

---

# 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

$$y = \mathbf{x}^T \boldsymbol{\theta}_* + \varepsilon, \quad (1.1)$$

where  $y \in \mathbb{R}$  is the response variable,  $\mathbf{x} \in \mathbb{R}^p$  consists of  $p$  random predictor variables, and  $\varepsilon$  is the error term independent of  $\mathbf{x}$ . The target is to estimate parameter  $\boldsymbol{\theta}_*$  based on the data  $\{(\mathbf{x}_t, y_t) :$

$t = 1, 2, \dots, N\}$ , 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

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

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

$$\boldsymbol{\theta}_{\text{prim}} = (\mathbb{E}[\mathbf{x}\mathbf{x}^T])^{-1} \mathbb{E}[y\mathbf{x}]. \quad (1.3)$$

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

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

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  $\{(\mathbf{x}_t, y_t) : t = 1, 2, \dots\}$  streaming in one at a time, in a sequential manner. The stochastic gradient of the population risk in (1.2) is

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

Then the online stochastic gradient descent at time  $t$  performs

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

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  $\alpha_X$  is the smallest eigenvalue of  $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ . 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):

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

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

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

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  $\boldsymbol{\theta}_{\text{ols}}$  is known to be non-identifiable. One popular way to encourage sparsity is to consider the following penalized regression

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

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

$$\tilde{\boldsymbol{\theta}}^{(t)} = \text{Truncate} \left( \boldsymbol{\theta}^{(t-1)} + \eta(y_t - \mathbf{x}_t^T \boldsymbol{\theta}^{(t-1)}) \mathbf{x}_t, k_0 \right), \quad 2$$

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  $S = \text{supp}(\boldsymbol{\theta}_*)$ . 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  $\eta$  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  $\hat{\boldsymbol{\theta}}_{\text{init}}$ , 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; Nedić & 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  $\lambda$ . 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)  $\mathbb{E}[\mathbf{x}] = 0, \mathbb{E}[\varepsilon | \mathbf{x}] = 0$ ;
- (ii) (Homogeneity of variance)  $\text{var}[\varepsilon | \mathbf{x}] = \sigma^2$  for all  $\mathbf{x}$ ;
- (iii) (Absence of multicollinearity)  $\mathbf{\Sigma}_X \equiv \mathbb{E}[\mathbf{xx}^T]$  has its smallest eigenvalue  $\alpha_X > 0$ ;
- (iv) (Normalized predictors) the predictor variables are normalized with  $\text{var}(x_i) = 1$ , 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  $\{(\mathbf{x}_i, y_i), i \leq t\}$ . 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

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

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

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

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

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

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

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

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

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

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

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)

$$d\boldsymbol{\Theta}(t) = -\eta \boldsymbol{\Sigma}_X (\boldsymbol{\Theta}(t) - \boldsymbol{\theta}_*) dt + \eta \sigma \boldsymbol{\Sigma}_X^{1/2} d\mathbf{W}(t), \quad (2.5)$$

where  $\mathbf{W}(t)$  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  $\eta$  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  $\eta \rightarrow 0^+$ , 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

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

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

- (ii) As  $\eta \rightarrow 0^+$ , 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

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

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

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

$$(X_k^\eta(t_1), \dots, X_k^\eta(t_n)) \Rightarrow (X_k(t_1), \dots, X_k(t_n)).$$

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-Uhlenbeck process:

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

The last term above is a manifestation of Itô integral

that has mean zero and second moment:

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

Therefore,  $\Theta(t)$  traverses first deterministically and rapidly from initial point  $\theta_0$  and then to a neighborhood of  $\theta_*$ . Recall that  $\alpha_X > 0$  is the smallest eigenvalue of  $\mathbf{\Sigma}_X$ . 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  $\eta \asymp \alpha_X^{-1} \log N/N$ ,  $\Theta(t)$  in (2.8) satisfies

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

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  $\theta_0 = \mathcal{O}(\eta^{0.5})$ , the SGD iteration (1.4) is close to the global minimizer  $\theta_*$  and keeps oscillating around such minimizer. We can further improve the bound

using trajectory average:

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

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  $\theta^{(t)} \approx \Theta(t)$  for each  $t = 1, \dots, N$ , we have roughly that  $\bar{\theta}^{(N)} \approx \bar{\Theta}(N)$  where

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

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

**Proposition 3.1.** Suppose  $\theta_0 = \theta_* + \mathcal{O}_{\mathbb{P}}(\eta^{0.5})$ . As  $\eta N \rightarrow \infty$ ,  $\bar{\Theta}(N)$  weakly converges:

$$\sqrt{N}(\bar{\Theta}(N) - \theta_*) \Rightarrow \mathcal{N}(0, \sigma^2 \mathbf{\Sigma}_X^{-1}), \quad \text{and} \quad (3.1)$$

$$\mathbb{E} \|\bar{\Theta}(N) - \theta_*\|^2 \lesssim \sigma^2 \cdot \frac{\text{tr}(\mathbf{\Sigma}_X^{-1})}{N}. \quad (3.2)$$

Immediately from Proposition 3.1, we obtain the following corollary.

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

$$\sqrt{N} \mathbf{x}_0^T (\bar{\Theta}(N) - \theta_*) \Rightarrow \mathcal{N}(0, (\mathbf{x}_0^T \mathbf{\Sigma}_X^{-1} \mathbf{x}_0) \sigma^2), \quad \text{and} \quad (3.3)$$

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

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

$$\mathbb{E} |\mathbf{x}^T (\bar{\Theta}(N) - \theta_*)|^2 \lesssim \frac{p \sigma^2}{N}.$$

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

**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  $\mathbb{E}[F(\theta; \zeta)]$  with a Lipschitz gradient, by choosing appropriate step sizes,

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



where  $\mathbf{A} = \mathbb{E} [\nabla^2 F(\boldsymbol{\theta}_*; \zeta)]$  is the Hessian matrix, and  $\mathbf{S}^2 = \mathbb{E} [\nabla F(\boldsymbol{\theta}_*; \zeta) \nabla F(\boldsymbol{\theta}_*; \zeta)^\top]$  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  $\mathcal{N}(\boldsymbol{\theta}_*, (\sigma^2/N) \boldsymbol{\Sigma}_X^{-1})$ . For simplicity, we consider the case where the two phases have the same number of samples. Suppose  $N_{cvg} = N_{avg} \sim 0.5 \alpha_X^{-1} \eta^{-1} \log(\eta^{-1})$ . From Proposition 3.1, we know that it takes

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

iterates to obtain an estimator satisfying

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

Corollary 3.2 implies

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

The error bound in (3.4) is  $\mathcal{O}(\sigma^2 \text{tr}(\boldsymbol{\Sigma}_X^{-1}) \cdot N^{-1})$ , which improves the bound  $\mathcal{O}(\sigma^2 p \alpha_X^{-1} \cdot (\log N) N^{-1})$  without averaging step (2.9) in two ways: (i) the Hessian-related factor  $p \alpha_X^{-1}$  in (3.4) is improved to  $\text{tr}(\boldsymbol{\Sigma}_X^{-1})$ , and (ii) the  $\log N$  factor is eliminated. Moreover, the rate is independent of  $\eta$  under the asymptotic regime  $\eta N \rightarrow \infty$ . 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  $y = \mathbf{x}^\top \boldsymbol{\theta}_* + \varepsilon$ . We consider the setting that  $\boldsymbol{\theta}_*$  is  $s$ -sparse, that is, the support set of  $\boldsymbol{\theta}_*$ ,  $S = \text{supp}(\boldsymbol{\theta}_*)$ , 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  $\hat{\boldsymbol{\theta}}_{\text{init}}$  with active set  $\mathcal{A}_{\text{init}} = \text{supp}(\hat{\boldsymbol{\theta}}_{\text{init}})$ .
- 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  $\mathbf{X}_{\text{init}} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ . In the first offline stage, we run the following nonconvex penalized regression

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

where  $p_{\lambda_{\text{init}}}(\cdot)$  is a folded concave penalty with regularization parameter  $\lambda_{\text{init}}$ , such as SCAD (Fan & Li, 2001) or MCP (Zhang, 2010). One can adopt iLMM 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  $\nu$  and  $c$  be some appropriate constants. Suppose  $p_\lambda(x)$  can be written as  $p_\lambda(x) = \lambda|x| + h_\lambda(x)$  such that

- $p'_\lambda(x) = 0$ , for  $|x| \geq \nu \geq c\lambda > 0$ .
- $h'_\lambda(x)$  is monotone and Lipschitz continuous, i.e., for  $x' \geq x$ , there exists a constant  $\xi_- \geq 0$  such that  $h'_\lambda(x') - h'_\lambda(x) \leq -\xi_-(x' - x)$ .
- $h_\lambda(x)$  and  $h_\lambda(x)$  are symmetric and pass through 0.
- $|h'_\lambda(x)| \leq \lambda$  for any  $x \in \mathbb{R}$ .

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

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

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

where  $\mathcal{C}(m, \gamma) \equiv \{\mathbf{u} : S \subseteq J, |J| \leq m, \|\mathbf{u}_{J^c}\|_1 \leq \gamma \|\mathbf{u}_J\|_1\}$  is a local  $\ell_1$  cone.

We say the RE condition holds if there exists  $m$  and  $\gamma$  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  $\min_{j \in S} |\theta_{*,j}| \geq 2\lambda = 2c\sqrt{\log p/n}$ .

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  $\hat{\theta}_{\text{init}}$ , when using iLMM for computation. The readers are referred to [Fan et al. \(2017\)](#) for more information regarding to computation using iLMM.

**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

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

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  $\theta_{\text{init}}$  and run the following truncated stochastic gradient descent algorithm:

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

where  $k_0 = |\mathcal{A}_{\text{init}}| \geq s$  is the size of the support set of  $\theta_{\text{init}}$ . 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  $t \leq T$

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

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  $\theta^{(t-1)}$  is supported on  $\mathcal{A}_{\text{init}}$

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

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  $\psi^{(t)} = \theta^{(t)}|_{\mathcal{A}_{\text{init}}} \in \mathbb{R}^{k_0}$  and  $\psi_* = \theta_*|_{\mathcal{A}_{\text{init}}} \in \mathbb{R}^{k_0}$  be separately the projected iteration and minimizer. Then by projection onto  $\mathcal{A}_{\text{init}}$ , the problem is translated to a low-dimensional problem

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

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

$$d\Psi(t) = -\eta \Sigma_Z (\Psi(t) - \psi_*) dt + \eta \sigma \Sigma_Z^{1/2} d\mathbf{W}(t). \tag{4.3}$$

The error bound in terms of the  $\ell_2$ -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

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

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

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  $y_i = \mathbf{x}_i^T \boldsymbol{\theta}_* + \varepsilon_i$  where the regression coefficient  $\boldsymbol{\theta}_* = (1, 1, \dots, 1)^T$ ,  $\varepsilon_i$  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  $\hat{\boldsymbol{\theta}}$  with step size  $\eta_i \equiv \log N/N$  being a constant for all  $1 \leq i \leq N$ . The above procedure is repeated for 100 times and the average  $\ell_2$  error  $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|_2$  is recorded.

Meanwhile, we also start the averaging process from  $k = 2000, 5000, 8000$  and  $10000$  in each repetition and record the mean  $\ell_2$  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.

## 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

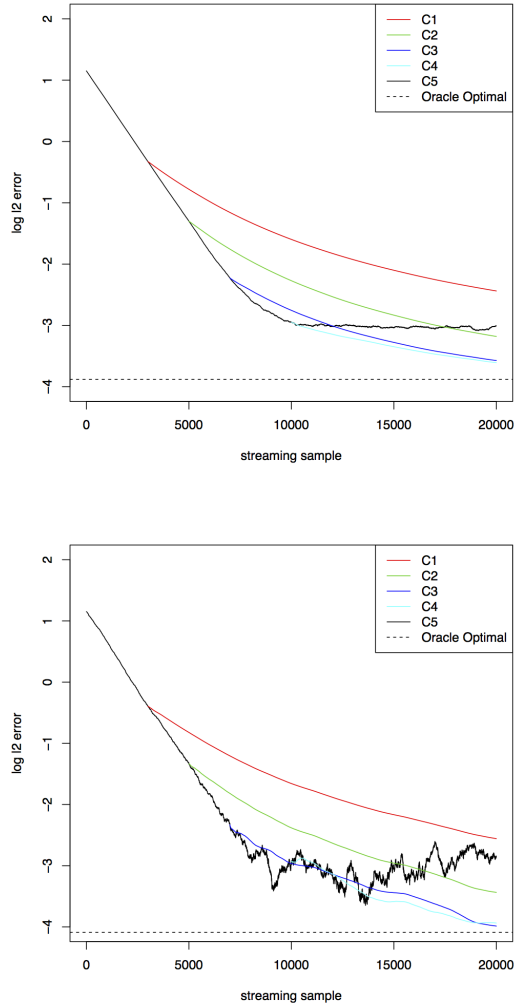


Figure 1: The  $\ell_2$  error in case C1 to C5 in the low-dimensional case. The dotted line is the oracle optimal error rate we can achieve when all samples are offline. Upper: the mean  $\ell_2$  error over 100 repetition; Lower: the  $\ell_2$  error in one repetition.

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 nonzeros 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 regularization.

- 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.