Statistical Sparse Online Regression: A Diffusion Approximation Perspective

Jianqing Fan
Princeton University
jqfan@princeton.edu

Wenyan Gong
Princeton University
wenyang@princeton.edu

Chris Junchi Li
Princeton University
junchil@princeton.edu

Qiang Sun
University of Toronto
qsun@utstat.toronto.edu

Abstract

In this paper we adopt the diffusion approximation perspective to investigate Stochastic Gradient Descent (SGD) for least squares, which allows us to characterize the exact dynamics of the online regression process. As a consequence, we show that SGD achieves the optimal rate of convergence, up to a logarithmic factor. We further show SGD combined with the trajectory average achieves a faster rate, by eliminating the logarithmic factor. We extend SGD to the high dimensional setting by proposing a two-step algorithm: a burn-in step using offline learning and a refinement step using a variant of truncated stochastic gradient descent. Under appropriate assumptions, we show the proposed algorithm produces near optimal sparse estimators. Numerical experiments lend further support to our obtained theory.

1 Introduction

In this paper we focus on the following random-design linear regression model

$$y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_* + \varepsilon, \tag{1.1}$$

where $y \in \mathbb{R}$ is the response variable, $\mathbf{x} \in \mathbb{R}^p$ consists of p random predictor variables, and ε is the error term independent of $\mathbf{x}$. The target is to estimate parameter $\boldsymbol{\theta}_*$ based on the data $\big\{(\mathbf{x}_t, y_t) : t = 1, 2, \ldots, N\big\}$, which are independent and identically distributed (i.i.d.) realizations of $(\mathbf{x}, y)$. The estimation problem can be formulated as solving the following convex optimization problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad \mathbb{E}_{\mathbf{x},y}\left\{\frac{1}{2}(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})^2\right\}. \tag{1.2}$$

which is known as population risk minimization. It is well known that when $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ is nondegenerate, (1.2) admits a closed-form solution

$$\boldsymbol{\theta}_{\mathrm{prm}} = \left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}\left[y\mathbf{x}\right]. \tag{1.3}$$

Correspondingly, the Ordinary Least Square (OLS) estimator can be viewed as a *sample average approximation*

$$\boldsymbol{\theta}_{\mathrm{ols}} = \left(\frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\right)^{-1}\frac{1}{N}\sum_{t=1}^{N}y_t\mathbf{x}_t.$$

However, directly calculating the above OLS estimator has at least two shortcomings. First, the computational complexity of inverting the empirical Hessian matrix is in the order of $\mathcal{O}(p^3)$, which is costly when p is large. Second, when the number of predictors p is strictly larger than the number of samples N, the OLS estimator using the idea of sample average approximation becomes nonidentifiable.

Stochastic Gradient Descent We focus on the setting where we only have access data samples $\big\{(\mathbf{x}_t, y_t) : t = 1, 2, \ldots\big\}$ streaming in one at a time, in a sequential manner. The stochastic gradient of the population risk in (1.2) is

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{1}{2}(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})^2\right] = -(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})\mathbf{x}.$$

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

Then the online stochastic gradient descent at time t performs

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \eta(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta})\mathbf{x}_t. \tag{1.4}$$

Here we assume that $\eta > 0$ is a small stepsize possibly depending on the streaming sample size N, and the availability of a proper initial point $\boldsymbol{\theta}^{(0)}$, which will be specified later. One of our main theorems in §2 states that the obtained online estimator achieves the following rate of convergence

$$\mathbb{E}\left\|\boldsymbol{\theta}^{(N)} - \boldsymbol{\theta}_*\right\|^2 \lesssim \frac{\sigma^2}{\alpha_X} \cdot \frac{p \log N}{N},$$

where α_X is the smallest eigenvalue of $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$. After the trajectory averaging, our theory in §3 suggests that the rate of convergence of the online estimator can be further improved (cf. Proposition 3.1):

$$\mathbb{E}\left\|\bar{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}_*\right\|^2 \lesssim \frac{\sigma^2}{\alpha_X} \cdot \frac{\mathrm{tr}\,(\boldsymbol{\Sigma}_X^{-1})}{N},$$

and the average prediction risk satisfies that when $\mathbf{x}$ is drawn from $\mathcal{D}$,

$$\mathbb{E}\left|\mathbf{x}^{\mathrm{T}}(\bar{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}_*)\right|^2 \lesssim \frac{p\sigma^2}{N}.$$

In addition, we also have asymptotic normality results in the continuous-time settings. See the details in Propositions 2.5, 3.1 and Corollary 3.2. The above rate of convergence is known to be statistically optimal (Raskutti et al., 2010) and partially recovers the existing results (Ruppert, 1988; Polyak & Juditsky, 1992); see for example Eq. (3) in Polyak & Juditsky (1992) or Theorem 1 in Ruppert (1988). See also more recent works by Chen et al. (2016); Su & Zhu (2018), among many others.

Stochastic Gradient Descent with Truncation When $p \gg N$, the minimizer $\boldsymbol{\theta}_{\mathrm{ols}}$ is known to be non-identifiable. One popular way to encourage sparsity is to consider the following penalized regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad \mathbb{E}_{\mathbf{x},y}\left\{\frac{1}{2}(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})^2\right\} + \lambda\|\boldsymbol{\theta}\|_1, \tag{1.5}$$

where $\lambda > 0$ is a regularization parameter. In this paper, we propose to use the following truncated stochastic gradient descent:

$$\widetilde{\boldsymbol{\theta}}^{(t)} = \mathrm{Truncate}\left(\boldsymbol{\theta}^{(t-1)} + \eta(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t, k_0\right),$$

where the operator "Truncate" ranks the coordinates and keeps the k_0 largest coordinates in absolute values. The truncation parameter k_0 is obtained by exploiting an offline burn-in step. Our main result states as follows:

$$\mathbb{E}\left\|\bar{\boldsymbol{\theta}}^{(N)} - \boldsymbol{\theta}_*\right\|^2 \lesssim \frac{\sigma^2}{\alpha_X} \cdot \frac{s \log N}{N}.$$

In above, s is the size of support $S = \mathrm{supp}(\boldsymbol{\theta}_*)$. See more in §4.

Contributions We make three important contributions towards understanding online regression.

(i) To our best knowledge, this work is among the first that employs the differential equation tools to analyze the online linear regression problem. Our analysis can be broke down into two phases. In the first phase the iteration follows a deterministic dynamics characterized by an ordinary differential equation (ODE), and in the second phase the iteration fluctuates around the global minimizer and behaves as a multidimensional Ornstein-Uhlenbeck (O-U) process, which is the solution to a Langevin-type stochastic differential equation (SDE). Moreover, the convergence rate is characterized by the second moment of the stationary distribution of the O-U process. The accuracy of the diffusion approximation, however, has not been studied and remains an open problem: see the remark at the end of §2.

(ii) We provide a trajectory average scheme to further improve the rate of convergence. This averaging idea has been exploited in early works (Ruppert, 1988; Polyak & Juditsky, 1992). However, our analysis suggests that the averaging should start with an initialization to be $\mathcal{O}(\eta^{0.5})$-distant from the local minimizer, where η is the stepsize. Such analysis allows us to introduce the two-phase training strategy and illustrates different phenomenon from Ruppert (1988) and Polyak & Juditsky (1992) that averages from the initial time.

(iii) We propose a two-stage algorithm for high-dimensional linear regression problems. The first stage is an offline algorithm which only uses a small of the data to learn an initial estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{init}}$, and the second stage refines this

coarse estimator in an online fashion such that it can achieve the optimal rate of convergence. To the best of the author's knowledge, we are among the first to provide statistical guarantees for the proposed algorithm. See §4.

More literatures In either the optimization or the statistics literature, stochastic gradient descent has been extensively studied as a first-order stochastic approximation method for minimizing an objective function. The online stochastic gradient method minimizes the sum of a large number of component functions update the working parameter of interest using one data point at a time through a gradient-type update (Kushner & Yin, 2003; Benveniste et al., 2012; Borkar, 2008; Bertsekas & Tsitsiklis, 1989; Nedić et al., 2001; Nedic & Bertsekas, 2001; Bertsekas, 2011; Nedić, 2011; Wang & Bertsekas, 2015, 2016). It has been shown that after N samples/iterations, the average of the iterates has $\mathcal{O}(1/N)$ optimization error for strongly convex objective, and $\mathcal{O}(1/\sqrt{N})$ error for general convex objective (Rakhlin et al., 2012; Shamir & Zhang, 2013), which is known to match with minimax information lower bounds (Agarwal et al., 2012a; Nemirovskii & Yudin, 1983) and hence *optimal* for convex optimization under the stochastic first-order oracle. The diffusion approximation characterization can also be interpreted from a variational bayesian perspective (Mandt et al., 2016), which is opposite to the direction in this paper, however.

In the offline setting, algorithms for Lasso problems have been extensively studied in the last decade. For example, Efron et al. (2004) proposed the LARS algorithm for computing the whole solution path of lasso penalized regression problems. Agarwal et al. (2012b) proposed the proximal gradient algorithm for solving convex regularization problems with fixed λ. Later, algorithms for nonconvex penalized regression problems have been proposed (Fan et al., 2017), among others. For sparse online learning problems, there has been quite some work in the optimization literature, where researchers study the iteration complexity of various online learning algorithms, see Bertsekas (2011); Duchi et al. (2011); Xiao (2010) among others. The idea of truncation, also known as *soft-thresholding operator* has been used for sparse regression. See e.g. Donoho (1995); Daubechies et al. (2004); Langford et al. (2009) and recently by Ma (2013); Yuan & Zhang (2013). In contrast, there is little literature on the statistical recovery properties for online algorithms until very recently, where Dieuleveut et al. (2016) analyzed an accelerated version of the online regression problem in the low-dimensional setting. However, for matching statistical rates purposes, such acceleration does *not* improve the $\mathcal{O}(\sqrt{p/N})$ lower bound.

The rest of this paper is organized as follows. §2 describes the continuum diffusion approximation method and provide explicit traverse time estimation in the low dimension regime. §3 analyzes averaged iterations along with finite-sample bounds. §4 examines a two-step online truncated regression and provide the finite-sample bound under certain assumptions. §5 provides some numerical experiments for simulated data that are consistent with our theory. We conclude the paper in §6. For the clarity in presentation, all proofs are deferred to Appendix.

2 Continuum Framework and Phase Transition

In this section, we will analyze the SGD algorithm (1.4) in the low dimensional ($p \ll N$) regime. The key tools are the diffusion approximation techniques and the goal is to obtain a finite-sample bound, matching that for the offline method.

We start with regularity assumptions.

Assumption 2.1. Suppose that $(\mathbf{x}, y)$ satisfy the following statements

(i) (Exogeneity) $\mathbb{E}[\mathbf{x}] = 0$, $\mathbb{E}[\varepsilon \mid \mathbf{x}] = 0$;

(ii) (Homogeneity of variance) $\mathrm{var}[\varepsilon \mid \mathbf{x}] = \sigma^2$ for all $\mathbf{x}$;

(iii) (Absence of multicollinearity) $\mathbf{\Sigma}_X \equiv \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ has its smallest eigenvalue $\alpha_X > 0$;

(iv) (Normalized predictors) the predictor variables are normalized with $\mathrm{var}(x_i) = 1$, where $\mathbf{\Sigma}_X$ has all diagonals being 1;

(v) (Subgaussian noise) $\varepsilon \sim \text{sub-Gaussian}(0, \sigma^2)$.

Let us consider the SGD iterates (1.4) and let $\mathcal{F}_t$ be the filtration generated by the first t samples $\{(\mathbf{x}_i, y_i), i \le t\}$. Our first lemma concerns the conditional expectation of the stochastic gradients.

Lemma 2.2. Under Assumption 2.1, the stochastic gradient at step t satisfies

$$\mathbb{E}\left[-\left(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)}\right)\mathbf{x}_t \mid \mathcal{F}_{t-1}\right] = \boldsymbol{\Sigma}_X(\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}_*). \tag{2.1}$$

The noise generated at iteration $t \geq 1$ is the increment subtracting its expectation:

$$\boldsymbol{e}_t \equiv \left(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)}\right)\mathbf{x}_t + \boldsymbol{\Sigma}_X(\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}_*). \tag{2.2}$$

From (2.1), we have $\mathbb{E}[\boldsymbol{e}_t \mid \mathcal{F}_{t-1}] = 0$, and hence $\boldsymbol{e}_t$ forms a martingale difference sequence with respect to $\mathcal{F}_t$. The stochastic gradient update (1.4) can then be expressed as

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \eta\boldsymbol{\Sigma}_X(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(t-1)}) + \eta\boldsymbol{e}_t. \tag{2.3}$$

Heuristically, (1.4) is approximated by the following ordinary differential equation

$$\frac{\mathrm{d}\boldsymbol{\Theta}(t)}{\mathrm{d}t} = -\eta\boldsymbol{\Sigma}_X(\boldsymbol{\Theta}(t) - \boldsymbol{\theta}_*). \tag{2.4}$$

The following lemma computes the conditional variance of the increment at time t.

Lemma 2.3. Under Assumption 2.1, we have that, when $\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_* + \mathcal{O}(\eta^{0.5})$

$$\begin{aligned}\mathbb{E}[\boldsymbol{e}_t\boldsymbol{e}_t^{\mathrm{T}} \mid \mathcal{F}_{t-1}] &= \mathrm{var}\left[\left(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)}\right)\mathbf{x}_t \mid \mathcal{F}_{t-1}\right] \\ &= \sigma^2\boldsymbol{\Sigma}_X + \mathcal{O}(\eta^{0.5}).\end{aligned}$$

Lemmas 2.2 and 2.3 allow us to conclude that, locally around $\boldsymbol{\theta}_*$, the following continuous-time dynamics serve as a good approximation for (2.3)

$$\mathrm{d}\boldsymbol{\Theta}(t) = -\eta\boldsymbol{\Sigma}_X\left(\boldsymbol{\Theta}(t) - \boldsymbol{\theta}_*\right)\mathrm{d}t + \eta\sigma\boldsymbol{\Sigma}_X^{1/2}\,\mathrm{d}\boldsymbol{W}(t), \tag{2.5}$$

where $\boldsymbol{W}(t)$ is the standard p-dimensional Brownian motion, and $\boldsymbol{\Sigma}_X^{1/2}$ denotes the unique positive semidefinite square root of $\boldsymbol{\Sigma}_X$.

2.1 Weak Convergence Theory and Traverse Time Estimates

We utilize the diffusion theory for Markov processes to derive the weak convergence result of the solutions to the ODE and SDE in Theorem 2.4. The derivation requires applications of Lemmas 2.2 and 2.3 under appropriate temporal and spatial scalings. To avoid possible ambiguity, we occasionally add a superscript η to the iterate $\boldsymbol{\theta}^{(t)}$ such that $\boldsymbol{\theta}^{\eta,(t)} \equiv \boldsymbol{\theta}^{(t)}$.

Theorem 2.4. Suppose that Assumption 2.1 holds. Then the following results hold:

(i) As $\eta \to 0^+$, if $\boldsymbol{\theta}^{\eta,(0)}$ converges weakly to some $\boldsymbol{\Theta}_0^a$, then the stochastic process $\boldsymbol{\theta}^{\eta,(\lfloor t\eta^{-1}\rfloor)}$ converges weakly to the solution to the following ODE

$$\frac{\mathrm{d}\boldsymbol{\Theta}^a(t)}{\mathrm{d}t} = -\boldsymbol{\Sigma}_X(\boldsymbol{\Theta}^a(t) - \boldsymbol{\theta}_*), \tag{2.6}$$

which is a time-rescaled version of the ODE (2.4).

(ii) As $\eta \to 0^+$, if $\eta^{-0.5}(\boldsymbol{\theta}^{\eta,(0)} - \boldsymbol{\theta}_*)$ converges weakly to some $\boldsymbol{\Theta}_0^b$ then $\eta^{-0.5}(\boldsymbol{\theta}^{\eta,(\lfloor t\eta^{-1}\rfloor)} - \boldsymbol{\theta}_*)$ converges weakly to the solution to the following SDE

$$\mathrm{d}\boldsymbol{\Theta}^b(t) = -\boldsymbol{\Sigma}_X\boldsymbol{\Theta}^b(t)\,\mathrm{d}t + \sigma\boldsymbol{\Sigma}_X^{1/2}\,\mathrm{d}\boldsymbol{W}(t), \tag{2.7}$$

which is a time-and-space-rescaled version of SDE (2.5).

In general, weak convergence for stochastic processes, $\boldsymbol{X}^\eta(t) \Rightarrow \boldsymbol{X}(t)$ in $\mathbb{R}^d$, is characterized by the following convergence for cylindrical sets: for each coordinate $k = 1, \ldots, d$ and any $0 \leq t_1 < \cdots < t_n < \infty$ the convergence in distribution holds as $\eta \to 0^+$ (Kushner & Yin, 2003):

$$\left(X_k^\eta(t_1), \ldots, X_k^\eta(t_n)\right) \Rightarrow \left(X_k(t_1), \ldots, X_k(t_n)\right).$$

Theorem 2.4 indicates that (2.4) and (2.5) serve as good approximations to the dynamics of the SGD updates when exploiting a rescaling argument: (2.4) is deterministic and serves as a global approximation, and (2.5) is local with the information of noise encoded. From the viewpoint of statisticians, the ODE approximation corresponds to rate of convergence, while the SDE dictates asymptotic normality. In other words, (2.5) serves as an asymptotic expansion of higher order term with error encoded. The solution to (2.5) is known to be a multivariate Ornstein-Ulenbeck process:

$$\begin{aligned}\boldsymbol{\Theta}(t) = \boldsymbol{\theta}_* &+ (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)\exp\left(-\eta t\boldsymbol{\Sigma}_X\right) \\ &+ \eta\sigma\int_0^t \exp\left(\eta(\tau - t)\boldsymbol{\Sigma}_X\right)\boldsymbol{\Sigma}_X^{1/2}\,\mathrm{d}\boldsymbol{W}(\tau).\end{aligned} \tag{2.8}$$

The last term above is a manifestation of Itô integral

that has mean zero and second moment:

$$\begin{aligned}&\operatorname{var}\left(\eta\sigma\int_0^t \exp\left(\eta(\tau-t)\boldsymbol{\Sigma}_X\right)\boldsymbol{\Sigma}_X^{1/2}\,\mathrm{d}\boldsymbol{W}(\tau)\right)\\&=\eta^2\sigma^2\int_0^t \exp\left(2\eta(\tau-t)\boldsymbol{\Sigma}_X\right)\boldsymbol{\Sigma}_X\,\mathrm{d}\tau\\&=\eta^2\cdot\frac{\sigma^2}{2\eta}\left(\mathbf{I}_p-\exp\left(-2\eta t\boldsymbol{\Sigma}_X\right)\right)\preccurlyeq\eta\cdot\frac{\sigma^2}{2}\mathbf{I}_p.\end{aligned}$$

Therefore, $\boldsymbol{\Theta}(t)$ traverses first deterministically and rapidly from initial point $\boldsymbol{\theta}_0$ and then to a neighborhood of $\boldsymbol{\theta}_*$. Recall that $\alpha_X>0$ is the smallest eigenvalue of $\boldsymbol{\Sigma}_X$. Further investigations to the aforementioned ODE and SDE approximations allows us to conclude the following proposition.

Proposition 2.5. Suppose that Assumption 2.1 holds. For N sufficiently large, by setting $\eta\asymp\alpha_X^{-1}\log N/N$, $\boldsymbol{\Theta}(t)$ in (2.8) satisfies

$$\mathbb{E}\|\boldsymbol{\Theta}(N)-\boldsymbol{\theta}_*\|^2\lesssim\frac{\sigma^2}{\alpha_X}\cdot\frac{p\log N}{N}. \tag{2.9}$$

Proposition 2.5 implies that, given N streaming samples, the estimation error is in the order of p/N up to a $\log N$ factor. In other words, SGD for least squares achieves minimax optimal rate up to a logarithmic factor.

Remark. Although Theorem 2.4 establishes the weak convergence of the SGD iteration to the ODE and SDE solutions under two different scalings, no such convergence rates have been provided, and hence the asymptotic analysis using the continuous-time dynamics does *not* give the analogous discrete-time dynamics. The recent revised work by Li et al. (2017) concludes that the continuous-time dynamics serve as the so-called order-1 weak approximation to the discrete-time dynamics. However, their bound of the discrete–continuous gap ignores the dependency of dimension d. Such explicit bound is left for further investigations.

3 Achieving Optimal Rate via Trajectory Average

The previous section indicates that, starting from $\boldsymbol{\theta}_0=\mathcal{O}(\eta^{0.5})$, the SGD iteration (1.4) is close to the global minimizer $\boldsymbol{\theta}_*$ and keeps oscillating around such minimizer. We can further improve the bound using trajectory average:

$$\bar{\boldsymbol{\theta}}^{(N)}=\frac{1}{N}\sum_{t=1}^N\boldsymbol{\theta}^{(t)}.$$

This averaging scheme provides a sharper bound of convergence rate in two aspects: (i) the $\log N$ factor is eliminated; (ii) the constant factor is improved. Since $\boldsymbol{\theta}^{(t)}\approx\boldsymbol{\Theta}(t)$ for each $t=1,\ldots,N$, we have roughly that $\bar{\boldsymbol{\theta}}^{(N)}\approx\bar{\boldsymbol{\Theta}}(N)$ where

$$\bar{\boldsymbol{\Theta}}(N)=\frac{1}{N}\int_0^N\boldsymbol{\Theta}(t)\mathrm{d}t.$$

Utilizing probabilistic tools related to the Ornstein-Uhlenbeck processes, we conclude that the following proposition concerning the convergence rate of $\bar{\boldsymbol{\Theta}}(N)$.

Proposition 3.1. Suppose $\boldsymbol{\theta}_0=\boldsymbol{\theta}_*+\mathcal{O}_{\mathbb{P}}(\eta^{0.5})$. As $\eta N\to\infty$, $\bar{\boldsymbol{\Theta}}(N)$ weakly converges:

$$\sqrt{N}(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*)\Rightarrow\mathcal{N}\left(0,\sigma^2\boldsymbol{\Sigma}_X^{-1}\right),\quad\text{and} \tag{3.1}$$

$$\mathbb{E}\left\|\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*\right\|^2\lesssim\sigma^2\cdot\frac{\operatorname{tr}\left(\boldsymbol{\Sigma}_X^{-1}\right)}{N}. \tag{3.2}$$

Immediately from Proposition 3.1, we obtain the following corollary.

Corollary 3.2. For any given $\mathbf{x}_0\in\mathbb{R}^p$, as $\eta N\to\infty$, we have:

$$\sqrt{N}\mathbf{x}_0^{\mathrm{T}}\left(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*\right)\Rightarrow\mathcal{N}\left(0,(\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1}\mathbf{x}_0)\sigma^2\right),\quad\text{and} \tag{3.3}$$

$$\mathbb{E}\left(\left|\mathbf{x}_0^{\mathrm{T}}(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*)\right|^2\mid\mathbf{x}_0\right)\lesssim\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1}\mathbf{x}_0\cdot\frac{\sigma^2}{N}.$$

Moreover, let $\mathbf{x}$ be drawn from the data distribution with bounded second moments, then the prediction error is bounded as

$$\mathbb{E}\left|\mathbf{x}^{\mathrm{T}}(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*)\right|^2\lesssim\frac{p\sigma^2}{N}.$$

Proposition 3.1 and Corollary 3.2 imply that, under the scaling condition that $\eta N\to\infty$, $\bar{\boldsymbol{\Theta}}(N)$ approximately follows $\mathcal{N}\left(\boldsymbol{\theta}_*,(\sigma^2/N)\boldsymbol{\Sigma}_X^{-1}\right)$. The striking fact of these results is that the rate of convergence is independent of the stepsize η.

Remark. The results in Proposition 3.1 is related to classical averaged SGD finite-sample bounds by Ruppert (1988) and Polyak & Juditsky (1992), who prove that when minimizing a strongly convex function $\mathbb{E}[F(\boldsymbol{\theta};\zeta)]$ with a Lipschitz gradient, by choosing appropriate step sizes,

$$\sqrt{N}\left(\bar{\boldsymbol{\theta}}^{(N)}-\boldsymbol{\theta}_*\right)\Rightarrow\mathcal{N}\left(0,\mathbf{A}^{-1}\mathbf{S}^2\mathbf{A}^{-1}\right),$$

where $\mathbf{A} = \mathbb{E}\left[\nabla^2 F(\boldsymbol{\theta}_*;\zeta)\right]$ is the Hessian matrix, and $\mathbf{S}^2 = \mathbb{E}\left[\nabla F(\boldsymbol{\theta}_*;\zeta)\nabla F(\boldsymbol{\theta}_*;\zeta)\right]^{\mathrm{T}}$ is the covariance of noise at a local minimizer. In our case, $\mathbf{A} = \boldsymbol{\Sigma}_X$ and $\mathbf{S}^2 = \sigma^2\boldsymbol{\Sigma}_X$ so $\mathbf{A}^{-1}\mathbf{S}^2\mathbf{A}^{-1} = \sigma^2\boldsymbol{\Sigma}_X^{-1}$. Seeing the approximation $\bar{\boldsymbol{\theta}}^{(N)} \approx \bar{\boldsymbol{\Theta}}(N)$, the results in Proposition 3.1 matches the best known sample bounds for SGD.

Two-Phase Phenomenon. We propose a two-phase training strategy as follows. In the first phase, we run the SGD for N_{cvg} iterations to get convergence to a $\mathcal{O}(\eta^{0.5})$-neighborhood. In the second phase, we compute the partial sum of the last N_{avg} SGD iterates and then divide it by N_{avg}, where the output $\bar{\boldsymbol{\theta}}^{(N)}$ approximately follows the distribution $\mathcal{N}\left(\boldsymbol{\theta}_*, (\sigma^2/N)\boldsymbol{\Sigma}_X^{-1}\right)$. For simplicity, we consider the case where the two phases have the same number of samples. Suppose $N_{cvg} = N_{avg} \sim 0.5\alpha_X^{-1}\eta^{-1}\log\left(\eta^{-1}\right)$. From Proposition 3.1, we know that it takes

$$N = N_{cvg} + N_{avg} \sim \alpha_X^{-1}\eta^{-1}\log\left(\eta^{-1}\right)$$

iterates to obtain an estimator satisfying

$$\mathbb{E}\|\bar{\boldsymbol{\Theta}}(N) - \boldsymbol{\theta}_*\|^2 \le \sigma^2 \cdot \frac{\mathrm{tr}(\boldsymbol{\Sigma}_X^{-1})}{N_{avg}} \le 2\sigma^2 \cdot \frac{\mathrm{tr}(\boldsymbol{\Sigma}_X^{-1})}{N}. \tag{3.4}$$

Corollary 3.2 implies

$$\mathbb{E}\left|\mathbf{x}_0^{\mathrm{T}}(\bar{\boldsymbol{\Theta}}(N) - \boldsymbol{\theta}_*)\right|^2 \lesssim \sigma^2\frac{p}{N_{avg}} = 2\sigma^2\frac{p}{N}. \tag{3.5}$$

The error bound in (3.4) is $\mathcal{O}\left(\sigma^2\,\mathrm{tr}(\boldsymbol{\Sigma}_X^{-1})\cdot N^{-1}\right)$, which improves the bound $\mathcal{O}\left(\sigma^2 p\alpha_X^{-1}\cdot(\log N)N^{-1}\right)$ without averaging step (2.9) in two ways: (i) the Hessian-related factor $p\alpha_X^{-1}$ in (3.4) is improved to $\mathrm{tr}(\boldsymbol{\Sigma}_X^{-1})$, and (ii) the $\log N$ factor is eliminated. Moreover, the rate is independent of η under the asymptotic regime $\eta N \to \infty$. We comment here that the numbers of data in the two phases are not required to be the same. However, the main results largely holds as long as the two sizes of data samples are of the same magnitude. Interested readers can find more on the variants of averaging schemes in Rakhlin et al. (2012) and Shamir & Zhang (2013).

4 Sparse Online Regression

Recall that our true model is $y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_* + \varepsilon$. We consider the setting that $\boldsymbol{\theta}_*$ is s-sparse, that is, the support set of $\boldsymbol{\theta}_*$, $S = \mathrm{supp}(\boldsymbol{\theta}_*)$, has size s.

We consider the setting where we have at hand a mini-batch data, $\{(\mathbf{x}_k, y_k)\}_{k=1}^n$, after which the data come in a streaming fashion. This resembles the setting of many scientific studies. For example, in genomic and biomedical imaging studies, a small-scale clinical trial is first carried out for some preliminary results, after which more and more studies are carried out in order to improve the detection power. Motivated by this setting, we propose and study the following two-step algorithm for sparse online regression:

- Step 1. Burn-in using offline learning: We use nonconvex penalized regression to obtain a burn-in estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{init}}$ with active set $\mathcal{A}_{\mathrm{init}} = \mathrm{supp}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{init}}\right)$.
- Step 2. Refinement using online learning: We start with the initial estimator in Step 1 and then run the truncated stochastic gradient update.

We analyze the above two steps separately.

4.1 Burn-in Step

Let $\mathbf{X}_{\mathrm{init}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$. In the first offline stage, we run the following nonconvex penalized regression

$$\widehat{\boldsymbol{\theta}}_{\mathrm{init}} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}}\left\{\underbrace{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}_{\mathrm{init}}\boldsymbol{\theta}\|_2^2}_{\mathcal{L}_{\mathrm{init}}(\boldsymbol{\theta})} + \sum_{j=1}^{p} p_\lambda(\boldsymbol{\theta}_j)\right\}, \tag{4.1}$$

where $p_{\lambda_{\mathrm{init}}}(\cdot)$ is a folded concave penalty with regularization parameter λ_{init}, such as SCAD (Fan & Li, 2001) or MCP (Zhang, 2010). One can adopt iLAMM for solving the optimization problem (4.1) (Fan et al., 2017), and we only require some common regularity conditions, as specified in the following assumption.

Assumption 4.1. Let ν and c be some appropriate constants. Suppose $p_\lambda(x)$ can be written as $p_\lambda(x) = \lambda|x| + h_\lambda(x)$ such that

(a) $p'_\lambda(x) = 0$, for $|x| \ge \nu \ge c\lambda > 0$.

(b) $h'_\lambda(x)$ is monotone and Lipschitz continuous, i.e., for $x' \ge x$, there exists a constant $\xi_- \ge 0$ such that $h'_\lambda(x') - h'_\lambda(x) \le -\xi_-(x' - x)$.

(c) $h_\lambda(x)$ and $h_\lambda(x)$ are symmetric and pass through 0.

(d) $|h'_\lambda(x)| \le \lambda$ for any $x \in \mathbb{R}$.

We need the following definition of restricted eigenvalues. Let $H_{\text{init}} = \mathbf{X}_{\text{init}}^{\mathrm{T}}\mathbf{X}_{\text{init}}/n$ be the Hessian matrix.

Definition 4.2 (Restricted Eigenvalue, RE)**.** The restricted eigenvalue is defined as

$$\kappa_+(m,\gamma) = \sup_{\mathbf{u}} \left\{ \mathbf{u}^{\mathrm{T}} H_{\text{init}} \mathbf{u}/n : \|\mathbf{u}\|_2 = 1, \mathbf{u} \in \mathcal{C}(m,\gamma) \right\};$$
$$\kappa_-(m,\gamma) = \inf_{\mathbf{u}} \left\{ \mathbf{u}^{\mathrm{T}} H_{\text{init}} \mathbf{u}/n : \|\mathbf{u}\|_2 = 1, \mathbf{u} \in \mathcal{C}(m,\gamma) \right\},$$

where $\mathcal{C}(m,\gamma) \equiv \left\{ \mathbf{u} : S \subseteq J, |J| \leq m, \|\mathbf{u}_{J^c}\|_1 \leq \gamma \|\mathbf{u}_J\|_1 \right\}$ is a local ℓ_1 cone.

We say the RE condition holds if there exists m and γ such that $0 < \kappa_* \leq \kappa_-(m,\gamma) \leq \kappa_+(m,\gamma) \leq \kappa^* < \infty$. Raskutti et al. (2010) showed that the RE condition holds with high probability when the design variables have sub-Gaussian tails. We also need the following assumption concerning minimal signal strength.

Assumption 4.3 (Minimal Signal Strength)**.** Assume that $\min_{j \in \mathcal{S}} |\boldsymbol{\theta}_{*,j}| \geq 2\lambda = 2c\sqrt{\log p/n}$.

Assumption 4.3 is rather mild since $\sqrt{\log p/n}$ diminishes quickly. In other words, Assumption 4.3 can be regarded as sample complexity requirement in the offline line step. Our first result in this section concerns the statistical property of the burn-in estimator $\widehat{\boldsymbol{\theta}}_{\text{init}}$, when using iLAMM for computation. The readers are referred to Fan et al. (2017) for more information regarding to computation using iLAMM.

Proposition 4.4. Assume that Assumption 4.1 holds. Assume that the RE condition holds with $m = 2s$, $\gamma = 3$, and $\kappa_* > \xi_-$. Suppose that $n \gtrsim s \log p$ and $\lambda = c\sqrt{\log p/n}$ for a constant c. Then, with probability at least $1 - 1/p$, we must have

$$\|\widehat{\boldsymbol{\theta}}_{\text{init}} - \boldsymbol{\theta}_*\| \lesssim s\sqrt{\log p/n} \quad \text{and} \quad |\mathcal{A}_{\text{init}}| \leq (C+1)s,$$

for a constant C. Moreover, if we assume Assumption 4.3, then $\mathcal{S} \subset \mathcal{A}_{\text{init}}$.

4.2 Refinement Step

In this section, we extend our analysis in §2 to the online updates after the offline burn-in step, namely, by adding in a truncation step, we obtain a rate that matches the finite-sample bound for batch method. We start from $\boldsymbol{\theta}_{\text{init}}$ and run the following truncated stochastic gradient descent algorithm:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}^{(t)} &= \boldsymbol{\theta}^{(t-1)} + \eta(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t, \\ \boldsymbol{\theta}^{(t)} &= \text{Truncate}(\widetilde{\boldsymbol{\theta}}^{(t)}, k_0). \end{aligned} \tag{4.2}$$

where $k_0 = |\mathcal{A}_{\text{init}}| \geq s$ is the size of the support set of $\boldsymbol{\theta}_{\text{init}}$. The first step is simply the stochastic gradient descent step. The second step is added, where we recall that *Truncate* operator keeps the maximal k_0 absolute coordinates and zero-out all others. To avoid ambiguity, if there are multiple k_0 maximals coordinates then the ones with least indices are selected. We establish the following lemma.

Lemma 4.5. Suppose we are under the assumptions in Proposition 4.4, and let $\mathcal{M}$ be a diagonal matrix with 1s on the (i,i) entries where $i \in \mathcal{A}_{\text{init}}$, and 0s otherwise. Then with probability $\geq 1 - 1/p$ we have for all $t \leq T$

$$\text{Truncate}(\widetilde{\boldsymbol{\theta}}^{(t)}, k_0) = \mathcal{M}\widetilde{\boldsymbol{\theta}}^{(t)}.$$

In other words, the algorithm focuses on the subset that has been selected from the true support under the assumptions in Proposition 4.4.

From the fact that $\boldsymbol{\theta}^{(t-1)}$ is supported on $\mathcal{A}_{\text{init}}$

$$\begin{aligned} \boldsymbol{\theta}^{(t)} &= \mathcal{M}\left[\boldsymbol{\theta}^{(t-1)} + \eta(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t\right] \\ &= \boldsymbol{\theta}^{(t-1)} + \eta(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathcal{M}\mathbf{x}_t. \end{aligned}$$

Let $\mathbf{z}_t = \mathbf{x}_t|_{\mathcal{A}_{\text{init}}}$ be the k_0-dimensional vector that project $\mathbf{x}_t \in \mathbb{R}^p$ onto active set $\mathcal{A}_{\text{init}}$, and similarly, let $\boldsymbol{\psi}^{(t)} = \boldsymbol{\theta}^{(t)}|_{\mathcal{A}_{\text{init}}} \in \mathbb{R}^{k_0}$ and $\boldsymbol{\psi}_* = \boldsymbol{\theta}_*|_{\mathcal{A}_{\text{init}}} \in \mathbb{R}^{k_0}$ be separately the projected iteration and minimizer. Then by projection onto $\mathcal{A}_{\text{init}}$, the problem is translated to a low-dimensional problem

$$\boldsymbol{\psi}^{(t)} = \boldsymbol{\psi}^{(t-1)} + \eta(y_t - \mathbf{z}_t^{\mathrm{T}}\boldsymbol{\psi}^{(t-1)})\mathbf{z}_t.$$

We can now apply results from §2. Analogous to (2.5), $\boldsymbol{\psi}^{(t)}$ can be approximated by $\boldsymbol{\Psi}(t)$ which is the solution to SDE

$$\mathrm{d}\boldsymbol{\Psi}(t) = -\eta\boldsymbol{\Sigma}_Z\left(\boldsymbol{\Psi}(t) - \boldsymbol{\psi}_*\right)\mathrm{d}t + \eta\sigma\boldsymbol{\Sigma}_Z^{1/2}\mathrm{d}\boldsymbol{W}(t). \tag{4.3}$$

The error bound in terms of the ℓ_2-loss is analogous to Proposition 2.5, as follows.

Proposition 4.6. Under Assumptions 2.1 and 4.3, when we know there are N streaming samples, by setting $\eta = \alpha_X^{-1}\log N/N$ then the rate is

$$\mathbb{E}\left[\|\boldsymbol{\theta}^{(N)} - \boldsymbol{\theta}_*\|^2; \mathcal{A}_*\right] \lesssim \frac{\sigma^2}{\alpha_X} \cdot \frac{s\log N}{N}, \tag{4.4}$$

where $\mathcal{A}_*$ is the event set that $S \subset \mathcal{A}_*$.

We believe the missing of $\log p$ compared to minimax rate is due to the minimal signal strength assumption, which results in a different class of minimax problems (Raskutti et al., 2010).

5 Simulation Results: Low-dimensional Case

In this section, we conduct some numerical experiments to evaluate the algorithm we proposed in the low-dimensional setting. Limited by space, we refer the readers to the supplementary material for simulations in high-dimensional setting.

In low-dimensional case, we focus on the utility of the averaging process in the proposed algorithm. We fix the dimension to be $p = 10$ and generate $N = 20000$ samples according to $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}_* + \varepsilon_i$ where the regression coefficient $\boldsymbol{\theta}_* = (1, 1, ..., 1)^{\mathrm{T}}$, ε_i follows $N(0, 1)$ and $\mathbf{x}_i$ follows a multivariate $N(\mathbf{0}, \mathbf{I}_p)$ so $\alpha_X = 1$. We choose for simplicity $\boldsymbol{\theta}_0 = \mathbf{0}$ (other values apply). We run the stochastic gradient descent to derive an estimator $\widehat{\boldsymbol{\theta}}$ with step size $\eta_i \equiv \log N/N$ being a constant for all $1 \le i \le N$. The above procedure is repeated for 100 times and the average ℓ_2 error $||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*||_2$ is recorded.

Meanwhile, we also start the averaging process from $k = 2000, 5000, 8000$ and 10000 in each repetition and record the mean ℓ_2 loss separately. This means that we consider 5 cases here altogether. The results are shown in Figure 1.

- C1: $T_{cvg} = 2000$, $T_{avg} = 18000$;
- C2: $T_{cvg} = 5000$, $T_{avg} = 15000$;
- C3: $T_{cvg} = 8000$, $T_{avg} = 12000$;
- C4: $T_{cvg} = 10000$, $T_{avg} = 10000$;
- C5: $T_{cvg} = 20000$, $T_{avg} = 0$.

From the upper panel in Figure 1, we can tell that averaging does help when T_{cvg} is large enough. This is consistent to the Two-phase Training Strategy since when T_{cvg} is small, the estimator has *not* yet converged to the $\mathcal{O}(N^{-0.5})$-neighborhood of the true coefficient. Simultaneously, we can see from the lower panel in Figure 1 that averaging helps reduce the oscillation of the estimator during SGD. This matches the discussion in §3 and shows that it can perform steadily well in experiments for simulated data.

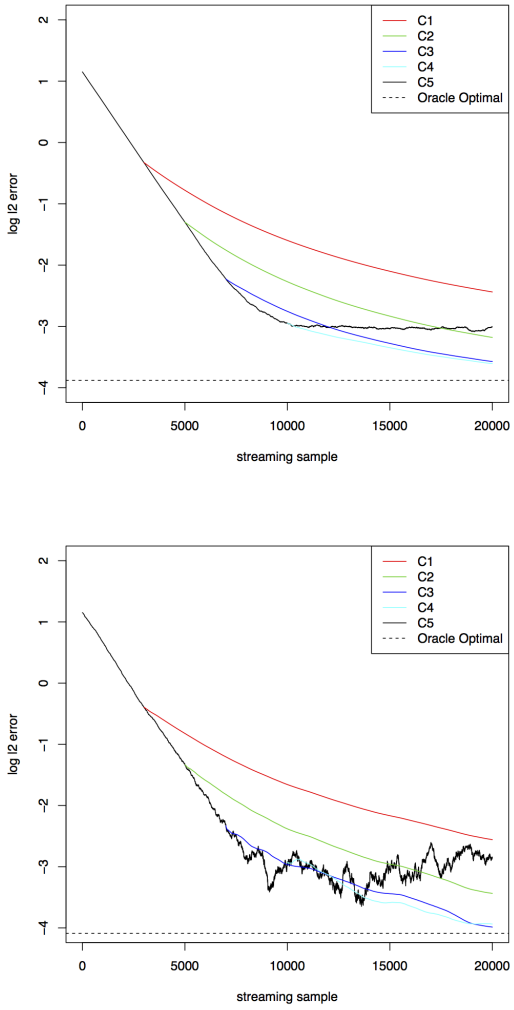

Figure 1: The ℓ_2 error in case C1 to C5 in the low-dimensional case. The dotted line is the oracle optimal error rate we can achieve when all samples are offline. Upper: the mean ℓ_2 error over 100 repetition; Lower: the ℓ_2 error in one repetition.

6 Conclusion

In this work, we propose a diffusion approximation approach to characterize the exact dynamics of online regression. This allows us to prove the near-optimal statistical rate of convergence along the streaming process. Using the idea of iteration average, we further improve the rate of convergence by eliminating the $\log N$ factor. Lastly, we propose a two-step algorithm for sparse online regression: a burn-in step using offline learning and a refinement step using truncated SGD. We show the proposed two-step algorithm produces near optimal sparse estimators, as if the locations of the nonzeroes were known in advance. We in addition conduct simulation experiments in both low-dimensional and high-dimensional settings, which substantiate our proposed theory.

References

Agarwal, A., Bartlett, P. L., Ravikumar, P., & Wainwright, M. J. (2012a). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5), 3235–3249.

Agarwal, A., Negahban, S., & Wainwright, M. J. (2012b). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, (pp. 2452–2482).

Benveniste, A., Métivier, M., & Priouret, P. (2012). *Adaptive Algorithms and Stochastic Approximations*. Springer.

Bertsekas, D. & Tsitsiklis, J. (1989). *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific.

Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming, Ser. B*, 129(2), 163–195.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

Chen, X., Lee, J. D., Tong, X. T., & Zhang, Y. (2016). Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*.

Da Prato, G. & Zabczyk, J. (2014). *Stochastic equations in infinite dimensions*. Cambridge university press.

Daubechies, I., Defrise, M., & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11), 1413–1457.

Dieuleveut, A., Flammarion, N., & Bach, F. (2016). Harder, better, faster, stronger convergence rates for least-squares regression. *arXiv preprint arXiv:1602.05419*.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3), 613–627.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul), 2121–2159.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2), 407–499.

Ethier, S. N. & Kurtz, T. G. (2005). *Markov Processes: Characterization and Convergence*. John Wiley & Sons.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348–1360.

Fan, J., Liu, H., Sun, Q., & Zhang, T. (2017). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics, to appear*.

Kushner, H. & Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer.

Langford, J., Li, L., & Zhang, T. (2009). Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(Mar), 777–801.

Li, C. J., Wang, M., Liu, H., & Zhang, T. (2018). Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming, Ser. B*, 167(1), 75–97.

Li, Q., Tai, C., & E, W. (2017). Stochastic modified equations and adaptive stochastic gradient algorithms. *arXiv preprint arXiv:1511.06251v3*.

Loh, P.-L. & Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics, to appear*.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2), 772–801.

Mandt, S., Hoffman, M., & Blei, D. (2016). A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning* (pp. 354–363).

Nedić, A. (2011). Random algorithms for convex minimization problems. *Mathematical Programming, Ser. B*, 129(2), 225–253.

Nedic, A. & Bertsekas, D. P. (2001). Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1), 109–138.

Nedić, A., Bertsekas, D. P., & Borkar, V. S. (2001). Distributed asynchronous incremental subgradient methods. *Studies in Computational Mathematics*, 8(C), 381–407.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., & Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regulariz-

ers. *Statistical Science*, 27(4), 538–557.

Nemirovskii, A. & Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley.

Oksendal, B. (2003). *Stochastic Differential Equations (6th edition)*. Springer.

Polyak, B. T. & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4), 838–855.

Rakhlin, A., Shamir, O., & Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 449–456).

Raskutti, G., Wainwright, M. J., & Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug), 2241–2259.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. *Technical Report, Cornell University Operations Research and Industrial Engineering*.

Shamir, O. & Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 71–79).

Su, W. & Zhu, Y. (2018). Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*.

Wang, M. & Bertsekas, D. P. (2015). Incremental constraint projection methods for variational inequalities. *Mathematical Programming, Ser. A*, 150(2), 321–363.

Wang, M. & Bertsekas, D. P. (2016). Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1), 681–717.

Xiao, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct), 2543–2596.

Yuan, X.-T. & Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14, 899–925.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 894–942.

A A Primer to SDE

To get readers familiar with Itô integral and stochastic differential equations, we provide in this section a quick sketch introducing the topic that is necessary to understand this paper. The readers are referred to Oksendal (2003) for fundamental topics on SDE.

We first remind readers the definition of the Brownian motion.

Definition A.1 (Brownian motion)**.** We call $B(t), t \geq 0$ a Brownian motion if it satisfies the following

(i) The $B(0) = 0$;

(ii) If $0 = t_0 < t_1 < \cdots < t_n$ then $B(t_k) - B(t_{k-1})$, $k = 1, \ldots, n$ are independent;

(iii) If $s, t \geq 0$ then $B(t+s) - B(t)$ is distributed as a normal distribution with mean 0 and variance s, i.e. for each $x \in \mathbb{R}$

$$\mathbb{P}\left(B(t+s) - B(t) \leq x\right) = \frac{1}{\sqrt{2\pi s}} \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2s}\right) \mathrm{d}y;$$

(iv) The trajectory $t \mapsto B(t)$ is continuous.

In the next part we introduce *Itô integral* in a special case where the deterministic function $f(t)$ is integrated with regards to the Brownian motion $B(t)$ on a finite interval. The definition is developed in two steps, firstly for $f(t)$ being piecewise constant and secondly for $f(t)$ being continuous.

(a) $f(t)$ is a piecewise constant function. Assume there is a partition $0 = t_0 < t_1 < \cdots < t_n = T$ and $c_1, \ldots, c_n \in \mathbb{R}$ so that $f(t)$ takes the value of c_i on the interval $(t_{i-1}, t_i]$ for each i, we define a random variable $I_T(f)$

$$I_T(f) \equiv \sum_{i=1}^{n} c_i \left(B(t_i) - B(t_{i-1})\right). \tag{A.1}$$

It is obvious that from the definition, $I_T(f)$ is a mean-zero Gaussian random variable.

(b) $f(t)$ is a continuous function. One can use a sequence of piecewise constant functions to approach $f(t)$ as follows. Let $f^n(t) = \sum_{i=1}^{n} \mathbf{1}_{t_{i-1} < t \leq t_i} f(t_{i-1})$ where the partition $0 = t_0 < t_1 < \cdots < t_n = T$ satisfies $\max_{1 \leq i \leq n} \left|t_i^n - t_{i-1}^n\right| \to 0$ as n goes to infinity. It is *not* hard to verify that the random variables $I_T(f^n)$ converges in L^2 metric in the space of random variables. Moreover, the L^2-limit is independent of the choice of partition sequences. In this case $I_T(f)$ is defined as

$$I_T(f) \equiv L^2\text{-}\lim_{n \to \infty} I_T(f^n), \tag{A.2}$$

which is a mean-zero Gaussian random variable.

Definition A.2. The Itô integral of function $f(t)$ with regards to the Brownian motion $B(t)$ on interval $[0, T]$ is defined as

$$\int_0^T f(t) \mathrm{d}B_t \equiv I_T(f),$$

where the mean-zero Gaussian random variable $I_T(f)$ was defined as in (A.2).

Proposition A.3. For deterministic, continuous function $f(t)$ defined on $[0, T]$ we have the following

$$\mathbb{E}\left(\int_0^T f(t) \mathrm{d}B_t\right)^2 = \int_0^T f(t)^2 \mathrm{d}t. \tag{A.3}$$

The result of Proposition A.3 is a special case of *Itô isometry* whose proof can be found in Oksendal (2003). Therefore the proof of Proposition A.3 is omitted.

We introduce the stochastic differential equation via one example, as

$$\mathrm{d}X(t) = \alpha X(t)\mathrm{d}t + \sigma \mathrm{d}B(t). \tag{A.4}$$

Here α, σ are two real numbers. (A.4) is called the *Ornstein-Uhlenbeck equation* (or in physicists' terminology the *Langevin equation*).

Definition A.4. We define $X(t)$ as the solution to the Ornstein-Uhlenbeck equation (A.4) with initial value $X(0) = X^o$, if $X(t)$ is a stochastic process that satisfies the following equation

$$X(t) - X^o = \alpha \int_0^T X(s)\mathrm{d}s + \sigma B(t). \tag{A.5}$$

The integral on the right hand of (A.5) is the standard Riemann integral.

In our settings, the language of differentials and integrals can be understood in the same way as in basic calculus. To solve the Ornstein-Uhlenbeck equation (A.4) we note that when $Z(t) = \exp(-\alpha t)$, the chain rule gives

$$\mathrm{d}\left(X(t)Z(t)\right) = Z(t)\mathrm{d}X(t) + X(t)\mathrm{d}Z(t), \tag{A.6}$$

and hence

$$\mathrm{d}\left(X(t)\exp(-\alpha t)\right) = \exp(-\alpha t)\left(\mathrm{d}X(t) - \alpha X(t)\mathrm{d}t\right) = \sigma \exp(-\alpha t)\mathrm{d}B(t).$$

Integrating from 0 to t we have when $X(0) = X^o$

$$X(t)\exp(-\alpha t) - X^o = \sigma \int_0^T \exp(-\alpha s)\mathrm{d}B(s).$$

Multiplying both sides by the exponential factor $\exp(\alpha t)$ we obtain the solution to (A.4) as

$$X(t) = X^o \exp\left(\alpha t\right) + \sigma \int_0^T \exp\left(\alpha(t-s)\right)\mathrm{d}B(s). \tag{A.7}$$

When $\alpha < 0$ the solution is called *(standard) Ornstein-Uhlenbeck process*, and when $\alpha > 0$ the solution is called *unstable Ornstein-Uhlenbeck process*.

Remark Note when $Z(t)$ is a stochastic process instead of a deterministic function, we do *not* have the chain rule illustrated in (A.6). Its stochastic calculus counterpart however, called *Itô's lemma* or *Itô's formula*, is invoked. Limited by space we refer the readers to (Oksendal, 2003, Sec. 4.2) for related materials.

B Deferred Proofs in §2

B.1 Proof of Lemma 2.2

Proof of Lemma 2.2. From (1.1) and Assumption 2.1 we have

$$\begin{aligned}\mathbb{E}\left[\left(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)}\right)\mathbf{x}_t \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\left(\mathbf{x}_t^{\mathrm{T}}(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(t-1)}) + \varepsilon_t\right)\mathbf{x}_t \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(t-1)}) \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\varepsilon_t\mathbf{x}_t \mid \mathcal{F}_{t-1}\right],\end{aligned}$$

where

$$\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(t-1)}) \mid \mathcal{F}_{t-1}\right] = -\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}\middle|\mathcal{F}_{t-1}\right](\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}_*) = -\boldsymbol{\Sigma}_X(\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}_*).$$

By iterated law of expectation and unbiasedness assumption

$$\mathbb{E}\left[\varepsilon_t \mathbf{x}_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left(\varepsilon_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}\right)\mathbf{x}_t \mid \mathcal{F}_{t-1}\right] = 0.$$

Combining the above three displays concludes the lemma.

□

B.2 Proof of Lemma 2.3

Proof of Lemma 2.3. The first equality is derived from the definition of $\boldsymbol{e}_t$ in (2.2). For the second, using Assumption 2.1 and given the knowledge of $\mathbf{x}_t$ and $\mathcal{F}_{t-1}$ one has

$$\begin{aligned}\operatorname{var}\left[(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}\right] &= \operatorname{var}\left[(\mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}_* + \varepsilon_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}\right] \\ &= \operatorname{var}\left(\varepsilon_t \mathbf{x}_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}\right) + \mathcal{O}(\eta^{0.5}) \\ &= \mathbf{x}_t \operatorname{var}(\varepsilon_t \mid \mathbf{x}_t)\mathbf{x}_t^{\mathrm{T}} + \mathcal{O}(\eta^{0.5}) = \sigma^2 \mathbf{x}_t\mathbf{x}_t^{\mathrm{T}} + \mathcal{O}(\eta^{0.5}),\end{aligned}$$

and

$$\mathbb{E}\left[(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}\right] = \mathbf{x}_t^{\mathrm{T}}(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(t-1)})\mathbf{x}_t = \mathcal{O}(\eta^{0.5}).$$

Using the iterated law of variance

$$\begin{aligned}&\operatorname{var}\left[(y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\operatorname{var}((y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}) \mid \mathcal{F}_{t-1}\right] + \operatorname{var}\left[\mathbb{E}((y_t - \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\theta}^{(t-1)})\mathbf{x}_t \mid \mathbf{x}_t, \mathcal{F}_{t-1}) \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\sigma^2\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}} + \mathcal{O}(\eta^{0.5}) \mid \mathcal{F}_{t-1}\right] + \operatorname{var}\left[\mathcal{O}(\eta^{0.5}) \mid \mathcal{F}_{t-1}\right] = \sigma^2\boldsymbol{\Sigma}_X + \mathcal{O}(\eta^{0.5}).\end{aligned}$$

This completes our proof.

□

B.3 Proof of Theorem 2.4

Proof of Theorem 2.4. We adopt the weak convergence theory of Markov processes, which essentially states that if the infinitesimal mean and variance converges to some limiting distribution, then weak convergence of Markov processes applies. The readers are referred to Ethier & Kurtz (2005, §7.4, Theorem 4.2) for more details.

(i) From (2.3)

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \eta\boldsymbol{\Sigma}_X(\boldsymbol{\theta}_* - \boldsymbol{\theta}^{(t-1)}) + \eta\boldsymbol{e}_t.$$

Lemma 2.2 implies that the stochastic gradient at step t has its infinitesimal mean

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\boldsymbol{\Theta}^a(t) \mid \boldsymbol{\Theta}^a(0) = \boldsymbol{\theta}] = -\boldsymbol{\Sigma}_X(\boldsymbol{\theta} - \boldsymbol{\theta}_*).$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[(\boldsymbol{\Theta}^a(t) - \boldsymbol{\Theta}^a(0))(\boldsymbol{\Theta}^a(t) - \boldsymbol{\Theta}^a(0))^{\mathrm{T}} \mid \boldsymbol{\Theta}^a(0) = \boldsymbol{\theta}] = \eta\sigma^2\boldsymbol{\Sigma}_X + \mathcal{O}(\eta^{1.5}) \to 0.$$

Using the weak convergence theory as mentioned before, the stochastic process $\boldsymbol{\theta}^{\eta,(\lfloor t\eta^{-1}\rfloor)}$ converges weakly to the solution to the ODE (2.6).

(ii) When $\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_* + \mathcal{O}(\eta^{0.5})$, similar to (i) one has from Lemma 2.2 that the infinitesimal mean of $\boldsymbol{\Theta}^b(t)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\boldsymbol{\Theta}^b(t) \mid \boldsymbol{\Theta}^b(0) = \boldsymbol{\theta}] = -\boldsymbol{\Sigma}_X(\boldsymbol{\theta} - \boldsymbol{\theta}_*). \tag{B.1}$$

Lemma 2.3 implies the infinitesimal variance of $\boldsymbol{\Theta}^b(t)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[(\boldsymbol{\Theta}^b(t) - \boldsymbol{\Theta}^b(0))(\boldsymbol{\Theta}^b(t) - \boldsymbol{\Theta}^b(0))^{\mathrm{T}} \mid \boldsymbol{\Theta}^b(0) = \boldsymbol{\theta}] = \sigma^2\boldsymbol{\Sigma}_X + \mathcal{O}(\eta^{0.5}) \to \sigma^2\boldsymbol{\Sigma}_X.$$

Again using the weak convergence theory of Markov processes, the stochastic process $\boldsymbol{\theta}^{\eta,(\lfloor t\eta^{-1}\rfloor)}$ converges weakly to the solution to the SDE (2.7).

□

B.4 Proof of Proposition 2.5

Proof of Proposition 2.5. Let δ be an arbitrary positive number that is no greater than $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_\infty$. Then starting from $\boldsymbol{\theta}_0$, the traverse time T such that $\|\boldsymbol{\Theta}(T) - \boldsymbol{\theta}_*\|_\infty \le \delta$ satisfies

$$\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_\infty \exp\left(-\eta\alpha_X \cdot T\right) = \delta.$$

By setting $\delta = \sqrt{\eta\sigma^2/2}$ we have when $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_\infty \ge \sqrt{\eta\sigma^2/2}$ it takes

$$T = \alpha_X^{-1}\eta^{-1}\log\left(\frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_\infty}{\sqrt{\eta\sigma^2/2}}\right) \sim 0.5\alpha_X^{-1}\eta^{-1}\log\left(\eta^{-1}\right) \tag{B.2}$$

iterations for the SGD dynamics to obtain $\boldsymbol{\Theta}(T)$ in a $\mathcal{O}(\eta^{0.5})$ neighborhood of $\boldsymbol{\theta}_*$, where $\mathbb{E}\|\boldsymbol{\Theta}(T) - \boldsymbol{\theta}_*\|^2 \le (p\sigma^2/2)\eta$. Plugging in $\eta = \alpha_X^{-1}\log N/N$ we conclude Proposition 2.5. □

C Deferred Proofs in §3

C.1 Proof of Proposition 3.1

Proof of Proposition 3.1. Writing $\boldsymbol{\Theta}(t) - \boldsymbol{\theta}_*$ we can assume WLOG $\boldsymbol{\theta}_* = 0$. Integrating both sides of (2.5) we have

$$\boldsymbol{\Theta}(N) - \boldsymbol{\theta}_0 = -\eta\boldsymbol{\Sigma}_X\int_0^N \boldsymbol{\Theta}(t)\mathrm{d}t + \eta\sigma\boldsymbol{\Sigma}_X^{1/2}\boldsymbol{W}(N)$$

Therefore

$$\sqrt{N}\bar{\boldsymbol{\Theta}}(N) = \frac{1}{\sqrt{N}}\int_0^N \boldsymbol{\Theta}(t)\mathrm{d}t = -\frac{\eta^{-1}}{\sqrt{N}}\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{\Theta}(N) - \boldsymbol{\theta}_0) + \frac{\sigma}{\sqrt{N}}\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{W}(N), \tag{C.1}$$

and since $\boldsymbol{\theta}_0 = \mathcal{O}(\eta^{0.5})$ and at stationarity $\boldsymbol{\Theta}(N) = \mathcal{O}(\eta^{0.5})$, the first term

$$\frac{\eta^{-2}}{N}\mathcal{O}(\eta) = \mathcal{O}\left((\eta N)^{-1}\right),$$

which as $\eta N \to \infty$ converges to 0. Hence $\sqrt{N}\bar{\boldsymbol{\Theta}}(N) = (\sigma/\sqrt{N})\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{W}(N) \Rightarrow \mathcal{N}(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$, indicating (3.1). Also

$$\mathrm{var}\left[\sqrt{N}\bar{\boldsymbol{\Theta}}(N)\right] = \frac{\eta^{-2}}{N}\mathcal{O}(\eta) + \sigma^2\boldsymbol{\Sigma}_X^{-1} = \mathcal{O}\left((\eta N)^{-1}\right) + \sigma^2\boldsymbol{\Sigma}_X^{-1},$$

which completes the proof of (3.2).

□

C.2 Proof of Corollary 3.2

Proof of Corollary 3.2. From (3.1), we have the rate in prediction error result as follows. Therefore for any fixed $\mathbf{x}_0$, left-multiplying both sides of (C.1) by $\mathbf{x}_0^{\mathrm{T}}$ allows us to conclude

$$\sqrt{N}\mathbf{x}_0^{\mathrm{T}}\bar{\boldsymbol{\Theta}}(N) = -(\eta N)^{-1}\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1}\cdot\sqrt{N}(\boldsymbol{\Theta}(N)-\boldsymbol{\theta}_0) + \frac{\sigma}{\sqrt{N}}\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{W}(N). \tag{C.2}$$

Thus as $\eta N\to\infty$ and from (3.1), the first term on RHS of (C.2) weakly converges to 0, and in the second $\boldsymbol{W}(N)/\sqrt{N}$ is a standard normal variable of p dimensions. Hence

$$\sqrt{N}\mathbf{x}_0^{\mathrm{T}}\left(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*\right)\Rightarrow\sigma\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1/2}\left(\frac{\boldsymbol{W}(N)}{\sqrt{N}}\right)\sim\mathcal{N}\left(0,\sigma^2\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1}\mathbf{x}_0\right),$$

where we uses the property of multivariate normal distribution, concluding (3.3). Moreover from (C.2) we conclude when N is sufficiently large

$$\mathbb{E}\left(\left|\mathbf{x}_0^{\mathrm{T}}(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*)\right|^2\mid\mathbf{x}_0\right)\lesssim\frac{\sigma^2}{N}\cdot\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1}\mathbf{x}_0.$$

Now randomize $\mathbf{x}_0\sim\mathcal{D}$ and we have

$$\begin{aligned}\mathbb{E}\left|\mathbf{x}_0^{\mathrm{T}}(\bar{\boldsymbol{\Theta}}(N)-\boldsymbol{\theta}_*)\right|^2&\lesssim\frac{\sigma^2}{N}\mathbb{E}\left[\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\Sigma}_X^{-1}\mathbf{x}_0\right]\\&=\frac{\sigma^2}{N}\mathbb{E}\,\mathrm{tr}\left[\boldsymbol{\Sigma}_X^{-1}\mathbf{x}_0\mathbf{x}_0^{\mathrm{T}}\right]=\frac{\sigma^2}{N}\mathbb{E}\,\mathrm{tr}\left(\mathbf{I}_p\right)=\frac{p\sigma^2}{N},\end{aligned}$$

completing our proof by choosing $C\geq 2$.

□

D Deferred Proofs in §4

D.1 Proof of Proposition 4.4

In order to prove this proposition, we need two additional lemmas. The first lemma concerns a deterministic error rate given the even that $\|\mathcal{L}_{\mathrm{init}}(\boldsymbol{\theta}_*)\|_\infty\leq\lambda/2$.

Lemma D.1. Assume that Assumption 4.1 holds. Further Assume that the RE condition holds with $m=2s$, $\gamma=3$, and $\kappa_*>\xi_-$. Suppose that $n\gtrsim s\log p$. If $\|\nabla\mathcal{L}_{\mathrm{init}}(\boldsymbol{\theta}_*)\|_\infty\leq\lambda/2$, we must have

$$\|\widehat{\boldsymbol{\theta}}_{\mathrm{init}}-\boldsymbol{\theta}_*\|_2\leq\lambda\sqrt{s},\ \text{ and }\ |\mathcal{A}_{\mathrm{init}}|\leq(C+1)s,$$

where C is a constant. If we further assume Assumption 4.3, then $\mathcal{S}\subset\mathcal{A}_{\mathrm{init}}$.

Proof of Lemma D.1. For the first part of this lemma, we can follow the usual analysis for lasso penalized regression, see, for example, the proof of Proposition 4.1 in Fan et al. (2017). This is because when $\kappa_*>\xi_-$, the concavity is dominated by the convexity and thus the problem becomes convex when restricted to the restricted ℓ_1-cone: $\|(\widehat{\boldsymbol{\theta}}_{\mathrm{init}})_{S^c}\|_1\leq 3\|(\widehat{\boldsymbol{\theta}}_{\mathrm{init}})_S\|_1$, which holds if $\|\nabla\mathcal{L}_{\mathrm{init}}(\boldsymbol{\theta}_*)\|_\infty\leq\lambda/2$. To prove $|\mathcal{A}_{\mathrm{init}}|\leq 2s$, that is, the nonconvex penalized estimator is sparse, similar arguments as in the proof of Lemma 5.4 in Fan et al. (2017) can be utilized. We do not repeat the steps here. Lastly, to prove $S\subset\mathcal{A}_{\mathrm{init}}$ under the minimal signal strength assumption, Assumption 4.3, we can basically follow the similar arguments in the proof of Theorem 2 in Loh & Wainwright (2017). □

Our second lemma gives a large probability bound for the even set $\left\{\|\mathcal{L}_{\mathrm{init}}(\boldsymbol{\theta}_*)\|_\infty\leq\lambda/2\right\}$, which is standard in the literature Negahban et al. (2012); Loh & Wainwright (2017); Fan et al. (2017), when the errors variables are sub-Gaussian distributed. We omit its proof here.

Lemma D.2. Suppose that $n \gtrsim s \log p$ and $\lambda \asymp \sqrt{\log p/n}$. We have

$$\mathbb{P}\big(\|\nabla \mathcal{L}_{\text{init}}(\boldsymbol{\theta}_*)\|_\infty \le \lambda/2\big) \ge 1 - 1/p.$$

Putting the above two lemmas together, we finish the proof of Proposition 4.4.

D.2 Proof of Lemma 4.5

Proof of Lemma 4.5. We have from Proposition 4.4 that $\mathcal{S} \subset \mathcal{A}_{\text{init}}$. Also from the Assumption 4.3, $\boldsymbol{\theta}^{(0)} = \widehat{\boldsymbol{\theta}}_{\text{init}}$ that $\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*\|_\infty \le \lambda$. By $\min_{j\in S} |\boldsymbol{\theta}_{*,j}| \ge 2\lambda$ and using the triangle inequality, we have for all $j \in S$ that

$$|\boldsymbol{\theta}_j^{(0)}| \ge |\boldsymbol{\theta}_{*,j}| - \left|\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_{*,j}\right| \ge 2\lambda - \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}_*\|_\infty \ge \lambda.$$

From the algorithm update (4.2) the following holds:

$$|\boldsymbol{\theta}_j^{(1)}| \begin{cases} \ge \lambda & \text{for } j \in \mathcal{A}_{\text{init}} \\ = 0 & \text{for } j \notin \mathcal{A}_{\text{init}} \end{cases}, \tag{D.1}$$

For $t \ge 1$, we turn to the SGD approximation and conclude for each $j = 1, \dots, d$, using Fernique's theorem (Da Prato & Zabczyk, 2014, §2.2) allows us to conclude that there are constants C, c depending on T such that

$$\mathbb{P}\left(\max_{0\le t\le T}\left|\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_{*,j}\right| \ge z\eta^{0.5}\right) \le Ce^{-cz^2}.$$

By setting $z = (2/c)^{1/2}\lambda\eta^{-0.5} = c'\sqrt{\log p}$ for a sufficiently large c', we have

$$\mathbb{P}\left(\max_{0\le t\le T}\left|\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_{*,j}\right| \ge z\eta^{0.5}\right) \le \frac{C}{p^2},$$

and the lemma is concluded.

□

D.3 Proof of Proposition 4.6

Proof of Proposition 4.6. Applying Proposition 2.5 to $\boldsymbol{\psi}^{(N)}$ we have

$$\mathbb{E}\|\boldsymbol{\psi}^{(N)} - \boldsymbol{\psi}_*\|^2 \lesssim \frac{\sigma^2}{\alpha_X} \cdot \frac{s\log N}{N},$$

and the proposition is concluded by the fact that $\mathbb{E}\|\boldsymbol{\theta}^{(N)} - \boldsymbol{\theta}_*\|^2 = \mathbb{E}\|\boldsymbol{\psi}^{(N)} - \boldsymbol{\psi}_*\|^2$.

□

E High-Dimensional Simulation Results

To conduct experiments for the high-dimensional case, we focus on the utility of truncation in the proposed algorithm. We fix the dimension to be $d = 1000$ and the intrinsic dimension to be $s = 10$, i.e. the true coefficient is sparse with 10 non-zero elements. As what we did in §5, we generate $n = 10000$ samples according to

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}_* + \varepsilon_i$$

where the regression coefficient $\boldsymbol{\theta}_* = (\underbrace{1, 1, ..., 1}_{10}, \underbrace{0, 0, ..., 0}_{990})^{\mathrm{T}}$, ε_i follows standard normal distribution and $\mathbf{x}_i$ follows multivariate distribution with mean zero and covariance matrix being identity matrix. We run the

stochastic gradient descent to derive an estimator $\widehat{\boldsymbol{\theta}}$ with step size $\eta_i \equiv \log n/n$ being a constant for all $1 \le i \le n$. Instead of starting from zero, the starting point of SGD comes from the extra $k = 50, 100, 150, 200$ and 500 burn-in samples. In the burn-in step, we apply LASSO to derive a sparse estimation of the coefficient and the tuning parameter is chosen by cross validation. As a controlled group, we also directly apply SGD on the steaming data without the burn-in process with the initial vector being zero. In all, we study 6 cases in this section with 100 repetitions and the mean ℓ_2 errors are recorded. The results are shown in Figure 2.

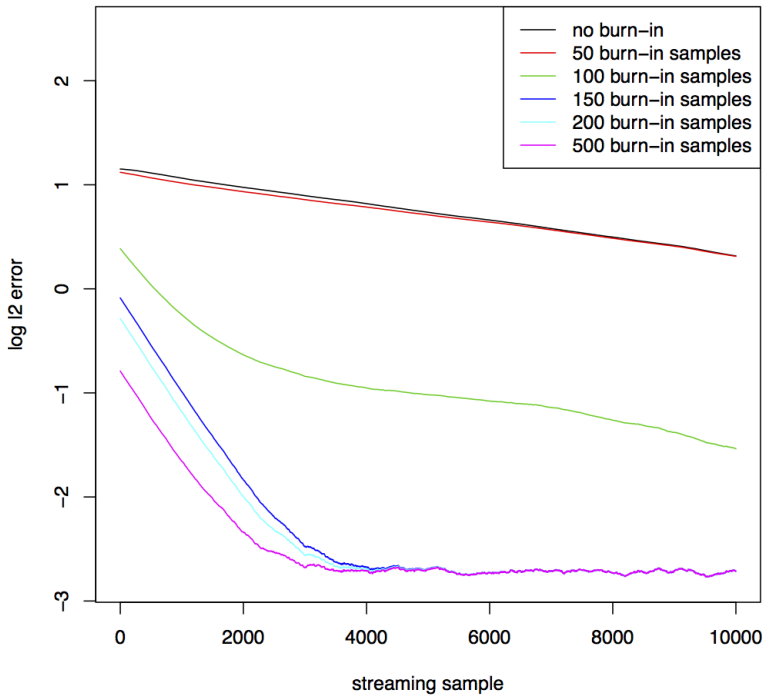

Figure 2: The ℓ_2 error in sparse online regression with different size of burn-in samples.

Generally speaking, a good initialization is very important for sparse online regression. When the burn-in sample size is large enough, i.e. the initial point is close to the true value, the algorithm would converge to the true coefficient; however, if the initialization is random or not close enough to the true value, the algorithm fails to do so. To deal with this problem, instead of truncate the estimator in every iteration, we can do the truncation periodically. In Figure 3, we illustrate the advantage of truncating the estimator every $k = 20, 50, 100$ times. We can see that when we truncate the estimator with a relatively long period, we are able to achieve consistent estimation.

F Further Remarks and Directions

Constant, small stepsizes For simplicity, we consider in this paper the regime where the stepsize is positively small. The choice of stepsize is arguably a central question of interest for many stochastic optimization problems. We adopt the converging stepsize so we can apply the techniques to high-dimensional case in §4, and also apply the tool of weak convergence to diffusions. After our initial submission, we found an updated version of the work Li et al. (2017) that proves a $\mathcal{O}(\eta)$ difference between the SGD and SDE processes up to $C\eta^{-1}$ steps. However, we require the time horizon to capture $C\eta^{-1}\log(\eta^{-1})$ steps when $\eta \to 0^+$, so the result does not apply to our case.

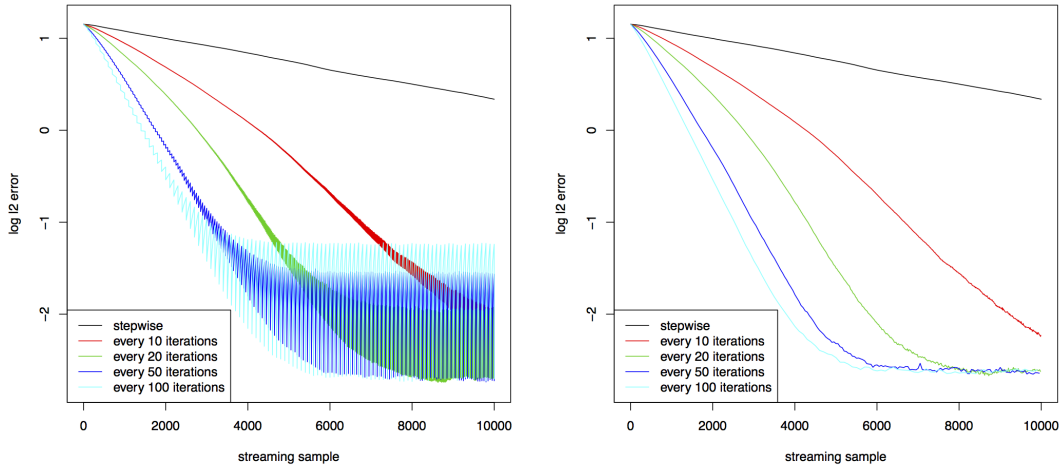

Figure 3: The ℓ_2 error in sparse online regression with different truncation period when there is no burn-in step. Left: the ℓ_2 error in each iteration; Right: the ℓ_2 error only in those iterations with truncation.

Knowledge of the minimum eigenvalue Proposition 2.5 and the Two-Phase Training Strategy require the knowledge of the minimum eigenvalue α_X of data matrix, which is usually not known in practice. To resolve this, we can first set the value to be a constant, say 1. Then we run the PCA iteration. If the iteration converges in equilibrium and performs oscillation early, it means our pick of α_X is small, then we increase it accordingly. If it does not even converge in observation, we decrease it accordingly. We can also run in parallel another iteration to estimate the minimal eigenvalue in an online fashion and adopt this α_X using a variant of power method (Li et al., 2018), which leads to another interesting direction and is omitted here, due to space limitation.

Knowledge of stream-size This can be conquered using an averaging scheme as in Proposition 3.1, we do not need to know the total streaming sample size in advance.

Knowledge of sparsity level The assumption of knowledge of sparsity level k_0 is known to truncate a specific number of nonzero coordinates. Without this assumption, there are no theoretical results. See e.g. Ma (2013); Langford et al. (2009); Yuan & Zhang (2013). However, in practice, we can always choose k_0 to be the support size estimated using the burn-in step.

Nonlinear framework We as a start using the diffusion approximation tools to analyzing the linear regression case, and will address the nonlinear case in later works.

Relaxation of assumption For Assumption 2.1(i), our assumption matches the online regression assumption in Langford et al. (2009). We believe the main results can hold for a relaxed version of this assumption. For Assumption 2.11(ii), in the heterogeneity case, we should expect similar behavior with more involved technical analysis, which is handled in the journal version of this article.