CTR D-STOP

Technical Report 159

Project Title:

Solving Perception Challenges for Autonomous Vehicles Using SGD

Research Supervisor: Constantine Caramanis
Wireless Networking and Communications Group

August 2020

Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin

CTR CENTER FOR TRANSPORTATION RESEARCH

WNCG Wireless Networking & Communications Group

D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

1. Report No.
D-STOP/2020/159

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Statistical Inference without Excess Data Using Only Stochastic Gradients: Volume 1

5. Report Date
August 2020

6. Performing Organization Code

7. Author(s)
Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis

8. Performing Organization Report No.
Report 159

9. Performing Organization Name and Address
Data-Supported Transportation Operations & Planning Center (D-STOP)
The University of Texas at Austin
3925 W. Braker Lane, 4th Floor
Austin, TX 78759

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DTRT13-G-UTC58

12. Sponsoring Agency Name and Address
United States Department of Transportation
University Transportation Centers
1200 New Jersey Avenue, SE
Washington, DC 20590

13. Type of Report and Period Covered

14. Sponsoring Agency Code

15. Supplementary Notes
Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program.
This is Volume 1; see Volume 2.

16. Abstract
We present a novel statistical inference framework for convex empirical risk minimization, using approximate stochastic Newton steps. The proposed algorithm is based on the notion of finite differences and allows the approximation of a Hessian-vector product from first-order information. In theory, our method efficiently computes the statistical error covariance in M-estimation, both for unregularized convex learning problems and high-dimensional LASSO regression, without using exact second order information, or resampling the entire data set. We also present a stochastic gradient sampling scheme for statistical inference in non-i.i.d. time series analysis, where we sample contiguous blocks of indices. In practice, we demonstrate the effectiveness of our framework on large-scale machine learning problems, that go even beyond convexity: as a highlight, our work can be used to detect certain adversarial attacks on neural networks.

17. Key Words
Statistical Inference; Frequentist Inference; M-estimation; High Dimensional Statistics; Time Series; Convex Optimization

18. Distribution Statement
No restrictions. This document is available to the public through NTIS (http://www.ntis.gov):
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

19. Security Classif.(of this report)
Unclassified

20. Security Classif.(of this page)
Unclassified

21. No. of Pages

22. Price

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers.

Proceedings of Machine Learning Research vol XX:1–66, 2020

Statistical Inference Without Excess Data Using Only Stochastic Gradients

Tianyang Li, and **Liu Liu** and **Anastasios Kyrillidis** and **Constantine Caramanis**

Abstract

We present a novel statistical inference framework for convex empirical risk minimization, using approximate stochastic Newton steps. The proposed algorithm is based on the notion of finite differences and allows the approximation of a Hessian-vector product from first-order information. In theory, our method efficiently computes the statistical error covariance in M-estimation, both for unregularized convex learning problems and high-dimensional LASSO regression, without using exact second order information, or resampling the entire data set. We also present a stochastic gradient sampling scheme for statistical inference in non-i.i.d. time series analysis, where we sample contiguous blocks of indices. In practice, we demonstrate the effectiveness of our framework on large-scale machine learning problems, that go even beyond convexity: as a highlight, our work can be used to detect certain adversarial attacks on neural networks.

Keywords: Statistical Inference; Frequentist Inference; M-estimation; High Dimensional Statistics; Time Series; Convex Optimization;

1. Introduction

Statistical inference is an important tool for assessing uncertainties, both for estimation and prediction purposes (Friedman et al., 2001; Efron and Hastie, 2016). *E.g.*, in unregularized linear regression and high-dimensional LASSO settings (van de Geer et al., 2014; Javanmard and Montanari, 2015; Tibshirani et al., 2015), we are interested in computing coordinate-wise confidence intervals and p-values of a p-dimensional variable, in order to infer which coordinates are active or not (Wasserman, 2013). Traditionally, the inverse Fisher information matrix (Edgeworth, 1908) contains the answer to such inference questions; however it requires storing and computing a $p \times p$ matrix structure, often prohibitive for large-scale applications (Tuerlinckx et al., 2006). Alternatively, the Bootstrap method is a popular statistical inference algorithm, where we solve an optimization problem per dataset replicate, but can be expensive for large data sets (Kleiner et al., 2014).

While optimization is mostly used for point estimates, recently it is also used as a means for statistical inference in large scale machine learning (Li et al., 2018; Chen et al., 2016; Su and Zhu, 2018; Fang et al., 2017). This manuscript follows this path: we propose an inference framework that uses stochastic gradients to approximate second-order, Newton steps. This is enabled by the fact that we only need to compute Hessian-vector products; in math, this can be approximated using $\nabla^2 f(\theta)v \approx \frac{\nabla f(\theta+\delta v)-\nabla f(\theta)}{\delta}$, where f is the objective function, and ∇f, $\nabla^2 f$ denote the gradient and Hessian of f. Our method can be interpreted as a generalization of the SVRG approach in optimization (Johnson and Zhang, 2013) (Appendix E); further, it is related to other stochastic Newton methods (e.g. (Agarwal et al., 2017)) when $\delta \to 0$. We defer the reader to Section 6 for more details. In this work, we apply our algorithm to unregularized M-estimation, and we use a similar approach, with proximal approximate Newton steps, in high-dimensional linear regression.

Our contributions can be summarized as follows; a more detailed discussion is deferred to Section 6:

- For the case of unregularized M-estimation, our method efficiently computes the statistical error covariance, useful for confidence intervals and p-values. Compared to state of the art, our scheme ***(i)*** guarantees consistency of computing the statistical error covariance, ***(ii)*** exploits better the available information (without wasting computational resources to compute quantities that are thereafter discarded), and ***(iii)*** converges to the optimum (without swaying around it).
- For high-dimensional linear regression, we propose a different estimator (see (13)) than the current literature. It is the result of a different optimization problem that is strongly convex with high probability. This permits the use of linearly convergent proximal algorithms (Xiao and Zhang, 2014; Lee et al., 2014) towards the optimum; in contrast, state of the art only guarantees convergence to a neighborhood of the LASSO solution within statistical error. Our model also does not assume that absolute values of the true parameter's non-zero entries are lower bounded.
- For statistical inference in non-i.i.d. time series analysis, we sample contiguous blocks of indices (instead of uniformly sampling) to compute stochastic gradients. This is similar to the Newey-West estimator (Newey and West, 1986) for HAC (heteroskedasticity and autocorrelation consistent) covariance estimation, but does not waste computational resources to compute the entire matrix.
- The effectiveness of our framework goes even beyond convexity. As a highlight, we show that our work can be used to detect certain adversarial attacks on neural networks.

2. Unregularized M-estimation

In unregularized, low-dimensional M-estimation problems, we estimate a parameter of interest:

$$\theta^\star = \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_{X\sim P}\left[\ell(X;\theta)\right], \quad \text{where } P(X) \text{ is the data distribution,}$$

using *empirical risk minimization* (ERM) on $n > p$ i.i.d. data points $\{X_i\}_{i=1}^n$:

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^p} \tfrac{1}{n}\sum_{i=1}^{n} \ell(X_i;\theta).$$

Statistical inference, such as computing one-dimensional confidence intervals, gives us information beyond the point estimate $\widehat{\theta}$, when $\widehat{\theta}$ has an asymptotic limit distribution (Wasserman, 2013). *E.g.*, under regularity conditions, the M-estimator satisfies asymptotic normality (van der Vaart, 1998, Theorem 5.21). *I.e.*, $\sqrt{n}(\widehat{\theta}-\theta^\star)$ weakly converges to a normal distribution:

$$\sqrt{n}\left(\widehat{\theta}-\theta^\star\right) \to \mathcal{N}\left(0, {H^\star}^{-1} G^\star {H^\star}^{-1}\right)$$

where $H^\star = \mathbb{E}_{X\sim P}[\nabla^2_\theta \ell(X;\theta^\star)]$ and $G^\star = \mathbb{E}_{X\sim P}[\nabla_\theta \ell(X;\theta^\star)\,\nabla_\theta \ell(X;\theta^\star)^\top]$. We can perform statistical inference when we have a good estimate of ${H^\star}^{-1} G^\star {H^\star}^{-1}$. In this work, we use the plug-in covariance estimator $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$ for ${H^\star}^{-1} G^\star {H^\star}^{-1}$, where:

$$\widehat{H} = \tfrac{1}{n}\sum_{i=1}^{n} \nabla^2_\theta \ell(X_i;\widehat{\theta}), \quad \text{and} \quad \widehat{G} = \tfrac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(X_i;\widehat{\theta})\,\nabla_\theta \ell(X_i;\widehat{\theta})^\top.$$

Observe that, in the naive case of directly computing $\widehat{G}$ and $\widehat{H}^{-1}$, we require both high computational- and space-complexity. Here, instead, we utilize approximate stochastic Newton motions from first order information to compute the quantity $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$.

2.1. Statistical inference with approximate Newton steps using only stochastic gradients

Based on the above, we are interested in solving the following p-dimensional optimization problem:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} f(\theta) := \frac{1}{n}\sum_{i=1}^{n} f_i(\theta), \quad \text{where } f_i(\theta) = \ell(X_i;\theta).$$

Notice that $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$ can be written as $\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{H}^{-1}\nabla_\theta \ell(X_i;\widehat{\theta})\right)\left(\widehat{H}^{-1}\nabla_\theta \ell(X_i;\widehat{\theta})\right)^\top$, which can be interpreted as the covariance of stochastic –inverse-Hessian conditioned– gradients at $\widehat{\theta}$. Thus, the covariance of stochastic Newton steps can be used for statistical inference.

Algorithm 1 Unregularized M-estimation statistical inference

```
1:  Parameters: $S_o, S_i \in \mathbb{Z}_+$; $\rho_0, \tau_0 \in \mathbb{R}_+$; $d_o, d_i \in \left(\frac{1}{2}, 1\right)$   Initial state: $\theta_0 \in \mathbb{R}^p$
2:  for $t = 0$ to $T-1$ do    // approximate stochastic Newton descent
3:      $\rho_t \leftarrow \rho_0 (t+1)^{-d_o}$
4:      $I_o \leftarrow$ uniformly sample $S_o$ indices with replacement from $[n]$
5:      $g_t^0 \leftarrow -\rho_t \left(\frac{1}{S_o}\sum_{i \in I_o} \nabla f_i(\theta_t)\right)$
6:      for $j = 0$ to $L-1$ do    // solving (1) approximately using SGD
7:          $\tau_j \leftarrow \tau_0 (j+1)^{-d_i}$ and $\delta_t^j \leftarrow O(\rho_t^4 \tau_j^4)$
8:          $I_i \leftarrow$ uniformly sample $S_i$ indices without replacement from $[n]$
9:          $g_t^{j+1} \leftarrow g_t^j - \tau_j \left(\frac{1}{S_i}\sum_{k \in I_i} \frac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j}\right) + \tau_j g_t^0$
10:     end for
11:     Use $\sqrt{S_o} \cdot \frac{\bar{g}_t}{\rho_t}$ for statistical inference, where $\bar{g}_t = \frac{1}{L+1}\sum_{j=0}^{L} g_t^j$
12:     $\theta_{t+1} \leftarrow \theta_t + g_t^L$
13: end for
```

Algorithm 1 approximates each stochastic Newton $\widehat{H}^{-1}\nabla_\theta \ell(X_i;\widehat{\theta})$ step using only first order information. We start from θ_0 which is sufficiently close to $\widehat{\theta}$, which can be effectively achieved using SVRG (Johnson and Zhang, 2013); a description of the SVRG algorithm can be found in Appendix E. Lines 4, 5 compute a stochastic gradient whose covariance is used as part of statistical inference. Lines 6 to 12 use SGD to solve the Newton step,

$$\min_{g \in \mathbb{R}^p} \left\langle \frac{1}{S_o}\sum_{i \in I_o} \nabla f_i(\theta_t), g \right\rangle + \frac{1}{2\rho_t}\left\langle g, \nabla^2 f(\theta_t) g \right\rangle, \tag{1}$$

which can be seen as a generalization of SVRG; this relationship is described in more detail in Appendix E. In particular, these lines correspond to solving (1) using SGD by uniformly sampling a random f_i, and approximating:

$$\nabla^2 f(\theta) g \approx \frac{\nabla f(\theta + \delta_t^j g) - \nabla f(\theta)}{\delta_t^j} = \mathbb{E}\left[\frac{\nabla f_i(\theta + \delta_t^j g) - \nabla f_i(\theta)}{\delta_t^j} \mid \theta\right] \tag{2}$$

Finally, the outer loop (lines 2 to 13) can be viewed as solving inverse Hessian conditioned stochastic gradient descent, similar to stochastic natural gradient descent (Amari, 1998).

In terms of parameters, similar to (Polyak and Juditsky, 1992; Ruppert, 1988), we use a decaying step size in Line 8 to control the error of approximating $H^{-1}g$. We set $\delta_t^j = O(\rho_t^4 \tau_j^4)$ to control the error of approximating Hessian vector product using a finite difference of gradients, so that it is smaller than the error of approximating $H^{-1}g$ using stochastic approximation. For similar reasons, we use a decaying step size in the outer loop to control the optimization error.

The following theorem characterizes the behavior of Algorithm 1.

Theorem 2.1 *For a twice continuously differentiable and convex function $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ where each f_i is also convex and twice continuously differentiable, assume f satisfies*

- ❑ *strong convexity: $\forall \theta_1, \theta_2$, $f(\theta_2) \geq f(\theta_1) + \langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle + \frac{1}{2}\alpha \|\theta_2 - \theta_1\|_2^2$;*
- ❑ *$\forall \theta$, each $\|\nabla^2 f_i(\theta)\|_2 \leq \beta_i$, which implies that f_i has Lipschitz gradient: $\forall \theta_1, \theta_2$, $\|\nabla f_i(\theta_1) - \nabla f_i(\theta_2)\|_2 \leq \beta_i \|\theta_1 - \theta_2\|_2$;*
- ❑ *each $\nabla^2 f_i$ is Lipschitz continuous: $\forall \theta_1, \theta_2$, $\|\nabla^2 f_i(\theta_2) - \nabla^2 f_i(\theta_1)\|_2 \leq h_i \|\theta_2 - \theta_1\|_2$.*

In Algorithm 1, we assume that batch sizes S_o—in the outer loop—and S_i—in the inner loops—are $O(1)$. The outer loop step size is

$$\rho_t = \rho_0 \cdot (t+1)^{-d_o}, \quad \text{where } d_o \in \left(\tfrac{1}{2}, 1\right) \text{ is the decaying rate.} \tag{3}$$

In each outer loop, the inner loop step size is

$$\tau_j = \tau_0 \cdot (j+1)^{-d_i}, \quad \text{where } d_i \in \left(\tfrac{1}{2}, 1\right) \text{ is the decaying rate.} \tag{4}$$

The scaling constant for Hessian vector product approximation is

$$\delta_t^j = \delta_0 \cdot \rho_t^4 \cdot \tau_j^4 = o\left(\frac{1}{(t+1)^2 (j+1)^2}\right) \tag{5}$$

Then, for the outer iterate θ_t we have

$$\mathbb{E}\left[\|\theta_t - \widehat{\theta}\|_2^2\right] \lesssim t^{-d_o}, \tag{6}$$

and

$$\mathbb{E}\left[\|\theta_t - \widehat{\theta}\|_2^4\right] \lesssim t^{-2d_o}. \tag{7}$$

In each outer loop, after L steps of the inner loop, we have:

$$\mathbb{E}\left[\left\|\frac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1} g_t^0\right\|_2^2 \mid \theta_t\right] \lesssim \frac{1}{L}\left\|g_t^0\right\|_2^2, \tag{8}$$

and at each step of the inner loop, we have:

$$\mathbb{E}\left[\left\|g_t^{j+1} - [\nabla^2 f(\theta_t)]^{-1} g_t^0\right\|_2^4 \mid \theta_t\right] \lesssim (j+1)^{-2d_i}\left\|g_t^0\right\|_2^4. \tag{9}$$

After T steps of the outer loop, we have a non-asymptotic bound on the "covariance":

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t \bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim T^{-\frac{d_o}{2}} + L^{-\frac{1}{2}}, \tag{10}$$

where $H = \nabla^2 f(\widehat{\theta})$ and $G = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widehat{\theta})\, \nabla f_i(\widehat{\theta})^\top$.

Some comments on the results in Theorem 2.1. The main outcome is that (10) provides a non-asymptotic bound and consistency guarantee for computing the estimator covariance using Algorithm 1. This is based on the bound for approximating the inverse-Hessian conditioned stochastic gradient in (8), and the optimization bound in (6). As a side note, the rates in Theorem 2.1 are very similar to classic results in stochastic approximation (Polyak and Juditsky, 1992; Ruppert, 1988); however the nested structure of outer and inner loops is different from standard stochastic approximation algorithms. Heuristically, calibration methods for parameter tuning in subsampling methods ((Efron and Tibshirani, 1994), Ch.18; (Politis et al., 2012), Ch. 9) can be used for hyper-parameter tuning in our algorithm.

In Algorithm 1, $\{{}^{\bar{g}_t}/_{\rho_t}\}_{i=1}^{n}$ does not have asymptotic normality. *I.e.*, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\bar{g}_t}{\rho_t}$ does not weakly converge to $\mathcal{N}\left(0, \frac{1}{S_o}H^{-1}GH^{-1}\right)$; we give an example using mean estimation in Appendix D.1. For a similar algorithm based on SVRG (Algorithm 6 in Appendix D), we show that we have asymptotic normality and improved bounds for the "covariance"; however, this requires a full gradient evaluation in each outer loop. In Appendix C, we present corollaries for the case where the iterations in the inner loop increase, as the counter in the outer loop increases (*i.e.*, $(L)_t$ is an increasing series). This guarantees consistency (convergence of the covariance estimate to $H^{-1}GH^{-1}$), although it is less efficient than using a constant number of inner loop iterations. Our procedure also serves as a general and flexible framework for using different stochastic gradient optimization algorithms (Toulis and Airoldi, 2017; Harikandeh et al., 2015; Loshchilov and Hutter, 2015; Daneshmand et al., 2016) in the inner and outer loop parts.

Finally, we present the following corollary that states that the average of consecutive iterates, in the outer loop, has asymptotic normality, similar to (Polyak and Juditsky, 1992; Ruppert, 1988).

Corollary 2.1 *In Algorithm 1's outer loop, the average of consecutive iterates satisfies*

$$\mathbb{E}\left[\left\|\frac{\sum_{t=1}^{T}\theta_t}{T}-\widehat{\theta}\right\|_2^2\right] \lesssim \frac{1}{T}, \qquad (11) \qquad \textit{and} \qquad \frac{1}{\sqrt{T}}\left(\frac{\sum_{t=1}^{T}\theta_t}{T}-\widehat{\theta}\right) = W+\Delta, \qquad (12)$$

where W weakly converges to $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$, and $\Delta = o_P(1)$ when $T \to \infty$ and $L \to \infty$ ($\mathbb{E}[\|\Delta\|_2^2] \lesssim T^{1-2d_o} + T^{d_o-1} + \frac{1}{L}$).

Corollary 2.1 uses 2nd , 4th moment bounds on individual iterates (eqs. (6), (7) in the above theorem), and the approximation of inverse Hessian conditioned stochastic gradient in (9).

3. High dimensional LASSO linear regression

In this section, we focus on the case of high-dimensional linear regression. Statistical inference in such settings, where $p \gg n$, is arguably a more difficult task: the bias introduced by the regularizer is of the same order with the estimator's variance. Recent works (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2015) propose statistical inference via de-biased LASSO estimators. Here, we present a new ℓ_1-norm regularized objective and propose an approximate stochastic *proximal* Newton algorithm, using only first order information.

We consider the linear model $y_i = \langle\theta^\star, x_i\rangle + \epsilon_i$, for some sparse $\theta^\star \in \mathbb{R}^p$. For each sample, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. noise. And each data point $x_i \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p$.

❏ *Assumptions on θ:* ***(i)*** $\theta^\star$ is s-sparse; ***(ii)*** $\|\theta^\star\|_2 = O(1)$, which implies that $\|\theta^\star\|_1 \lesssim \sqrt{s}$.

❏ *Assumptions on* Σ*:* ***(i)*** Σ is sparse, where each column (and row) has at most b non-zero entries;[1] ***(ii)*** Σ is well conditioned: all of Σ's eigenvalues are $\Theta(1)$; ***(iii)*** Σ is diagonally dominant ($\Sigma_{ii} - \sum_{j\neq i}|\Sigma_{ij}| \geq D_\Sigma > 0$ for all $1 \leq i \leq p$), and this will be used to bound the ℓ_∞ norm of $\widehat{S}^{-1}$ (Varah, 1975). A commonly used design covariance that satisfies all of our assumptions is I.

We estimate $\theta^\star$ using:

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2}\left\langle \theta, \left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\theta \right\rangle + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left(x_i^\top\theta - y_i\right)^2 + \lambda\|\theta\|_1, \tag{13}$$

where $\widehat{S}_{jk} = \operatorname{sign}\left(\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)_{jk}\right)\left(\left|\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)_{jk}\right| - \omega\right)_+$ is an estimate of Σ by soft-thresholding each element of $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ with $\omega = \Theta\left(\sqrt{\frac{\log p}{n}}\right)$ (Rothman et al., 2009). Under our assumptions, $\widehat{S}$ is positive definite with high probability when $n \gg b^2 \log p$ (Lemma F.3), and this guarantees that the optimization problem (13) is well defined. *I.e.*, we replace the degenerate Hessian in regular LASSO regression with an estimate, which is positive definite with high probability under our assumptions.

We set the regularization parameter

$$\lambda = \Theta\left((\sigma + \|\theta^\star\|_1)\sqrt{\frac{\log p}{n}}\right),$$

which is similar to LASSO regression (Bühlmann and van de Geer, 2011; Negahban et al., 2012) and related estimators using thresholded covariance (Yang et al., 2014; Jeng and Daye, 2011).

Point estimate. Theorem 3.1 provides guarantees for our proposed point estimate (13).

Theorem 3.1 *When* $n \gg b^2 \log p$*, the solution* $\widehat{\theta}$ *in* (13) *satisfies*

$$\left\|\widehat{\theta} - \theta^\star\right\|_1 \lesssim s\left(\sigma + \|\theta^\star\|_1\right)\sqrt{\frac{\log p}{n}} \lesssim s\left(\sigma + \sqrt{s}\right)\sqrt{\frac{\log p}{n}}, \tag{14}$$

$$\left\|\widehat{\theta} - \theta^\star\right\|_2 \lesssim \sqrt{s}\left(\sigma + \|\theta^\star\|_1\right)\sqrt{\frac{\log p}{n}} \lesssim \sqrt{s}\left(\sigma + \sqrt{s}\right)\sqrt{\frac{\log p}{n}}, \tag{15}$$

with probability at least $1 - p^{-\Theta(1)}$.

Confidence intervals. We next present a de-biased estimator $\widehat{\theta}^{\mathrm{d}}$ (16), based on our proposed estimator. $\widehat{\theta}^{\mathrm{d}}$ can be used to compute confidence intervals and p-values for each coordinate of $\widehat{\theta}^{\mathrm{d}}$, which can be used for false discovery rate control (Javanmard and Javadi, 2018). The estimator satisfies:

$$\widehat{\theta}^{\mathrm{d}} = \widehat{\theta} + \widehat{S}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^\top\widehat{\theta}\right)x_i\right]. \tag{16}$$

Our de-biased estimator is similar to (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014, 2015). however, we have different terms, since we need to de-bias covariance

[1] This is satisfied when Σ is block diagonal or banded. Covariance estimation under this sparsity assumption has been extensively studied (Bickel and Levina, 2008; Bickel et al., 2009; Cai and Zhou, 2012), and soft thresholding is an effective yet simple estimation method (Rothman et al., 2009).

estimation. Our estimator assumes $n \gg b^2 \log p$, since then $\widehat{S}$ is positive definite with high probability (Lemma F.3). The assumption that Σ is diagonally dominant guarantees that the ℓ_∞ norm $\|\widehat{S}^{-1}\|_\infty$ is bounded by $O\left(\frac{1}{D_\Sigma}\right)$ with high probability when $n \gg \frac{1}{{D_\Sigma}^2} \log p$.

Theorem 3.2 shows that we can compute valid confidence intervals for each coordinate when $n \gg (\frac{1}{D_\Sigma} s\,(\sigma + \|\theta^\star\|_1) \log p)^2$. This is satisfied when $n \gg (\frac{1}{D_\Sigma} s\,(\sigma + \sqrt{s}) \log p)^2$. And the covariance is similar to the sandwich estimator (Huber, 1967; White, 1980).

Theorem 3.2 *Under our assumptions, when $n \gg \max\{b^2, \frac{1}{{D_\Sigma}^2}\} \log p$, we have:*

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^\star) = Z + R, \tag{17}$$

where the conditional distribution satisfies $Z \mid \{x_i\}_{i=1}^n \sim \mathcal{N}\left(0, \sigma^2 \cdot \left[\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{S}^{-1}\right]\right)$, and $\|R\|_\infty \lesssim \frac{1}{D_\Sigma} s\,(\sigma + \|\theta^\star\|_1)\,\frac{\log p}{\sqrt{n}} \lesssim \frac{1}{D_\Sigma} s\,(\sigma + \sqrt{s})\,\frac{\log p}{\sqrt{n}}$ with probability at least $1 - p^{-\Theta(1)}$.

Our estimate in (13) has similar error rates to the estimator in (Yang et al., 2014); however, no confidence interval guarantees are provided, and the estimator is based on inverting a large covariance matrix. Further, although it does not match minimax rates achieved by regular LASSO regression (Raskutti et al., 2011), and the sample complexity in Theorem 3.2 is slightly higher than other methods (van de Geer et al., 2014; Javanmard and Montanari, 2014, 2015), our criterion is strongly convex with high probability: this allows us to use linearly convergent proximal algorithms (Xiao and Zhang, 2014; Lee et al., 2014), whereas provable linearly convergent optimization bounds for LASSO only guarantees convergence to a neighborhood of the LASSO solution within statistical error (Agarwal et al., 2010). This is crucial for computing the de-biased estimator, as we need the optimization error to be much less than the statistical error.

In Appendix A, we present our algorithm for statistical inference in high dimensional linear regression using stochastic gradients. It estimates the statistical error covariance using the plug-in estimator:

$$\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top\right)\widehat{S}^{-1},$$

which is related to the empirical sandwich estimator (Huber, 1967; White, 1980). Algorithm 2 computes the statistical error covariance. Similar to Algorithm 1, Algorithm 2 has an outer loop part and an inner loop part, where the outer loops correspond to approximate proximal Newton steps, and the inner loops solve each proximal Newton step using proximal SVRG (Xiao and Zhang, 2014). To control the variance, we use SVRG and proximal SVRG to solve the Newton steps. This is because in the high dimensional setting, the variance is too large when we use SGD (Moulines and Bach, 2011) and proximal SGD (Atchadé et al., 2017) for solving Newton steps. However, since we have $p \gg n$, instead of sampling *by sample*, we sample *by feature*. When we set $L_o^t = \Theta(\log(p) \cdot \log(t))$, we can estimate the statistical error covariance with element-wise error less than $O\left(\frac{\max\{1,\sigma\}\mathrm{polylog}(n,p)}{\sqrt{T}}\right)$ with high probability, using $O\left(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T)\right)$ numerical operations. And Algorithm 3 calculates the de-biased estimator $\widehat{\theta}^{\mathrm{d}}$ (16) via SVRG. For more details, we defer the reader to Appendix A.

4. Time series analysis

In this section, we present a sampling scheme for statistical inference in time series analysis using M-estimation, where we sample contiguous blocks of indices, instead of uniformly.

We consider a linear model $y_i = \langle x_i, \theta^\star \rangle + \epsilon_i$, where $\mathbb{E}[\epsilon_i x_i] = 0$, but $\{x_i, y_i\}_{i=1}^n$ may not be i.i.d. as this is a time series. And we use ordinary least squares (OLS) $\widehat{\theta} = \arg\min_\theta \sum_{i=1}^n \frac{1}{2}\left(\langle x_i, \theta \rangle - y_i\right)^2$ to estimate $\theta^\star$. Applications include multifactor financial models for explaining returns (Bender et al., 2013; Rosenberg and McKibben, 1973). For non-i.i.d. time series data, OLS may not be the optimal estimator, as opposed to the maximum likelihood estimator (Shumway and Stoffer, 2011), but OLS is simple yet often robust, compared to more sophisticated models that take into account time series dynamics. And it is widely used in econometrics for time series analysis (Berndt, 1991). To perform statistical inference, we use the asymptotic normality

$$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) \to \mathcal{N}\left(0, {H^\star}^{-1} G^\star {H^\star}^{-1}\right) \tag{18}$$

where $H^\star = \lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^n \nabla^2 f_i(\theta^\star)\right)$ and $G^\star = \lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^n \sum_{j=1}^n \nabla f_i(\theta^\star)\, \nabla f_j(\theta^\star)^\top\right)$, with $f_i(\theta) = \frac{1}{2}\left(\langle x_i, \theta\rangle - y_i\right)^2$. The difference compared with the i.i.d. case (Section 2) is that $G^\star$ now includes autocovariance terms. We use the plug-in estimate $\widehat{H} = \frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\widehat{\theta})$ as before, and we estimate $G^\star$ using the Newey-West covariance estimator (Newey and West, 1986) for HAC (heteroskedasticity and autocorrelation consistent) covariance estimation

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\widehat{\theta})\, f_i(\widehat{\theta})^\top + \sum_{j=1}^{\mathbf{l}} w(j,\mathbf{l}) \sum_{i=j+1}^n \left(\nabla f_i(\widehat{\theta})\, \nabla f_{i-j}(\widehat{\theta})^\top + \nabla f_{i-j}(\widehat{\theta})\, \nabla f_i(\widehat{\theta})^\top\right) \tag{19}$$

where $w(j,\mathbf{l})$ is sample autocovariance weight, such as Bartlett weight $w(j,\mathbf{l}) = 1 - {}^{j}/_{(\mathbf{l}+1)}$ (Bartlett, 1946), and $\mathbf{l}$ is the *lag* parameter, which captures data dependence across time. Note that this is an essential building block in time series statistical inference procedures, such as Driscoll-Kraay standard errors (Driscoll and Kraay, 1998; Kraay and Driscoll, 1999), moving block bootstrap (Kunsch, 1989), and circular bootstrap (Politis and Romano, 1992, 1994).

In our framework, we solve OLS using our approximate Newton procedure with a slight modification to Algorithm 1. Instead of uniformly sampling indices as in line 4 of Algorithm 1, we uniformly select some $i_o \in [n]$, and set the outer mini-batch indexes I_o to the random contiguous block $\{i_o, i_o+1, \ldots, i_o+\mathbf{l}-1\} \mod n$, where we let the indexes circularly wrap around, as in line 4 of Algorithm 5, and this sampling scheme is similar to *circular bootstrap*. Here $\mathbf{l}$ is the lag parameter, similar to the Newey-West estimator. And the stochastic gradient's expectation is still the full gradient. The complete algorithm is in Algorithm 5, and its guarantees are given in Corollary B.1. Our approximate Newton statistical inference procedure is equivalent to using weight $w(j,\mathbf{l}) = 1 - {}^{j}/_{\mathbf{l}}$ in the Newey-West covariance estimator (19), with negligible terms for blocks that wrap around, and this is the same as circular bootstrap. Note that the connection between sampling scheme and Newey-West estimator was also observed in (Kunsch, 1989). Following (Politis and Romano, 1992), we can set the lag parameter such that $\mathbf{l} \cdot n^{-1/3} \to 0$, and run at least n outer loops. In practice, other methods for tuning the lag parameter can be used, such as (Newey and West, 1994). For more details, we refer the reader to Appendix B.

	Approximate Newton	Bootstrap	Inverse Fisher information	Averaged SGD
Lin1	(0.906, 0.289)	(0.933, 0.294)	(0.918, 0.274)	(0.458, 0.094)
Lin2	(0.915, 0.321)	(0.942, 0.332)	(0.921,0.308)	(0.455 0.103)

Table 1: Linear regression (low dimensional): synthetic data confidence interval (coverage, length)

	Approximate Newton	Jackknife	Inverse Fisher information	Averaged SGD
Log1	(0.902, 0.840)	(0.966 1.018)	(0.938, 0.892)	(0.075 0.044)
Log2	(0.925, 1.006)	(0.979, 1.167)	(0.948, 1.025)	(0.065 0.045)

Table 2: Logistic regression (low dimensional): synthetic data confidence interval (coverage, length)

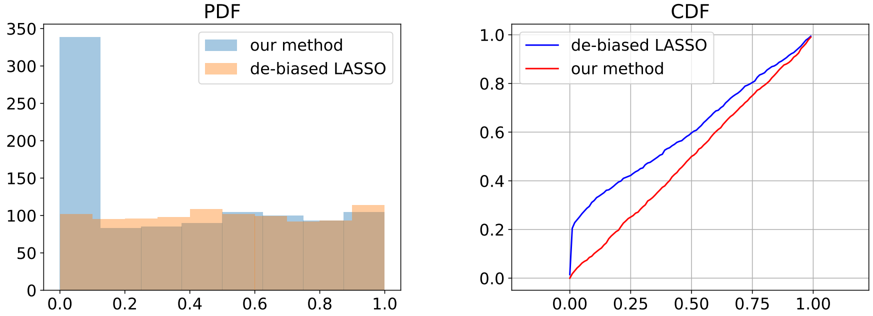

Figure 1: Distribution of two-sided Z-test p-values under the null hypothesis (high dimensional)

5. Experiments

5.1. Synthetic data

The coverage probability is defined as $\frac{1}{p}\sum_{i=1}^{p}\mathbb{P}[\theta_i^\star \in \hat{C}_i]$, where $\hat{C}_i$ is the estimated confidence interval for the i^{th} coordinate. The average confidence interval length is defined as $\frac{1}{p}\sum_{i=1}^{p}(\hat{C}_i^u - \hat{C}_i^l)$, where $[\hat{C}_i^l, \hat{C}_i^u]$ is the estimated confidence interval for the i^{th} coordinate. In our experiments, coverage probability and average confidence interval length are estimated through simulation. Result given as a (α, β) indicates (coverage probability, confidence interval length).

Low dimensional problems. Table 1 and Table 2 show 95% confidence interval's coverage and length of 200 simulations for linear and logistic regression. The exact configurations for linear/logistic regression examples are provided in Appendix H.1.1. Compared with Bootstrap and Jackknife (Efron and Tibshirani, 1994), Algorithm 1 uses less numerical operations, while achieving similar results. Compared with the averaged SGD method (Li et al., 2018; Chen et al., 2016), our algorithm performs much better, while using the same amount of computation, and is much less sensitive to the choice hyper-parameters. And we observe that calibrated approximate Newton confidence intervals (Efron and Tibshirani, 1994; Politis et al., 2012) are better than bootstrap and inverse Fisher information (Table 3).

High dimensional linear regression. Figure 1 shows p-value distribution under the null hypothesis for our method and the de-biased LASSO estimator with known covariance, using 600 i.i.d. samples generated from a model with $\Sigma = I$, $\sigma = 0.7$, and we can see that it is close to a uniform distribution, similar results are observed for other high dimensional statistical inference procedures such as (Candes et al., 2018). And visualization of confidence intervals computed by our algorithm is shown in Figure 3.

Time series analysis. In our linear regression simulation, we generate i.i.d. random explanatory variables, and the observation noise is a 0-mean moving average (MA) process independent of the explanatory variables. Results on average 95% confidence interval coverage and length are given in Appendix H.1.3, and they validate our theory.

5.2. Real data

Neural network adversarial attack detection. Here we use ideas from statistical inference to detect certain adversarial attacks on neural networks. A key observation is that neural networks are effective at representing low dimensional manifolds such as natural images (Basri and Jacobs, 2016; Chui and Mhaskar, 2016), and this causes the risk function's Hessian to be degenerate (Sagun et al., 2017). From a statistical inference perspective, we interpret this as meaning that the confidence intervals in the null space of $H^{+}GH^{+}$ is infinity, where H^{+} is the pseudo-inverse of the Hessian (see Section 2). When we make a prediction $\Psi(x;\widehat{\theta})$ using a fixed data point x as input (i.e., conditioned on x), using the delta method (van der Vaart, 1998), the confidence interval of the prediction can be derived from the asymptotic normality of $\Psi(x;\widehat{\theta})$

$$\sqrt{n}\left(\Psi(x;\widehat{\theta}) - \Psi(x;\theta^{\star})\right) \to \mathcal{N}\left(0, \nabla_\theta \Psi(x;\widehat{\theta})^\top \left[\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}\right] \nabla_\theta \Psi(x;\widehat{\theta})\right)$$

To detect adversarial attacks, we use the score

$$\frac{\left\|\left(I - P_{H^{+}GH^{+}}\right)\nabla_\theta \Psi(x;\widehat{\theta})\right\|_2}{\left\|\nabla_\theta \Psi(x;\widehat{\theta})\right\|_2},$$

to measure how much $\nabla_\theta \Psi(x;\widehat{\theta})$ lies in null space of $H^{+}GH^{+}$, where $P_{H^{+}GH^{+}}$ is the projection matrix onto the range of $H^{+}GH^{+}$. Conceptually, for the same image, the randomly perturbed image's score should be larger than the original image's score, and the adversarial image's score should be larger than the randomly perturbed image's score.

We train a binary classification neural network with 1 hidden layer and softplus activation function, to distinguish between "Shirt" and "T-shirt/top" in the Fashion MNIST data set (Xiao et al., 2017). Figure 2 shows distributions of scores of original images, adversarial images generated using the fast gradient sign method (Goodfellow et al., 2014), and randomly perturbed images. Adversarial and random perturbations have the same ℓ_∞ norm. The adversarial perturbations and example images are shown in Appendix H.2.1. Although the scores' values are small, they are still significantly larger than 64-bit floating point precision ($2^{-53} \approx 1.11 \times 10^{-16}$). We observe that scores of randomly perturbed images is an order of magnitude larger than scores of original images, and scores of adversarial images is an order of magnitude larger than scores of randomly perturbed images.

High dimensional linear regression. We apply our high dimensional inference procedure to the dataset in (Rhee et al., 2006) to detect mutations related to HIV drug resistance, where we randomly sub-sample the dataset so that the number of features is larger than the number of samples. When we control the family-wise error rate (FWER) at 0.05 using the Bonferroni correction (Bonferroni,

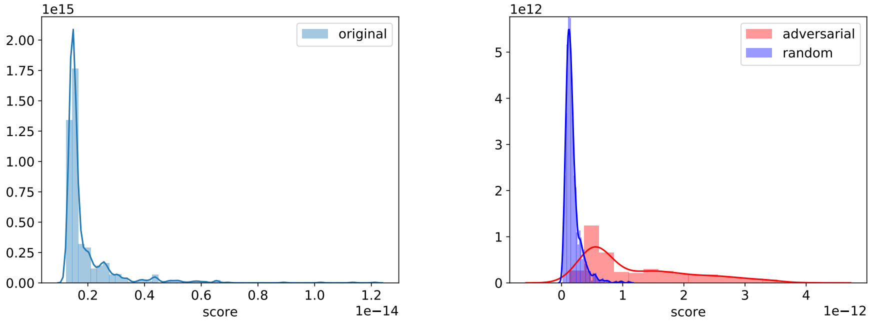

Figure 2: Distribution of scores for original, randomly perturbed, and adversarially perturbed images

1936), our procedure is able to detect verified mutations in an expert dataset (Johnson et al., 2005) (Table 4), and the details are given in Appendix H.2.2. Another experiment with a genomic data set concerning riboflavin (vitamin B2) production rate (Bühlmann et al., 2014) is given in the appendix.

Time series analysis. Using monthly equities returns data from (Frazzini and Pedersen, 2014), we use our approximate Newton statistical inference procedure to show that the correlation between US equities market returns and non-US global equities market returns is statistically significant, which validates the capital asset pricing model (CAPM) (Sharpe, 1964; Lintner, 1965; Fama and French, 2004). The details are given in Appendix H.2.3.

6. Related work

Unregularized M-estimation. This work provides a general, flexible framework for *simultaneous* point estimation and statistical inference, and improves upon previous methods, based on averaged stochastic gradient descent (Li et al., 2018; Chen et al., 2016).

Compared to (Chen et al., 2016) (and similar works (Su and Zhu, 2018; Fang et al., 2017) using SGD with decreasing step size), our method does not need to increase the lengths of “segments” (inner loops) to reduce correlations between different “replicates”. Even in that case, if we use T replicates and increasing “segment” length (number of inner loops is $t^{\frac{d_o}{1-d_o}} \cdot L$) with a total of $O(T^{\frac{1}{1-d_o}} \cdot L)$ stochastic gradient steps, (Chen et al., 2016) guarantees $O(L^{-\frac{1-d_o}{2}} + T^{-\frac{1}{2}} + T^{\max\{\frac{1}{2}-\frac{d_o}{4(1-d_o)},0\}-\frac{1}{2}} \cdot L^{-\frac{d_o}{4}} + T^{\max\{\frac{1-2d_o}{2(1-d_o)},0\}-\frac{1}{2}} \cdot L^{\frac{1-2d_o}{2}})$, whereas our method guarantees $O(T^{-\frac{d_o}{2}})$. Further, (Chen et al., 2016) is inconsistent, whereas our scheme guarantees consistency of computing the statistical error covariance.

(Li et al., 2018) uses fixed step size SGD for statistical inference, and discards iterates between different “segments” to reduce correlation, whereas we do not discard any iterates in our computations. Although (Li et al., 2018) states empirically constant step SGD performs well in statistical inference, it has been empirically shown (Dieuleveut et al., 2017) that averaging consecutive iterates in constant step SGD does not guarantee convergence to the optimal – the average will be “wobbling” around the optimal, whereas decreasing step size stochastic approximation methods ((Polyak and Juditsky, 1992; Ruppert, 1988) and our work) will converge to the optimal, and averaging consecutive iterates guarantees “fast” rates.

Finally, from an optimization perspective, our method is similar to stochastic Newton methods (e.g. (Agarwal et al., 2017)); however, our method only uses first-order information to approximate a Hessian vector product ($\nabla^2 f(\theta) v \approx \frac{\nabla f(\theta+\delta v)-\nabla f(\theta)}{\delta}$). Algorithm 1's outer loops are similar to stochastic natural gradient descent (Amari, 1998). Also, we demonstrate an intuitive view of SVRG (Johnson and Zhang, 2013) as a special case of approximate stochastic Newton steps using first order information (Appendix E).

High dimensional linear regression. (Chen et al., 2016)'s high dimensional inference algorithm is based on (Agarwal et al., 2012), and only guarantees that optimization error is at the same scale as the statistical error. However, proper de-biasing of the LASSO estimator requires the optimization error to be much less than the statistical error, otherwise the optimization error introduces additional bias that de-biasing cannot handle. Our optimization objective is strongly convex with high probability: this permits the use of linearly convergent proximal algorithms (Xiao and Zhang, 2014; Lee et al., 2014) towards the optimum, which guarantees the optimization error to be much smaller than the statistical error.

Our method of de-biasing the LASSO in Section 3 is similar to (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014, 2015). Our method uses a new ℓ_1 regularized objective (13) for high dimensional linear regression, and we have different de-biasing terms, because we also need to de-bias the covariance estimation. In Algorithm 2, our covariance estimate is similar to the classic *sandwich estimator* (Huber, 1967; White, 1980). Previous methods require $O(p^2)$ space which unsuitable for large scale problems, whereas our method only requires $O(p)$ space.

Similar to our ℓ_1-norm regularized objective, (Yang et al., 2014; Jeng and Daye, 2011) shows similar point estimate statistical guarantees for related estimators; however there are no confidence interval results. Further, although (Yang et al., 2014) is an elementary estimator in closed form, it still requires computing the inverse of the thresholded covariance, which is challenging in high dimensions, and may not computationally outperform optimization approaches.

Finally, for feature selection, we do not assume that absolute values of the true parameter's non-zero entries are lower bounded ("beta-min" condition). (Fan et al., 2018; Loh, 2017; Loh and Wainwright, 2017; Bühlmann and van de Geer, 2011; Wainwright, 2009).

Time series analysis. Our approach of sampling contiguous blocks of indices to compute stochastic gradients for statistical inference in time series analysis is similar to resampling procedures in *moving block* or *circular* bootstrap (Carlstein, 1986; Kunsch, 1989; Bühlmann, 2002; Davison and Hinkley, 1997; Efron and Tibshirani, 1994; Lahiri, 2013; Politis and Romano, 1992, 1994; Kreiss and Lahiri, 2012), and *conformal prediction* (Balasubramanian et al., 2014; Shafer and Vovk, 2008; Vovk et al., 2005). Also, our procedure is similar to Driscoll-Kraay standard errors (Driscoll and Kraay, 1998; Kraay and Driscoll, 1999; Hoechle, 2007), but does not waste computational resources to explicitly store entire matrices, and is suited for large scale time series analysis.

References

Alekh Agarwal, Sahand Negahban, and Martin Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems*, pages 1538–1546, 2012.

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res*, 18(1):310–342, 2017.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Maurice Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41, 1946.

Ronen Basri and David Jacobs. Efficient representation of low-dimensional manifolds using deep networks. *arXiv preprint arXiv:1602.04723*, 2016.

Jennifer Bender, Remy Briand, Dimitris Melas, and Raman Subramanian. Foundations of factor investing. 2013.

Ernst Berndt. *The practice of econometrics: classic and contemporary*. Addison Wesley Publishing Company, 1991.

Peter Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

Peter Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Peter Bühlmann. Bootstraps for time series. *Statistical science*, pages 52–72, 2002.

Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

Tony Cai and Harrison Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, pages 2389–2420, 2012.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:model-xknockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14(3):1171–1179, 1986.

Xi Chen, Jason Lee, Xin Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

Charles Chui and Hrushikesh Narhar Mhaskar. Deep nets for local manifold learning. *arXiv preprint arXiv:1607.07110*, 2016.

Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Starting small-learning with adaptive sample sizes. In *International conference on machine learning*, pages 1463–1471, 2016.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *arXiv preprint arXiv:1707.06386*, 2017.

Jürgen Dippon. Asymptotic expansions of the Robbins-Monro process. *Mathematical Methods of Statistics*, 17(2):138–145, 2008a.

Jürgen Dippon. Edgeworth expansions for stochastic approximation theory. *Mathematical Methods of Statistics*, 17(1):44–65, 2008b.

John Driscoll and Aart Kraay. Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 80(4):549–560, 1998.

Francis Ysidro Edgeworth. On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2):381–397, 1908.

B. Efron and T. Hastie. *Computer age statistical inference*. Cambridge University Press, 2016.

Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Eugene F. Fama and Kenneth R. French. The capital asset pricing model: Theory and evidence. *Journal of economic perspectives*, 18(3):25–46, 2004.

Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1017–1026, 2018.

Yixin Fang, Jinfeng Xu, and Lei Yang. On Scalable Inference with Stochastic Gradient Descent. *arXiv preprint arXiv:1707.00192*, 2017.

Andrea Frazzini and Lasse Heje Pedersen. Betting against beta. *Journal of Financial Economics*, 111(1):1–25, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Reza Harikandeh, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt, Jakub Konecny, and Scott Sallinen. StopWasting My Gradients: Practical SVRG. In *Advances in Neural Information Processing Systems 28*, pages 2251–2259, 2015.

Daniel Hoechle. Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal*, 7(3):281–312, 2007.

Peter Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability*, pages 221–233, Berkeley, CA, 1967. University of California Press. ISBN 0097-0433.

Adel Javanmard and Hamid Javadi. False Discovery Rate Control via Debiased Lasso. *arXiv preprint arXiv:1803.04464*, 2018.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Adel Javanmard and Andrea Montanari. De-biasing the Lasso: Optimal sample size for Gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

Jessie Jeng and John Daye. Sparse covariance thresholding for high-dimensional variable selection. *Statistica Sinica*, pages 625–657, 2011.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

Victoria A Johnson, Francoise Brun-Vezinet, Bonaventura Clotet, Brian Conway, Daniel R. Kuritzkes, Deenan Pillay, Jonathan Schapiro, Amalio Telenti, and Douglas Richman. Update of the drug resistance mutations in HIV-1: 2005. *Topics in HIV medicine: a publication of the International AIDS Society, USA*, 13(1):51–57, 2005.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4): 795–816, 2014.

Aart Kraay and John Driscoll. *Spatial Correlations in Panel Data*. The World Bank, 1999.

Jens-Peter Kreiss and Soumendra Nath Lahiri. Bootstrap methods for time series. In *Handbook of statistics*, volume 30, pages 3–26. Elsevier, 2012.

Hans Kunsch. The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241, 1989.

Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2013.

Jason Lee, Yuekai Sun, and Michael Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using SGD. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, pages 13–37, 1965.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

Po-Ling Loh and Martin Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Sahand Negahban, Pradeep Ravikumar, Martin Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, pages 538–557, 2012.

Whitney Newey and Kenneth West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *National Bureau of Economic Research Cambridge, Mass., USA*, 1986.

Whitney Newey and Kenneth West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer New York, 2012. ISBN 9781461215547.

Dimitris Politis and Joseph Romano. A circular block-resampling procedure for stationary data. *Exploring the limits of bootstrap*, pages 263–270, 1992.

Dimitris Politis and Joseph Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Garvesh Raskutti, Martin Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L. Brutlag, and Robert W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

Barr Rosenberg and Walt McKibben. The prediction of systematic and specific risk in common stocks. *Journal of Financial and Quantitative Analysis*, 8(2):317–333, 1973.

Adam Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. *arXiv preprint arXiv:1706.04454*, 2017.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

Robert Shumway and David Stoffer. Time series regression and exploratory data analysis. In *Time series analysis and its applications*, pages 47–82. Springer, 2011.

Weijie Su and Yuancheng Zhu. Statistical Inference for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent. *arXiv preprint arXiv:1802.04876*, 2018.

R. Tibshirani, M. Wainwright, and T. Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Joel Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

F. Tuerlinckx, F. Rijmen, G. Verbeke, and P. Boeck. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255, 2006.

Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Aad W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

James Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world: conformal prediction*. Springer, 2005.

Martin Wainwright. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. To appear, 2017.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *International Conference on Machine Learning*, pages 388–396, 2014.

Cun-Hui Zhang and Stephanie Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.