = \Omega(1)$, we are done. ■

References

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

[19] Observe that our bounds depend on $1/\delta$. It is possible to achieve $\log(1/\delta)$ dependence, but with more technical proofs. To keep the presentation simpler, we chose not to optimize this dependence.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on learning theory*, pages 113–149. PMLR, 2015.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Arnab Bhattacharyya, Rathin Desai, Sai Ganesh Nagarajan, and Ioannis Panageas. Efficient statistics for sparse graphical models from truncated samples. *arXiv preprint arXiv:2006.09735*, 2020.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

A Clifford Cohen. *Truncated and censored samples: theory and applications*. CRC press, 1991.

Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*, pages 955–960, 2019.

Constantinos Daskalakis, Dhruv Rohatgi, and Manolis Zampetakis. Truncated linear regression in high dimensions. *arXiv preprint arXiv:2007.14539*, 2020.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. Efficient parameter estimation of truncated boolean product distributions. In *Conference on Learning Theory*, pages 1586–1600. PMLR, 2020.

Francis Galton. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1898.

Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. *arXiv preprint arXiv:2002.02064*, 2020.

Jerry A Hausman and David A Wise. Social experimentation, truncated distributions, and efficient estimation. *Econometrica: Journal of the Econometric Society*, pages 919–938, 1977.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pages 408–421. PMLR, 2020.

James Honaker and Gary King. What to do about missing values in time-series cross-section data. *American journal of political science*, 54(2):561–581, 2010.

Andrew Ilyas, Emmanouil Zampetakis, and Constantinos Daskalakis. A theoretical and practical framework for regression and classification from truncated samples. In *International Conference on Artificial Intelligence and Statistics*, pages 4463–4473. PMLR, 2020.

Benjamin Kramer Johannsen and Elmar Mertens. A time series model of interest rates with the effective lower bound. 2018.

Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lung-Fei Lee and GS Maddala. The common structure of tests for selectivity bias, serial correlation, heteroscedasticity and non-normality in the tobit model. *International Economic Review*, pages 1–20, 1985.

Lennart Ljung. System identification. *Wiley encyclopedia of electrical and electronics engineering*, pages 1–19, 1999.

Sai Ganesh Nagarajan and Ioannis Panageas. On the analysis of em for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, pages 634–659. PMLR, 2020.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Jung Wook Park, Marc G Genton, and Sujit K Ghosh. Censored time series analysis with autoregressive moving average models. *Canadian Journal of Statistics*, 35(1):151–168, 2007.

Karl Pearson. On the systematic fitting of frequency curves. *Biometrika*, 2:2–7, 1902.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. *arXiv preprint arXiv:2006.05910*, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Fuwen Yang and Yongmin Li. Set-membership filtering for systems with sensor saturation. *Automatica*, 45(8):1896–1902, 2009.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Scott L Zeger and Ron Brookmeyer. Regression analysis with censored autocorrelated data. *Journal of the American Statistical Association*, 81(395):722–729, 1986.