Subgradient Method for System Identification with Non-Smooth Objectives

Baturalp Yalcin BYALCIN@BERKELEY.EDU
Jihun Kim JIHUN.KIM@BERKELEY.EDU
Javad Lavaei LAVAEI@BERKELEY.EDU
Department of Industrial Engineering and Operations Research, University of California, Berkeley

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

This paper investigates a subgradient-based algorithm to solve the system identification problem for linear time-invariant systems with non-smooth objectives. This is essential for robust system identification in safety-critical applications. While existing work provides theoretical exact recovery guarantees using optimization solvers, the design of fast learning algorithms with convergence guarantees for practical use remains unexplored. We analyze the subgradient method in this setting, where the optimization problems to be solved evolve over time as new measurements are collected, and we establish linear convergence to the ground-truth system for both the best and Polyak step sizes after a burn-in period. We further characterize sublinear convergence of the iterates under constant and diminishing step sizes, which require only minimal information and thus offer broad applicability. Finally, we compare the time complexity of standard solvers with the subgradient algorithm and support our findings with experimental results. This is the first work to analyze subgradient algorithms for system identification with non-smooth objectives.

Keywords: Subgradient Algorithm, System Identification, Exact Recovery

1. Introduction

Dynamical systems form the foundation of modern problems, such as reinforcement learning, auto-regressive models, and control systems. These systems evolve based on their current state, with the system dynamics equations determining the next state. Due to the complexity of real-world systems, these dynamics are often unknown or difficult to model precisely. Learning the unknown dynamics of a dynamical system is known as the *system identification* problem.

The system identification problem has been extensively studied in the literature (Chen and Guo, 2012; Lennart, 1999). The least-squares estimator (LSE) is the most widely studied approach for this problem as it provides a closed-form solution. Early research focused on the asymptotic (infinite-time) convergence properties of LSE (Ljung, 1978). Recently, finite-time (non-asymptotic) analyses gained popularity, leveraging high-dimensional statistical techniques for both linear (Simchowitz et al., 2018; Faradonbeh et al., 2018; Tsiamis and Pappas, 2019; Dean et al., 2020) and nonlinear dynamical systems (Foster et al., 2020; Sattar and Oymak, 2022; Ziemann et al., 2022). See Tsiamis et al. (2023) for a survey. Traditionally, both asymptotic and non-asymptotic analyses of LSE focused on systems with independent and identically distributed (sub)-Gaussian disturbances. However, this assumption is often unrealistic for safety-critical systems such as autonomous vehicles, unmanned aerial vehicles, and power grids. In these settings, disturbances may be designed by an adversarial agent, leading to dependencies across time and arbitrary variations.

Since LSE is susceptible to outliers and temporal dependencies, recent research focused on robust non-smooth estimators for safety-critical systems. Feng and Lavaei (2021) and Feng et al. (2023) analyzed the properties of these estimators through the lens of the Null Space Property. Later works modeled disturbances as zero with probability $1-p$ and nonzero with probability p. For any p between 0 and 1 with a directional symmetry, Yalcin et al. (2024) and Zhang et al. (2025) proved that non-smooth estimators can exactly recover system dynamics for linear and parameterized nonlinear systems, respectively. Moreover, Kim and Lavaei (2025a) showed that asymmetric disturbances with nonzero probability p could be transformed into directional-symmetric disturbances with probability $2p$, enabling exact recovery for any disturbance structure where $p < 1/2$. Building on these works, Kim and Lavaei (2025c) and Kim et al. (2025) respectively demonstrated the recovery of input-output mapping of partially observed systems and parameterized nonlinear systems. These non-asymptotic analyses leverage techniques from modern high-dimensional probability and statistics (Vershynin, 2018; Wainwright, 2019). Despite their strong learning properties, these approaches rely on convex optimization solvers, which become computationally prohibitive as the system state dimension and time horizon grow. Safety-critical systems require rapid estimation to ensure real-time control without delays. Thus, it is imperative to design fast algorithms that efficiently solve non-smooth estimators within a reasonable time frame.

Due to the non-smoothness of the objective, gradient- and Hessian-based first-order and second-order algorithms are not viable in this context. Instead, we leverage the *subgradient method*, a well-established technique for non-smooth convex optimization. Unlike gradient-based methods, subgradient methods do not guarantee descent at each iteration. Instead, existing results show that the minimum sub-optimality gap asymptotically converges to zero over time with both constant and diminishing step sizes. For a comprehensive overview of subgradient methods, see Polyak (1987); Bertsekas (1997); Nesterov (2009); Shor (2012); Nesterov (2013); Kim (2025). Closed-form solutions and gradient-based methods are widely used in system identification problems, as the literature has primarily focused on LSE or other smooth estimators for linear systems (Keesman, 2011). Nonlinear system identification is inherently more complex than its linear counterpart (Schoukens and Ljung, 2019), yet existing estimation techniques still rely on smooth penalty functions. In contrast, subgradient methods have not been explored in the system identification context. Thus, this paper is the first to analyze the subgradient algorithm for the system identification problem. It is important to note that we do not apply the subgradient method to a single optimization problem but instead to a sequence of time-varying problems, since the optimization problem to be solved evolves as new measurements are sequentially taken. Within this framework, ***our contribution is twofold***: we show that, with high probability, the algorithm iterates (1) *converge linearly* to the ground-truth system under the best and Polyak step sizes after a burn-in period, and (2) *converge sublinearly* to the ground-truth system under constant and diminishing step sizes after a burn-in period.

We note that, after the first appearance of this paper, Kim and Lavaei (2025b) extended our work to further reduce computational complexity using a stochastic subgradient method, where the expected estimation error is provably guaranteed to converge to the ground-truth system.

Notation: For a matrix Z, $\|Z\|_F$ denotes its Frobenius norm. For two matrices Z_1 and Z_2, we use $\langle Z_1, Z_2 \rangle$ to denote the inner product. For a vector z, $\|z\|_2$ denotes its ℓ_2-norm. Given two functions f and g, the notation $f(x) = \Theta(g(x))$ means that there exist universal positive constants c_1 and c_2 such that $c_1 g(x) \leq f(x) \leq c_2 g(x)$. Similarly, $f(x) = \mathcal{O}(g(x))$ implies that there exists $c_3 > 0$ such that $f(x) \leq c_3 g(x)$, and $f(x) = \Omega(g(x))$ implies that there exists $c_4 > 0$ such that $f(x) \geq c_4 g(x)$. I_n stands for the $n \times n$ identity matrix.

2. Non-smooth Estimator and Disturbance Model

We consider a linear time-invariant (LTI) dynamical system of order n with the update equation

$$x_{t+1} = \bar{A}x_t + \bar{d}_t, \quad t = 0, 1, \ldots, T-1, \tag{1}$$

where $\bar{A} \in \mathbb{R}^{n\times n}$ is the unknown system matrix and $\bar{d}_t \in \mathbb{R}^n$ are unknown system disturbances. Our goal is to estimate $\bar{A}$ using the state samples $\{x_t\}_{t=0}^T$ obtained from a single system initialization. Note we only study the autonomous case due to space restrictions and the generalization to the case with inputs is straightforward. Traditionally, the following LSE has been widely studied in the literature:

$$\min_{A\in\mathbb{R}^{n\times n}} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t\|_2^2. \tag{LSE}$$

However, the LSE is vulnerable to outliers and adversarial disturbances. In safety-critical system identification, a new popular alternative is the following non-smooth estimator:

$$\min_{A\in\mathbb{R}^{n\times n}} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t\|_2. \tag{NSE}$$

The primary objective of this non-smooth estimator is to achieve exact recovery in finite time, which is an essential feature for safety-critical applications such as power-grid monitoring, unmanned aerial vehicles, and autonomous vehicles. In these scenarios, the system should be learned from a single trajectory (since multiple initialization is often infeasible) and, moreover, it can be exposed to adversarial agents. Even under i.i.d. disturbance vectors, if $\bar{d}_t$ is non-zero at every time step t, the exact recovery of the $\bar{A}$ is not attainable in finite time. Motivated by Zhang et al. (2025), we, therefore, assume a disturbance model in which the disturbance vector $\bar{d}_t$ is zero at certain time instances.

Assumption 1 (Probabilistic sparsity model) *For each time instance t, the disturbance vector $\bar{d}_t$ is zero with probability $1-p$, where $p \in (0,1)$. Furthermore, the time instances at which the disturbances are nonzero are independent of each other (while the nonzero disturbance values could be correlated).*

It prevails the system identification literature that persitent excitation is crucial to successfully learn the system (Narendra and Annaswamy, 1987). As a result, lots of literatures in (LSE) also adopt the concept of persistent excitation and so do us. For example, Zhang et al. (2025) showed in their Theorem 3 that it is infeasible to learn the full system dynamics if the disturbance vectors take arbitrary values and are chosen from a low-dimensional subspace. To this end, we define the filtration $\mathcal{F}_t = \boldsymbol{\sigma}\{x_0, \ldots, x_t\}$ and present a persistent excitation condition, which ensures exploration of the entire state space.

Assumption 2 (Persistent Excitation) *Conditional on the past information $\mathcal{F}_t = \boldsymbol{\sigma}\{x_0, \ldots, x_t\}$ and the event that $\bar{d}_t \neq 0$, the disturbance vector satisfies*

$$\mathbb{E}[(\bar{A}x_t + \bar{d}_t)(\bar{A}x_t + \bar{d}_t)^T \mid \mathcal{F}_t,\ \bar{d}_t \neq 0] \succeq \lambda^2 I_n, \quad \forall t = 0, \ldots, T-1, \tag{2}$$

where $\lambda > 0$ is a constant.

We now impose a sufficient condition for the system stability of the LTI system below to guarantee learning under possibly large correlated disturbances.

Assumption 3 *The ground-truth $\bar{A}$ satisfies $\rho := \left\|\bar{A}\right\|_2 < 1$ (note that this assumption can easily be extended to the stability of $\bar{A}$—spectral radius being less than 1—due to Gelfand's formula).*

We also assume below that the disturbance vectors $\bar{d}_t$ are sub-Gaussian, which limits the tail behavior of $\bar{d}_t$ (note that bounded disturbances are automatically sub-Gaussian). However, they are temporally correlated since each disturbance $\bar{d}_t$ may depend on previous disturbances. This correlation better reflects real-world attacks, where an adversarial agent can inject disturbances that depend on prior injections.

Assumption 4 (Direction-restricted sub-Gaussian disturbances) *Conditional on the filtration $\mathcal{F}_t$ and the event that $\bar{d}_t \neq 0$, the attack vector $\bar{d}_t$ is defined by the product $\ell_t \hat{d}_t$, where*

1. *$\hat{d}_t$ is a zero-mean unit vector whenever $\bar{d}_t \neq 0$, for any $\mathcal{F}_t$;*
2. *ℓ_t is a finite, positive variable that the adversary can arbitrarily select;*
3. *$\bar{d}_t = \ell_t \hat{d}_t$ is sub-Gaussian with parameter σ, for any $\mathcal{F}_t$.*

Remark 1 *In Appendix A, we show that Assumptions 1, 2, 3 and 4 suffice to derive the same finite sample complexity bounds for exact recovery in Zhang et al. (2025). Note that $\hat{d}_t$ and ℓ_t implies a direction and a length of the disturbance, respectively. At time t, the adversary can select any positive ℓ_t that is not necessarily independent of $\hat{d}_t$, and even select the maximally adversarial ℓ_t based on the previous information $\mathcal{F}_t$. However, one may notice that the restriction on the direction $\hat{d}_t$ limits the behavior of disturbances. In fact, Theorem 3 in Kim and Lavaei (2025a) showed that when the probabilistic sparsity parameter $p \in (0, 1/2)$, this restrictive assumption can be removed and an arbitrary distribution $\bar{d}_t$ is approximately within the class of disturbances defined in Assumption 4. This incorporates realistic scenarios where the attack happens infrequently but the magnitude of attacks can be arbitrary.*

In recent years, non-smooth estimators for system identification gained traction (Feng and Lavaei, 2021; Feng et al., 2023; Yalcin et al., 2024; Zhang et al., 2025; Kim and Lavaei, 2025a,c; Kim et al., 2025). These studies analyzed non-smooth estimators in a non-asymptotic framework under various disturbance models. However, the aforementioned works do not propose numerical algorithms and rely on existing convex optimization solvers, which suffer from two drawbacks. The first issue is that in practice (NSE) should be repeatedly solved for $T = 1, 2, ...$ as new measurements are collected until a satisfactory model is obtained. This means that the methods in the aforementioned works require solving a convex optimization problem at each time step. In addition, the problem (NSE) at time T is translated into a second-order conic program with $T + n^2$ scalar variables and T conic inequalities. Hence, the solver performance degrades significantly as the order of the system and the time horizon grow. Since safety-critical applications demand fast and efficient identification, convex optimization solvers are inefficient for real-time computation. To address these limitations, we investigate the subgradient method for learning system parameters.

3. Subgradient Method

We let $f_T(A)$ denote the objective function of (NSE); *i.e.*, $f_T(A) = \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t\|_2$, and $\partial f_T(A)$ denote its subdifferential at a point A and a time period T. We denote a subgradient in the subdifferential by $G_{A,T} \in \partial f_T(A)$. We first define the subdifferential of the ℓ_2 norm.

Definition 2 *The subdifferential of the ℓ_2 norm is defined as* $\partial\|x\|_2 = \begin{cases} \frac{x}{\|x\|_2}, & \text{if } x \neq 0, \\ \{e : \|e\|_2 \leq 1\}, & \text{if } x = 0. \end{cases}$

The subdifferential $\partial\|x\|_2$ is a unique point when x is nonzero, but it consists of all points within the unit ball when $x = 0$. Using the previous definition, we obtain the subdifferential of the objective function, $\partial f_T(A)$.

Lemma 3 *The subdifferential of $f_T(\cdot)$ at a point A is* $\partial f_T(A) = \Big\{ -\sum_{t=0}^{T-1} \partial\|x_{t+1} - Ax_t\|_2 \cdot x_t^T \Big\}$, *and the subdifferential of the $f_T(\cdot)$ at the point $\bar{A}$ is*

$$\partial f_T(\bar{A}) = \left\{ -\sum_{t\in\mathcal{K}} \hat{d}_t x_t^T - \sum_{t\in\mathcal{K}^c} e_t x_t^T : \|e_t\|_2 \leq 1, \forall t \in \mathcal{K}^c \right\},$$

where $\mathcal{K} := \{t \in \{0,\dots,T-1\} \mid \bar{d}_t \neq 0\}$ is the set of time indices of nonzero disturbances, $\mathcal{K}^c := \{0,\dots,T-1\} \setminus \mathcal{K}$ is the complement of $\mathcal{K}$, and $\hat{d}_t := \bar{d}_t/\|\bar{d}_t\|_2, \forall t \in \mathcal{K}$, are the normalized disturbance directions.

The proof of Lemma 3 simply follows from a property of the Minkowski sum over convex sets.

To find the matrix $\bar{A}$, we propose Algorithm 1 that relies on the subgradient method. In Step (i), the subdifferential of the objective function at the current estimate is calculated. Since this subdifferential possibly contains multiple values, we choose a subgradient from the subdifferential in Step (ii). After that, a step size is selected in Step (iii). In Step (iv), a subgradient update is performed. Lastly, we update the objective function with the new data point in Step (v). The main feature of this algorithm is that we do not apply the subgradient algorithm to solve the current optimization problem (NSE) and instead we take one iteration to update our estimate of the matrix value at time T. Then, after a new measurement is taken, the optimization problem at time $T+1$ changes since an additional term is added to its objective function. Thus, as we update our estimates, the optimization problems also change over time. We analyze the convergence of Algorithm 1 under different step size selection rules and show that although the objective function changes at each iteration, our algorithm converges to the $\bar{A}$. In particular, Subsections 3.1 and 3.2 will establish the *linear convergence* achieved with the best and Polyak step sizes, while Subsection 3.3 will explore the polynomial rates of *minimal-information approaches*: constant and diminishing step sizes.

3.1. Subgradient Method with Best Step Size

We analyze the subgradient method using the best step size, which is defined as the step size that maximizes improvement at each iteration. However, computing this step size requires knowledge of the ground-truth matrix $\bar{A}$, making it impractical in real-world applications. Nonetheless, studying this case offers valuable insights into the behavior of subgradient updates. Let the step size in Algorithm 1 be chosen as

$$\beta^{(T)} = \frac{\langle G_{\hat{A}^{(T)},T}, \hat{A}^{(T)} - \bar{A} \rangle}{\|G_{\hat{A}^{(T)},T}\|_F^2}.$$

Algorithm 1 Subgradient Method for (NSE)

Let $f_1(A) = \|x_1 - Ax_0\|_2$. Initialize $\hat{A}^{(1)}$ randomly.
for $T = 1, 2, \ldots$ **do**
(i) Find $\partial f_T(\hat{A}^{(T)}) = -\sum_{t=0}^{T-1} \partial \|(\bar{A} - \hat{A}^{(T)})x_t + \bar{d}_t\|_2 \cdot x_t^T$.
(ii) Choose $G_{\hat{A}^{(T)},T}$ from $\partial f_T(\hat{A}^{(T)})$.
(iii) Choose $\beta^{(T)}$ using an appropriate step size rule.
(iv) Update $\hat{A}^{(T+1)} = \hat{A}^{(T)} - \beta^{(T)} G_{\hat{A}^{(T)},T}$.
(v) Update $f_{T+1}(A) = f_T(A) + \|x_{T+1} - Ax_T\|_2$.
end for
Return $\hat{A}^{(2)}, \hat{A}^{(3)}, \hat{A}^{(4)}, \ldots$

It is shown in Appendix B that this is the best step size possible. According to Theorem 4, the distance to the ground-truth matrix decreases by a factor of $\sqrt{1 - \cos^2(\hat{\theta}_T)}$ at each update with the best step size. As long as this angle is not $\pm 90^\circ$, Algorithm 1 makes progress toward $\bar{A}$.

Theorem 4 *Under the best step size for Algorithm 1, it holds that*

$$\|\hat{A}^{(T+1)} - \bar{A}\|_F \leq \sqrt{1 - \cos^2(\hat{\theta}_T)} \|\hat{A}^{(T)} - \bar{A}\|_F, \quad \forall T \geq 1,$$

where $\hat{\theta}_T$ denotes the angle between $\hat{A}^{(T)} - \bar{A}$ and $G_{\hat{A}^{(T)},T}$.

If $\hat{\theta}_T$ is obtuse, the step size will be negative, which must be avoided. To this end, we establish a positive lower bound on $\cos(\hat{\theta}_T)$ in Theorem 6 and show that this angle remains acute. To guarantee that, we define the burn-in period as the number of samples required for a theoretical exact recovery. With high probability, $\bar{A}$ turns out to be the unique optimal solution of $\min_A f_T(A)$ for all time steps after the burn-in period.

Definition 5 *Consider an arbitrary $\delta \in (0, 1)$ and constants p, λ, ρ, σ each from Assumptions 1, 2, 3 and 4. Let $\kappa := \sigma/\lambda$. Let $T^{(burn)}$ be defined as*

$$T^{(burn)} := \Theta\left(\max\left\{ \frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)} \right\} \cdot \left[n^2 \log\left(\frac{\kappa}{p(1-p)(1-\rho)} \right) + \log\left(\frac{1}{\delta} \right) \right] \right).$$

Then, with probability at least $1 - \delta$, the ground-truth matrix $\bar{A}$ is the unique optimal solution to (NSE) for all $T \geq T^{(burn)}$ (see Appendix A for details).

We analyze the algorithm's behavior after the burn-in period because it happens that the iterates of Algorithm 1 do not move toward $\bar{A}$ when T is not large enough or equivalently when $\bar{A}$ is not the unique solution of (NSE). The following theorem establishes that if the number of samples exceeds $T^{(burn)}$, the cosine of the angle can be bounded below by a positive term.

Theorem 6 *Define $\hat{\theta}_T$ to be the angle between $\hat{A}^{(T)} - \bar{A}$ and $G_{\hat{A}^{(T)},T}$. Suppose that Assumptions 1, 2, 3 and 4 hold. Then, with probability at least $1 - \delta$, the term $\cos(\hat{\theta}_T)$ can be bounded below by*

$$\cos(\hat{\theta}_T) = \Omega\left(\frac{p(1-p)(1-\rho)}{\kappa^5} \right) > 0.$$

for all $T \geq T^{(burn)}$ such that $\hat{A}^{(T)} \neq \bar{A}$.

The analysis of Theorem 6 exploits the sharpness of $f_T(\cdot)$ with respect to $\bar{A}$ when $T \geq T^{(burn)}$; *i.e.*, $\exists \bar{c} > 0$ such that $f_T(\hat{A}^{(T)}) - f_T(\bar{A}) \geq \bar{c}\|\hat{A}^{(T)} - \bar{A}\|_F$ (see Appendix C for details). Theorem 6 provides a probabilistic lower bound on the cosine of the angle of interest when the sample complexity T is sufficiently large. Together with Theorem 4, it can be readily shown that the algorithm iterates converge toward the ground-truth $\bar{A}$ after a sufficiently large time $T^{(burn)}$.

Corollary 7 *Suppose that Assumptions 1, 2, 3 and 4 hold. Let $D = \|\hat{A}^{(T^{(burn)})} - \bar{A}\|_F$ denote the distance of the solution of the last iterate to the ground-truth $\bar{A}$ at the end of the burn-in time. Define γ to be*

$$\gamma := \sqrt{1 - \Omega\left(\frac{p^2(1-p)^2(1-\rho)^2}{\kappa^{10}}\right)}.$$

Then, given $\delta \in (0,1)$ and an arbitrary $\epsilon > 0$, with probability at least $1-\delta$, Algorithm 1 under the best step size achieves $\|\hat{A}^{(T)} - \bar{A}\|_F \leq \epsilon$ when the number of iterations satisfies $T \geq T^{(burn)} + T^{(conv)}(\epsilon)$, where

$$T^{(conv)}(\epsilon) := \frac{\log(D/\epsilon)}{-\log(\gamma)}.$$

The proof simply follows from identifying T such that $D \cdot \gamma^T \leq \epsilon$. Corollary 7 implies that to achieve an ϵ-level estimation error, the total number of required samples is the sum of those needed for the burn-in phase $T^{(burn)}$ and the linear convergence phase $T^{(conv)}(\epsilon)$. With high probability, the burn-in period scales as $\max\{(1-p)^{-2}, p^{-1}(1-p)^{-1}\}$ with respect to p and as n^2 with respect to the system dimension. After the burn-in phase, the algorithm converges to the ground-truth system at a linear rate because $\cos(\hat{\theta}_T)$ is bounded below by a positive factor, and thus the linear convergence rate is bounded above by $\gamma < 1$. This ensures that the sample complexity scales logarithmically as $\log(D/\epsilon)$ in the convergence phase. In the next subsections, we now generalize the results to more practical step size rules.

3.2. Subgradient Method with Polyak Step Size

In this subsection, we analyze the Polyak step size. Unlike the best step size selection, the Polyak step size does not require knowledge of $\bar{A}$. However, it requires information about the optimal objective value at $\bar{A}$ at time T; *i.e.*, $f_T(\bar{A})$. Under this step size rule, we can establish the result similar to Corollary 7.

Theorem 8 *Suppose that Assumptions 1, 2, 3 and 4 hold. Let $\delta \in (0,1)$ and $\epsilon > 0$ be arbitrary. Consider Algorithm 1 under the Polyak step size defined as*

$$\beta^{(T)} = \frac{f_T(\hat{A}^{(T)}) - f_T(\bar{A})}{\|G_{\hat{A}^{(T)},T}\|_F^2}.$$

With probability at least $1-\delta$, Algorithm 1 under the Polyak step size achieves $\|\hat{A}^{(T)} - \bar{A}\|_F \leq \epsilon$ when $T \geq T^{(burn)} + T^{(conv)}(\epsilon)$, where $T^{(conv)}(\epsilon)$ is defined in Corollary 7.

The proof of Theorem 8 can be found in Appendix D. Although the Polyak step size relies on less information about the ground-truth system dynamics than the best step size (*i.e.*, it needs to know an optimal scalar rather than an optimal matrix), it requires the same number of iterations as the best step size selection. These step sizes are implementable in practice (subject to some approximations) if some prior knowledge about the system dynamics is available.

3.3. Subgradient Method with Constant and Diminishing Step Sizes

The step sizes discussed in previous subsections require knowledge of the ground-truth matrix $\bar{A}$ or the optimal objective value $f_T(\bar{A})$. While theoretically useful, these step sizes are impractical in real-world settings. We therefore explore the subgradient method with constant and diminishing step sizes, which are broadly applicable. Unlike the best and Polyak step sizes, constant and diminishing step sizes do not guarantee descent at every iteration. Instead, we analyze the *minimum solution gap*, $\min_{T\in[T^{(burn)},\bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F$, and show that it converges to zero when $\bar{T}$ is sufficiently large.

Now, we demonstrate that under a constant step size, the minimum solution gap of the subgradient algorithm polynomially converges at a rate of $\bar{T}^{-1/2}$, thereby showing sublinear convergence. Specifically, Theorem 9 states that when $\bar{T}$ is significantly larger than $T^{(burn)}$, the minimum solution gap can be simplified as $\mathcal{O}\big(\frac{D\kappa^5}{p(1-p)(1-\rho)\bar{T}^{1/2}}\big)$.

Theorem 9 *Suppose that Assumptions 1, 2, 3 and 4 hold. Let $D = \|\hat{A}^{(T^{(burn)})} - \bar{A}\|_F$, and fix a sufficiently large $\bar{T}$. Consider Algorithm 1 with the step size $\beta^{(T)}$ set to the constant value*

$$\beta^{(T)} = \beta = \Theta\left(\frac{D(1-\rho)}{\sigma(\sum_{t=T^{(burn)}}^{\bar{T}} t^2)^{1/2}}\right), \quad \forall T \geq T^{(burn)}.$$

Then, with probability at least $1-\delta$, the minimum solution gap at time $\bar{T}$ is bounded as

$$\min_{T\in[T^{(burn)},\bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F = \mathcal{O}\left(\frac{D\kappa^5}{p(1-p)(1-\rho)} \cdot \frac{(\sum_{t=T^{(burn)}}^{\bar{T}} t^2)^{1/2}}{\sum_{t=T^{(burn)}}^{\bar{T}} t}\right).$$

Furthermore, we establish a similar result for the diminishing step size, which is chosen to satisfy $\sum_T \beta^{(T)} \to \infty$ and $\sum_T (\beta^{(T)})^2 < \infty$, ensuring sufficient exploration and stable convergence without excessive oscillations. Theorem 10 establishes that, under a diminishing step size, the minimum solution gap polynomially converges at a rate of $\bar{T}^{-1/2}$, which also yields sublinear convergence.

Theorem 10 *Under the same setting as Theorem 9, suppose that the step size $\beta^{(T)}$ takes the diminishing form $\beta^{(T)} = \beta/T$ for all $T \geq T^{(burn)}$, where β is defined as*

$$\beta := \Theta\left(\frac{D(1-\rho)}{\sigma(\bar{T} - T^{(burn)})^{1/2}}\right).$$

Then, with probability at least $1-\delta$, the minimum solution gap at time $\bar{T}$ is bounded as

$$\min_{T\in[T^{(burn)},\bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F = \mathcal{O}\left(\frac{D\kappa^5}{p(1-p)(1-\rho)} \cdot \frac{1}{(\bar{T} - T^{(burn)})^{1/2}}\right).$$

The analysis leverages the sharpness of $f_T(\cdot)$ used in Theorem 6 (see Appendix E for details).

4. Time Complexity Analysis

We compare the time complexities of the subgradient method and the exact solution of (NSE) using CVX. For the subgradient algorithm, each time period involves computing the subgradient, which has a per-iteration cost of $\mathcal{O}(Tn^2)$. Running the algorithm for T iterations results in an overall

complexity of $\mathcal{O}(T^2n^2)$. On the other hand, solving (NSE) using CVX reduces the problem to a second-order conic program (SOCP) with $T + n^2$ variables with T second-order cone constraints (minimization of the sum of norms can be written as a SOCP (Alizadeh and Goldfarb, 2003)). Defining $[T] = \{0, 1, \ldots, T-1\}$, this problem is formulated as

$$\min_{\substack{\text{vec}(A)\in\mathbb{R}^{n^2} \\ f_t\in\mathbb{R},\ t\in[T]}} \quad \sum_{t=0}^{T} f_t \quad \text{s.t.} \quad f_t \geq \|x_{t+1} - (x_t \otimes I_n)\,\text{vec}(A)\|_2,\ t \in [T], \tag{SOCP}$$

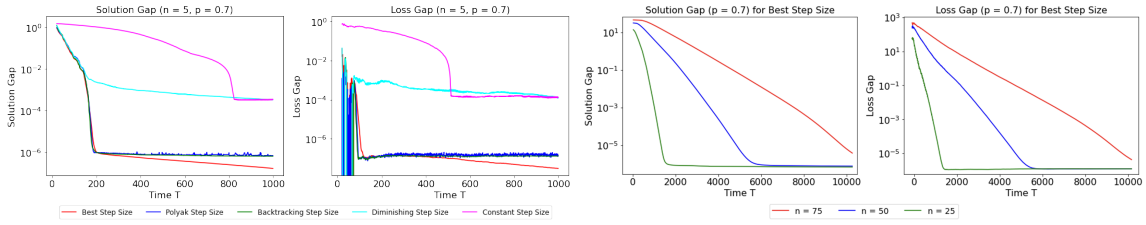
where $\otimes$ denotes the Kronecker product and $\text{vec}(A)$ is the vectorization of A obtained by stacking its columns into a single vector. The CVX solver employs primal-dual interior point methods to solve (SOCP). The number of iterations required to reach an ϵ-accurate solution is $\mathcal{O}(T^{1/2}\log(1/\epsilon))$ (Wright, 1997). Since solving conic constraints has a worst-case complexity of $\mathcal{O}(n^6)$, the per time period complexity is $\mathcal{O}(T^{1/2}n^6)$, leading to an overall complexity of $\mathcal{O}(T^{3/2}n^6)$. We know that T and $T^{(burn)}$ scale as $\Theta(n^2)$. Consequently, the per-time period complexity of the subgradient method and the exact SOCP solution are $\mathcal{O}(n^4)$ and $\mathcal{O}(n^7)$, respectively. This highlights that solving (SOCP) exactly requires significantly more computational power than the subgradient approach.

5. Numerical Experiments

In this section, we present numerical experiments to validate the convergence results. We generate a system with ground-truth matrix $\bar{A} \in \mathbb{R}^{n\times n}$ whose singular values are uniform in $(0, 1)$. Given the probability p, disturbance vectors are set to zero with probability $1 - p$. Then, we sample $\ell_t \sim \mathcal{N}(0, \sigma_t^2)$, where $\sigma_t^2 := \min\{\|x_t\|_2^2, 1/n\}$, and sample $\hat{d}_t$ from a uniform distribution over a unit sphere. When an attack occurs, the disturbance is set to $\bar{d}_t := |\ell_t|\hat{d}_t$, which conforms to Assumption 4 while the disturbance values are still correlated over time. We generate 10 independent trajectories of the subgradient algorithm using 10 different systems. Then, we report the average solution gap and the average loss gap of Algorithm 1 at each period. The *solution gap* and the *loss gap* for Algorithm 1 at time T are $\|\hat{A}^{(T)} - \bar{A}\|_F$ and $f_T(\hat{A}^{(T)}) - f_T(\bar{A})$, respectively.

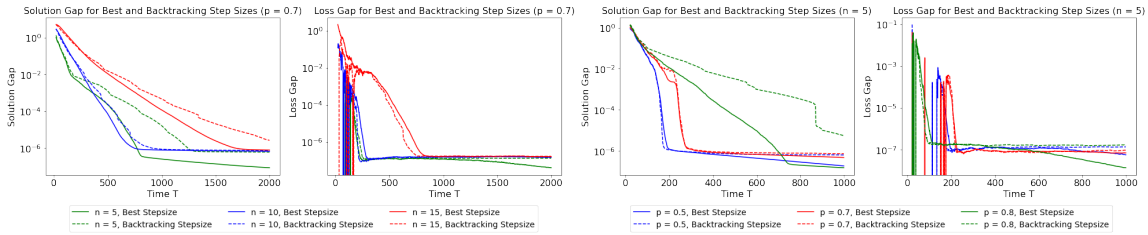
Algorithm 1 is experimented with different step size rules. In addition to the best, Polyak, diminishing, and constant step sizes, we also introduce the backtracking (Armijo) step size, where the initial step size is repeatedly multiplied by a factor $\alpha \in (0, 1)$ until we achieve a sufficient descent (see Armijo (1966)). Figure 1(*a*) depicts the performance of Algorithm 1 under these different step sizes for $n = 5$ and $p = 0.7$. Best and Polyak step sizes achieve fast and linear convergence as proven theoretically. Moreover, despite slower rates, the diminishing and constant step sizes converge to a neighborhood of the ground-truth system A. Interestingly, the backtracking step size performs similarly to the best and Polyak step sizes even without knowledge of $\bar{A}$ or $f_T(\bar{A})$. Thus, the theoretical analysis of the backtracking step size would be a promising future direction.

We test the performance of Algorithm 1 with different n and p. We choose backtracking and best step sizes to test the impact of the system dimension, $n \in \{5, 10, 15\}$, on the solution gap and the loss gap when $p = 0.7$. Figure 2(*a*) shows that as n grows, the number of samples needed for convergence grows as n^2. This aligns with the theoretical results from earlier sections. Moreover, the backtracking step size performs similarly to the best step size. Likewise, Figure 2(*b*) depicts the performance of the subgradient algorithm for different probabilities of nonzero disturbance $p \in \{0.5, 0.7, 0.8\}$ with $n = 5$. As p increases, the convergence for the estimator needs more samples, as supported by the theoretical results.



(*a*) Different Step Sizes for $n = 5$, and $p = 0.7$ (*b*) Best Step Size, $p = 0.7$, $n \in \{50, 75, 100\}$

Figure 1: Solution Gap and Loss Gap of Algorithm 1 with Different Step Sizes and Dimensions n.



(*a*) $p = 0.7,\ n \in \{5, 10, 15\}$ (*b*) $n = 5,\ p \in \{0.5, 0.7, 0.8\}$

Figure 2: Solution Gap and Loss Gap of Algorithm 1 with Best and Backtracking Step Sizes, Different Dimensions n and Probabilistic Sparsity Parameters p.

Last but not least, we provide simulation results using the best step size for large systems of order $n \in \{25, 50, 75\}$. We run these experiments for $T = 10000$. This is not viable with the standard CVX solver due to the high computational complexity of (SOCP). Nevertheless, the subgradient algorithm provides an accurate estimation for large systems within a reasonable time frame, as seen in Figure 1(*b*).

6. Conclusion

We propose and analyze a subgradient algorithm for non-smooth system identification of linear-time invariant systems with disturbances following a probabilistic sparsity model. While the non-asymptotic theory for this setting is well-established in the literature, existing methods rely on solvers that use primal-dual interior-point methods for second-order conic programs, leading to high computational costs. To address this, we propose a subgradient method that enables fast exact recovery. We demonstrate that the subgradient method with the best and Polyak step sizes converges linearly to the ground-truth system dynamics after a burn-in period. Moreover, a greater practical approach is demonstrated by our finding that under both constant and diminishing step sizes, the minimum solution gap converges to zero at a polynomial rate. These theoretical results are supported by numerical simulations. Our experiments also highlight the practical effectiveness of the backtracking step size, suggesting it as a promising alternative. However, its theoretical properties remain an open question. Moreover, our current implementation computes the full subgradient at each iteration. To further improve time complexity, exploring the convergence behavior of partial subgradient approaches, such as mini-batch or stochastic gradient descent, within the system identification context would be an important theoretical direction.

Acknowledgments

This work was supported by the U. S. Army Research Laboratory and the U. S. Army Research Office under Grant W911NF2010219, Office of Naval Research under Grant N000142412673, AFOSR, NSF, and the UC Noyce Initiative.

References

Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical Programming*, 95(1):3–51, 2003.

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Han-Fu Chen and Lei Guo. *Identification and stochastic adaptive control*. Springer Science & Business Media, 2012.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4): 633–679, 2020. doi: 10.1007/s10208-019-09426-y.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Han Feng and Javad Lavaei. Learning of dynamical systems under adversarial attacks. In *IEEE Conference on Decision and Control (CDC)*, pages 3010–3017, 2021.

Han Feng, Baturalp Yalcin, and Javad Lavaei. Learning of dynamical systems under adversarial attacks-null space property perspective. In *American Control Conference (ACC)*, pages 4179–4184, 2023.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861, 2020.

Karel J Keesman. *System identification: an introduction*. Springer Science & Business Media, 2011.

Jihun Kim. Revisiting the geometrically decaying step size: Linear convergence for smooth or non-smooth functions. *arXiv preprint arXiv:2508.13569*, 2025.

Jihun Kim and Javad Lavaei. Prevailing against adversarial noncentral disturbances: Exact recovery of linear systems with the l_1-norm estimator. In *American Control Conference (ACC)*, pages 1161–1168, 2025a.

Jihun Kim and Javad Lavaei. Bridging batch and streaming estimations to system identification under adversarial attacks. *arXiv preprint arXiv:2509.15794*, 2025b.

Jihun Kim and Javad Lavaei. System identification from partial observations under adversarial attacks. In *IEEE Conference on Decision and Control (CDC)*, pages 1384–1391, 2025c.

Jihun Kim, Yuchen Fang, and Javad Lavaei. On the sharp input-output analysis of nonlinear systems under adversarial attacks. *arXiv preprint arXiv:2505.11688*, 2025.

Ljung Lennart. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, 28:540, 1999.

Lennart Ljung. Convergence analysis of parametric identification methods. *IEEE Transactions on Automatic Control*, 23(5):770–783, 1978.

Kumpati S. Narendra and Anuradha M. Annaswamy. Persistent excitation in adaptive systems. *International Journal of Control*, 45(1):127–160, 1987.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., New York, 1987.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022.

Johan Schoukens and Lennart Ljung. Nonlinear system identification: A user-oriented road map. *IEEE Control Systems Magazine*, 39(6):28–99, 2019.

Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference on Learning Theory (COLT)*, pages 439–473, 2018.

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

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J. Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Stephen J Wright. *Primal-dual interior-point methods*. SIAM, 1997.

Baturalp Yalcin, Haixiang Zhang, Javad Lavaei, and Murat Arcak. Exact recovery for system identification with more corrupt data than clean data. *IEEE Open Journal of Control Systems*, 2024.

Haixiang Zhang, Baturalp Yalcin, Javad Lavaei, and Eduardo D. Sontag. Exact recovery guarantees for parameterized nonlinear system identification problem under sparse disturbances or semi-oblivious attacks. *Transactions on Machine Learning Research*, 2025.

Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *Conference on Learning Theory (COLT)*, pages 3333–3364, 2022.

Appendix

A. Exact Recovery under Disturbance-Restriction

We first state the assumption on the disturbances presented in Zhang et al. (2025).

Assumption 5 (Semi-oblivious sub-Gaussian disturbances) *Conditional on the filtration $\mathcal{F}_t$ and the event that $\bar{d}_t \neq 0$, the attack vector $\bar{d}_t$ is defined by the product $\ell_t \hat{d}_t$, where*

1. *$\hat{d}_t$ is a zero-mean unit vector whenever $\bar{d}_t \neq 0$, for any $\mathcal{F}_t$;*
2. *ℓ_t is a zero-mean and sub-Gaussian with parameter σ;*
3. *ℓ_t and $\hat{d}_t$ are independent whenever $\bar{d}_t \neq 0$, for any $\mathcal{F}_t$.*

Consider $T^{(burn)}$ in Definition 5. Zhang et al. (2025) showed that under Assumptions 1, 2, 3 and 5, if $T \geq T^{(burn)}$, then $\bar{A}$ is the unique solution to the estimator $\min_A \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t\|_2$ with probability at least $1-\delta$ (consider $L=1$). In this section, we will show that the estimator recovers $\bar{A}$ for *all* sufficiently large T with probability at least $1-\delta$.

Lemma 11 *Suppose that Assumptions 1, 2, 3 and 4 hold. Consider $T^{(burn)}$ in Definition 5 and $\delta \in (0,1)$. Then, with probability at least $1-\delta$, $\bar{A}$ is the unique solution to the estimator $\min_A \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t\|_2$ for all $T \geq T^{(burn)}$.*

Proof *We first note that the claim in equations (57)-(58) of Zhang et al. (2025) is also valid under Assumptions 1, 2, 3 and 4, with Assumption 5 replaced by 4. Zhang et al. (2025) relied on the independence between ℓ_t and $\hat{d}_t$ to show that the term of interest, $\sum_j \hat{d}_{k_j}^T Z g_{k_j}^{k_\ell}$ (for $j \geq \ell+1$) in equation (50), has zero mean. However, this independence is unnecessary, as the same result follows by applying the tower rule with respect to the filtration $\mathcal{F}_{k_j}$, using the fact that $\mathbb{E}[\hat{d}_{k_j} \mid \mathcal{F}_{k_j}] = 0$ under the zero-mean unit-vector assumption on $\hat{d}_t$ given $\mathcal{F}_t$. Thus, the sub-Gaussian parameter of the term $\sum_j \hat{d}_{k_j}^T Z g_{k_j}^{k_\ell}$ remains $\frac{\sigma}{1-\rho}$ (with $L=1$), which yields the same conclusion in equations (57)-(58).*

For simplicity, define $\eta := \Theta\Big(\max\Big\{\frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)}\Big\}\Big)$. Equations (57)-(58) imply that the desirable event in equation (58) at time T holds with probability at least $1-\exp(-\frac{T}{\eta})$. Then, due to the union bound, the probability that such events hold for all $T \geq T^*$ is bounded below by

$$1 - \sum_{t=T^*}^{\infty} \exp\left(-\frac{t}{\eta}\right) = 1 - \frac{e^{-T^*/\eta}}{1-e^{-1/\eta}} = 1 - \Theta\left(\eta e^{-T^*/\eta}\right), \tag{3}$$

since $1-e^{-1/\eta} = \Theta(1/\eta)$. By setting $T^* := \Theta\big(\eta \log(\frac{\eta}{\delta})\big)$, the probability (3) is lower bounded by $1-\delta$. We substitute η to present the required sample complexity T^*:

$$\Theta\left(\max\left\{\frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)}\right\} \log\left(\frac{\kappa}{\delta \cdot p(1-p)(1-\rho)}\right)\right).$$

Now, we apply the covering number argument in Zhang et al. (2025). With the ϵ-covering number of the unit sphere $N = (1+\frac{2}{\epsilon})^{n^2}$ (see Vershynin (2018); Wainwright (2019)), we require

the sample complexity of

$$\Theta\left(\max\left\{\frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)}\right\}\log\left(\frac{2N\cdot\kappa}{\delta\cdot p(1-p)(1-\rho)}\right)\right)$$
$$= \Theta\left(\max\left\{\frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)}\right\}\cdot\left[\log(2N)+\log\left(\frac{\kappa}{p(1-p)(1-\rho)}\right)+\log\left(\frac{1}{\delta}\right)\right]\right)$$
$$= \Theta\left(\max\left\{\frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)}\right\}\cdot\left[n^2\log\left(1+\frac{2}{\epsilon}\right)+\log\left(\frac{\kappa}{p(1-p)(1-\rho)}\right)+\log\left(\frac{1}{\delta}\right)\right]\right)$$
$$= \Theta\left(\max\left\{\frac{\kappa^{10}}{(1-p)^2(1-\rho)^3}, \frac{\kappa^4}{p(1-p)}\right\}\cdot\left[n^2\log\left(\frac{\kappa}{p(1-p)(1-\rho)}\right)+\log\left(\frac{1}{\delta}\right)\right]\right),$$

where the last equality comes from selecting $\log(\frac{1}{\epsilon})$ *to be* $\Theta\big(\log\big(\frac{\kappa}{p(1-p)(1-\rho)}\big)\big)$ *(see (62)-(64) in Zhang et al. (2025)). This completes the proof.*

B. Proof of Theorem 4

Proof *Using the subgradient update iteration at time* T, $\hat{A}^{(T+1)} = \hat{A}^{(T)} - \beta^{(T)}G_{\hat{A}^{(T)},T}$, *we have*

$$\begin{aligned}\|\hat{A}^{(T+1)} - \bar{A}\|_F^2 &= \|(\hat{A}^{(T)} - \bar{A}) - \beta^{(T)}G_{\hat{A}^{(T)},T}\|_F^2 \\ &= \|\hat{A}^{(T)} - \bar{A}\|_F^2 + (\beta^{(T)})^2\|G_{\hat{A}^{(T)},T}\|_F^2 - 2\beta^{(T)}\langle G_{\hat{A}^{(T)},T}, \hat{A}^{(T)} - \bar{A}\rangle. \quad (4)\end{aligned}$$

Note that the step size $\beta^{(T)} = \frac{\langle G_{\hat{A}^{(T)},T}, \hat{A}^{(T)}-\bar{A}\rangle}{\|G_{\hat{A}^{(T)},T}\|_F^2}$ *proposed in the theorem minimizes (4). Substituting the step size yields*

$$\|\hat{A}^{(T+1)} - \bar{A}\|_F^2 = \|\hat{A}^{(T)} - \bar{A}\|_F^2 - \frac{(\langle G_{\hat{A}^{(T)},T}, \hat{A}^{(T)} - \bar{A}\rangle)^2}{\|G_{\hat{A}^{(T)},T}\|_F^2} \le \sqrt{1-\cos^2(\hat{\theta}_T)}\|\hat{A}^{(T)} - \bar{A}\|_F.$$

The last inequality uses the fact that $\langle G_{\hat{A}^{(T)},T}, \hat{A}^{(T)} - \bar{A}\rangle = \|G_{\hat{A}^{(T)},T}\|_F\|\hat{A}^{(T)} - \bar{A}\|_F\cos(\hat{\theta}_T)$.

C. Proof of Theorem 6

Proof *Since* $f_T(\cdot)$ *is a sum of* ℓ_2 *norm functions and is therefore convex, one can leverage subgradient inequality (Boyd and Vandenberghe, 2004) to derive*

$$f_T(\bar{A}) - f_T(\hat{A}^{(T)}) \ge \langle \bar{A} - \hat{A}^{(T)}, G_{\hat{A}^{(T)},T}\rangle. \quad (5)$$

Thus, it holds that

$$\cos(\hat{\theta}_T) = \frac{\langle \hat{A}^{(T)} - \bar{A}, G_{\hat{A}^{(T)},T}\rangle}{\|\hat{A}^{(T)} - \bar{A}\|_F\|G_{\hat{A}^{(T)},T}\|_F} \ge \frac{f_T(\hat{A}^{(T)}) - f_T(\bar{A})}{\|\hat{A}^{(T)} - \bar{A}\|_F\|G_{\hat{A}^{(T)},T}\|_F}. \quad (6)$$

The subgradient inequality also implies

$$f_T(\hat{A}^{(T)}) - f_T(\bar{A}) \ge \langle \hat{A}^{(T)} - \bar{A}, G_{\bar{A},T}\rangle. \quad (7)$$

Thus, using Lemma 3, we can lower bound the difference between the objectives as

$$\begin{aligned} f_T(\hat{A}^{(T)}) - f_T(\bar{A}) &\geq \sup_{G_{\bar{A},T} \in \partial f_T(\bar{A})} \left\{ \langle \hat{A}^{(T)} - \bar{A}, G_{\bar{A},T} \rangle \right\} \\ &= \sup_{\|e_t\| \leq 1, t \in \mathcal{K}^c} \left\{ \left\langle \hat{A}^{(T)} - \bar{A}, -\sum_{t \in \mathcal{K}} \hat{d}_t \cdot x_t^T - \sum_{t \in \mathcal{K}^c} e_t \cdot x_t^T \right\rangle \right\} \\ &= \sum_{t \in \mathcal{K}^c} \|(\hat{A}^{(T)} - \bar{A}) x_t\|_2 - \sum_{t \in \mathcal{K}} \left\langle (\hat{A}^{(T)} - \bar{A}) x_t, \frac{\bar{d}_t}{\|\bar{d}_t\|_2} \right\rangle, \end{aligned} \tag{8}$$

where the term on the right-hand side being positive is exactly the exact recovery condition in Corollary 3 in Zhang et al. (2025) with $f(x_t) = x_t$ and $Z = \hat{A}^{(T)} - \bar{A}$. Thus, from (64) in Zhang et al. (2025), along with Lemma 11, we obtain the relevant lemma below.

Lemma 12 *Suppose that Assumptions 1, 2, 3 and 4 hold. Consider any given $B \in \mathbb{R}^{n \times n}$ and $\delta \in (0,1)$. Then, with probability at least $1 - \delta$, the event*

$$\sum_{t \in \mathcal{K}^c} \|(B - \bar{A}) x_t\|_2 - \sum_{t \in \mathcal{K}} \left\langle B - \bar{A}, \frac{\bar{d}_t}{\|\bar{d}_t\|_2} \right\rangle = \Omega\left(\frac{\lambda p(1-p) \|B - \bar{A}\|_F T}{\kappa^4} \right)$$

holds for all $T \geq T^{(burn)}$.

Using Lemma 12 with $B = \hat{A}^{(T)}$ and applying it to (8), we have that, with probability at least $1 - \delta$,

$$f_T(\hat{A}^{(T)}) - f_T(\bar{A}) = \Omega\left(\frac{\lambda p(1-p) \|\hat{A}^{(T)} - \bar{A}\|_F T}{\kappa^4} \right) \tag{9}$$

holds for all $T \geq T^{(burn)}$. In other words, for a sufficiently large T, $f_T(\cdot)$ has a sharpness with respect to $\bar{A}$; i.e., $\exists \bar{c} > 0$ such that $f_T(\hat{A}^{(T)}) - f_T(\bar{A}) \geq \bar{c} \|\hat{A}^{(T)} - \bar{A}\|_F$.

Next, we upper-bound the norm of the subgradient as below:

$$\begin{aligned} \|G_{\hat{A}^{(T)},T}\|_F^2 &= \Big\langle -\sum_{t=0}^{T-1} \partial \|(\bar{A} - \hat{A}^{(T)}) x_t + \bar{d}_t\|_2 \cdot x_t^T, -\sum_{t=0}^{T-1} \partial \|(\bar{A} - \hat{A}^{(T)}) x_t + \bar{d}_t\|_2 \cdot x_t^T \Big\rangle \\ &\leq \sum_{t=0}^{T-1} \sum_{m=0}^{T-1} \|x_t\|_2 \|x_m\|_2 = \left(\sum_{t=0}^{T-1} \|x_t\|_2 \right)^2. \end{aligned} \tag{10}$$

The inequality holds due to the norm of subgradients being at most 1 and the Cauchy-Schwarz inequality. In the next lemma, we upper bound $\sum_{t=0}^{T-1} \|x_t\|_2$ with high probability.

Lemma 13 *Suppose that Assumptions 3 and 4 hold and consider $\delta \in (0,1)$. Then, with probability at least $1 - \delta$, we have*

$$\sum_{t=0}^{T-1} \|x_t\|_2 = \mathcal{O}\left(\frac{\sigma T}{1 - \rho} \right) \tag{11}$$

for all $T = \Omega\big(\log(\frac{1}{\delta})\big)$.

We defer the proof of Lemma 13 to the end of this section. We now consider (9) and (11) to hold with probability at least $1-\frac{\delta}{2}$. Since this does not affect the order of the sample complexity, applying the union bound implies that both events hold for all $T \geq T^{(burn)}$ with probability at least $1-\delta$ (note that we already have $T^{(burn)} = \Omega(\log(\frac{1}{\delta}))$).

We substitute (9), (10), (11) into (6) to arrive at a lower bound on $\cos(\hat{\theta}_T)$:

$$\cos(\hat{\theta}_T) \geq \frac{f_T(\hat{A}^{(T)}) - f_T(\bar{A})}{\|\hat{A}^{(T)} - \bar{A}\|_F \|G_{\hat{A}^{(T)},T}\|_F} = \Omega\left(\frac{\lambda p(1-p)\|\hat{A}^{(T)} - \bar{A}\|_F T}{\kappa^4} \cdot \frac{1}{\|\hat{A}^{(T)} - \bar{A}\|_F \cdot \frac{\sigma}{1-\rho} T}\right)$$
$$= \Omega\left(\frac{\lambda p(1-p)(1-\rho)}{\sigma\kappa^4}\right) = \Omega\left(\frac{p(1-p)(1-\rho)}{\kappa^5}\right) > 0 \qquad (12)$$

for all $T \geq T^{(burn)}$ with probability at least $1-\delta$. This completes the proof.

Proof of Lemma 13 *Due to system dynamics (1), we have*

$$\sum_{t=0}^{T-1} \|x_t\|_2 = \sum_{t=0}^{T-1} \left\| \bar{A}^t x_0 + \sum_{i=0}^{t-1} \bar{A}^{t-1-i} \bar{d}_i \right\|_2 < \sum_{i=0}^{\infty} \rho^i \left[\|x_0\|_2 + \sum_{t=0}^{T-2} \|\bar{d}_t\|_2 \right]$$
$$= \frac{1}{1-\rho}\left[\|x_0\|_2 + \sum_{t=0}^{T-2} \|\bar{d}_t\|_2 \right] = \frac{1}{1-\rho}\left[\|x_0\|_2 + \sum_{t=0}^{T-2} \ell_t \right] \qquad (13)$$

due to the triangle inequality. Under Assumption 4, the sub-Gaussian parameter of ℓ_t is σ. The centering lemma in Vershynin (2018) implies that $\ell_t - \mathbb{E}[\ell_t]$ is a zero-mean sub-Gaussian variable with parameter $\mathcal{O}(\sigma)$. It follows that

$$\mathbb{E}\left[\exp\left(\nu \sum_{t=0}^{T-2} (\ell_t - \mathbb{E}[\ell_t])\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\nu \sum_{t=0}^{T-2} (\ell_t - \mathbb{E}[\ell_t])\right) \,\middle|\, \mathcal{F}_{T-2}\right]\right]$$
$$= \mathbb{E}\left[\exp\left(\nu \sum_{t=0}^{T-3} (\ell_t - \mathbb{E}[\ell_t])\right) \mathbb{E}\left[\exp\left(\nu(\ell_{T-2} - \mathbb{E}[\ell_{T-2}])\right) \mid \mathcal{F}_{T-2}\right]\right]$$
$$\leq \mathbb{E}\left[\exp\left(\nu \sum_{t=0}^{T-3} (\ell_t - \mathbb{E}[\ell_t])\right)\right] \exp(\nu^2 \mathcal{O}(\sigma^2))$$
$$\leq \cdots \leq \exp(\nu^2 (T-1)\mathcal{O}(\sigma^2))$$

for all $\nu \in \mathbb{R}$, which implies that $\sum_{t=0}^{T-2} \ell_t - \mathbb{E}[\ell_t]$ is a zero-mean sub-Gaussian variable with parameter $\mathcal{O}(\sqrt{T}\sigma)$, and thus $\|x_0\|_2 - \mathbb{E}[\|x_0\|_2] + \sum_{t=0}^{T-2} \ell_t - \mathbb{E}[\ell_t]$ also has a sub-Gaussian parameter of $\mathcal{O}(\sqrt{T}\sigma)$.

We can leverage Hoeffding's inequality to arrive at

$$\mathbb{P}\Big(\|x_0\|_2 - \mathbb{E}[\|x_0\|_2] + \sum_{t=0}^{T-2} \ell_t - \mathbb{E}[\ell_t] \leq \sigma T\Big) \geq 1 - \exp\Big(-\Omega\Big(\frac{\sigma^2 T^2}{\sigma^2 T}\Big)\Big) = 1 - \exp(-\Omega(T)).$$

Note that the sub-Gaussian parameter of ℓ_t being σ yields $\mathbb{E}[\|x_0\|_2] + \sum_{t=0}^{T-2} \mathbb{E}[\ell_t] = \mathcal{O}(\sigma T)$. This concludes that

$$\mathbb{P}\Big(\|x_0\|_2 + \sum_{t=0}^{T-2} \ell_t \leq \mathcal{O}(\sigma T)\Big) \geq 1 - \exp(-\Omega(T)). \qquad (14)$$

Due to the union bound, the probability that (14) holds for all $T \geq \tilde{T}$ is bounded below by

$$1 - \sum_{t=\tilde{T}}^{\infty} \exp(-cT) = 1 - \frac{e^{-c\tilde{T}}}{1-e^{-c}} \tag{15}$$

for some $c > 0$. Setting $\tilde{T} := \frac{1}{c}\log(\frac{1}{\delta(1-e^{-c})}) = \Theta\big(\log(\frac{1}{\delta})\big)$ makes (15) equal to $1-\delta$, which implies that the event in (14) holds for all $T \geq \tilde{T}$ with probability at least $1-\delta$. Finally, considering (13) completes the proof.

D. Proof of Theorem 8

Proof *We first note that $f_T(\hat{A}^{(T)}) - f_T(\bar{A}) \geq 0$ when $T \geq T^{(burn)}$ with high probability due to Lemma 11. It follows that the proposed step size in the theorem satisfies $\beta^{(T)} \geq 0$.*

Now, using the subgradient update iteration at time T, $\hat{A}^{(T+1)} = \hat{A}^{(T)} - \beta^{(T)} G_{\hat{A}^{(T)},T}$, we have

$$\begin{aligned}
\|\hat{A}^{(T+1)} - \bar{A}\|_F^2 &= \|(\hat{A}^{(T)} - \bar{A}) - \beta^{(T)} G_{\hat{A}^{(T)},T}\|_F^2 \\
&= \|(\hat{A}^{(T)} - \bar{A})\|_F^2 + (\beta^{(T)})^2 \|G_{\hat{A}^{(T)},T}\|_F^2 - 2\beta^{(T)} \langle G_{\hat{A}^{(T)},T}, \hat{A}^{(T)} - \bar{A}\rangle \\
&\leq \|(\hat{A}^{(T)} - \bar{A})\|_F^2 + (\beta^{(T)})^2 \|G_{\hat{A}^{(T)},T}\|_F^2 - 2\beta^{(T)} (f_T(\hat{A}^{(T)}) - f_T(\bar{A})),
\end{aligned} \tag{16}$$

where the last inequality leverages (5) and $\beta^{(T)} \geq 0$. Substituting $\beta^{(T)}$ yields

$$\begin{aligned}
\|\hat{A}^{(T+1)} - \bar{A}\|_F^2 &\leq \|\hat{A}^{(T)} - \bar{A}\|_F^2 - \frac{(f_T(\hat{A}^{(T)}) - f_T(\bar{A}))^2}{\|G_{\hat{A}^{(T)},T}\|_F^2} \\
&= \|\hat{A}^{(T)} - \bar{A}\|_F^2 \left(1 - \left(\frac{f_T(\hat{A}^{(T)}) - f_T(\bar{A})}{\|\hat{A}^{(T)} - \bar{A}\|_F \|G_{\hat{A}^{(T)},T}\|_F}\right)^2\right),
\end{aligned}$$

where the right-hand side can be bounded above by γ^2 with high probability when $T \geq T^{(burn)}$, which follows from (12). Thus, we have convergence to the ground-truth system with the linear rate γ since

$$\|\hat{A}^{(T+1)} - \bar{A}\|_F \leq \gamma \|\hat{A}^{(T)} - \bar{A}\|_F.$$

The rest of the arguments are the same to those in Corollary 7.

E. Proof of Theorems 9 and 10

Proof *We first define the following lemma and defer its proof to the end of this section.*

Lemma 14 *Let $D = \|\hat{A}^{(T^{(burn)})} - \bar{A}\|_F$ be the distance of the iterate to the ground-truth $\bar{A}$ at the end of the burn-in time $T^{(burn)}$. Then, for $\bar{T} \geq T^{(burn)}$, it holds that*

$$\min_{T \in [T^{(burn)}, \bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F = \mathcal{O}\left(\frac{\kappa^4 \big[D^2 + (\frac{\sigma}{1-\rho})^2 \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)} T)^2\big]}{\lambda p(1-p) \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)} T)}\right) \tag{17}$$

with probability at least $1-\delta$.

For Theorem 9, setting the step size to be $\beta^{(T)} = \beta$ for all $T \geq T^{(burn)}$ implies the term on the right-hand side to be

$$\mathcal{O}\left(\frac{\kappa^4\left[D^2 + (\frac{\sigma}{1-\rho})^2\beta^2 \sum_{t=T^{(burn)}}^{\bar{T}} t^2\right]}{\lambda p(1-p)\beta \sum_{t=T^{(burn)}}^{\bar{T}} t}\right). \tag{18}$$

Minimizing the right-hand side with respect to β results in the following constant step size:

$$\beta^{(T)} = \beta = \Theta\left(\frac{D(1-\rho)}{\sigma(\sum_{t=T^{(burn)}}^{\bar{T}} t^2)^{1/2}}\right).$$

Substituting this constant step size back into (18) provides the simplified upper bound that holds with probability at least $1-\delta$:

$$\begin{aligned}\min_{T\in[T^{(burn)},\bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F &= \mathcal{O}\left(\frac{\kappa^4}{\lambda p(1-p)} \cdot \frac{D\sigma}{1-\rho} \cdot \frac{(\sum_{t=T^{(burn)}}^{\bar{T}} t^2)^{1/2}}{\sum_{t=T^{(burn)}}^{\bar{T}} t}\right) \\ &= \mathcal{O}\left(\frac{D\kappa^5}{p(1-p)(1-\rho)} \cdot \frac{(\sum_{t=T^{(burn)}}^{\bar{T}} t^2)^{1/2}}{\sum_{t=T^{(burn)}}^{\bar{T}} t}\right).\end{aligned}$$

This completes the proof of Theorem 9. Now, to prove Theorem 10, setting the step size to be $\beta^{(T)} = \beta/T$ implies the right-hand side of (17) to be

$$\mathcal{O}\left(\frac{\kappa^4\left[D^2 + (\frac{\sigma}{1-\rho})^2\beta^2(\bar{T} - T^{(burn)})\right]}{\lambda p(1-p)\beta(\bar{T} - T^{(burn)})}\right). \tag{19}$$

Minimizing the right-hand side with respect to β results in the following constant:

$$\beta := \Theta\left(\frac{D(1-\rho)}{\sigma(\bar{T} - T^{(burn)})^{1/2}}\right).$$

Substituting this constant back into (19) ensures the following upper bound with probability at least $1-\delta$:

$$\begin{aligned}\min_{T\in[T^{(burn)},\bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F &= \mathcal{O}\left(\frac{\kappa^4}{\lambda p(1-p)} \cdot \frac{D\sigma}{1-\rho} \cdot \frac{(\bar{T} - T^{(burn)})^{1/2}}{\bar{T} - T^{(burn)}}\right) \\ &= \mathcal{O}\left(\frac{D\kappa^5}{p(1-p)(1-\rho)} \cdot \frac{1}{(\bar{T} - T^{(burn)})^{1/2}}\right).\end{aligned}$$

This completes the proof of Theorem 10.

Proof of Lemma 14 *We note that the proof hinges on the sharpness condition given in (9). Using (16) and the iterative substitution between the time periods from $T^{(burn)}$ to $\bar{T}$, we obtain*

$$\begin{aligned}&\|\hat{A}^{(\bar{T}+1)} - \bar{A}\|_F^2 \\ &\leq \|\hat{A}^{(T^{(burn)})} - \bar{A}\|_F^2 + \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)})^2 \|G_{\hat{A}^{(T)},T}\|_F^2 - 2\sum_{T=T^{(burn)}}^{\bar{T}} \beta^{(T)}(f_T(\hat{A}^{(T)}) - f_T(\bar{A})) \\ &\leq D^2 + \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)})^2 \left(\frac{\sigma T}{1-\rho}\right)^2 - 2\sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)}) \cdot \Omega\left(\frac{\lambda p(1-p)\|\hat{A}^{(T)} - \bar{A}\|_F \cdot T}{\kappa^4}\right).\end{aligned}$$

The last inequality follows from the fact that $\|G_{\hat{A}^{(T)},T}\|_F$ *is bounded above by* $\frac{\sigma T}{1-\rho}$ *considering* (10)-(11)*, and* $f_T(\hat{A}^{(T)}) - f_T(\bar{A})$ *is bounded below by* (9) *with high probability for all* $T \geq T^{(burn)}$*. Since* $\|\hat{A}^{(\bar{T}+1)} - \bar{A}\|_F^2$ *is nonnegative, we arrive at*

$$2 \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)}T) \cdot \Omega \left(\frac{\lambda p(1-p)\|\hat{A}^{(T)} - \bar{A}\|_F}{\kappa^4} \right) \leq D^2 + \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)}T)^2 \left(\frac{\sigma}{1-\rho} \right)^2,$$

where the left-hand side can further be bounded below by

$$2\Omega \left(\frac{\lambda p(1-p)}{\kappa^4} \right) \cdot \min_{T \in [T^{(burn)}, \bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F \cdot \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)}T).$$

Thus, it follows that

$$\min_{T \in [T^{(burn)}, \bar{T}]} \|\hat{A}^{(T)} - \bar{A}\|_F = \mathcal{O} \left(\frac{D^2 + (\frac{\sigma}{1-\rho})^2 \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)}T)^2}{\frac{\lambda p(1-p)}{\kappa^4} \sum_{T=T^{(burn)}}^{\bar{T}} (\beta^{(T)}T)} \right)$$

holds with high probability. This completes the proof.