# Online Statistical Inference of Constrained Stochastic Optimization via Random Scaling

**Xinchen Du**[1]**, Wanrong Zhu**[2]**, Wei Biao Wu**[3]**, Sen Na**[1]
1. School of Industrial and Systems Engineering, Georgia Institute of Technology
2. Department of Statistics, University of California, Irvine
3. Department of Statistics, The University of Chicago

## Abstract

Constrained stochastic nonlinear optimization problems have attracted significant attention for their ability to model complex machine learning phenomena. As datasets continue to grow, online inference methods have become crucial for enabling real-time decision-making without the need to store historical data. In this work, we develop an online inference procedure for constrained stochastic optimization by leveraging a method called Adaptive Inexact Stochastic Sequential Quadratic Programming (AI-SSQP), which can be considered as a generalization of (sketched) Newton methods to constrained problems. We first establish the asymptotic normality of *averaged* AI-SSQP iterates. Then we propose a *random scaling* method that constructs parameter-free pivotal statistics through appropriate normalization. Our online inference approach offers two key advantages: (i) it enables the construction of *asymptotically valid* and *statistically efficient* confidence intervals, while existing work based on last iterates are less efficient and rely on a covariance estimator that is inconsistent; and (ii) it is *matrix-free*, i.e., the computation involves only primal-dual iterates without any matrix inversions, making its computational cost match that of first-order methods for unconstrained problems.

## 1 Introduction

We consider the following constrained stochastic optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \;\; f(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{P}}[F(\boldsymbol{x};\xi)], \quad \text{s.t. } \; c(\boldsymbol{x}) = \mathbf{0}, \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is the objective function, $F(\boldsymbol{x},\xi)$ is a noisy measurement of $f(\boldsymbol{x})$ with the randomness $\xi$ following distribution $\mathcal{P}$, and $c:\mathbb{R}^d\to\mathbb{R}^m$ denotes the equality constraints. Problem (1) arises in various real-world applications, such as portfolio allocation [15, 13], PDE-constrained optimization [30, 20], constrained deep neural network [6], and physics-informed neural networks [9].

In recent years, a growing body of literature has focused on developing *online* methods for constrained stochastic optimization problems, motivated by their computational and memory efficiency compared to offline methods. One simple approach is projection-based Stochastic Gradient Descent (SGD). [11] proved almost sure convergence for projection-based SGD, while [14] and [12] established asymptotic normality for its iterates. Building on this distributional result, [17] proposed a batch-means estimator of the limiting covariance to enable practical statistical inference. However, the projection operator, which plays a key role in these methods, can be computationally expensive in practice.

To avoid computing projection operators, various SSQP methods, which generalize stochastic Newton methods for unconstrained problems, have been proposed to efficiently solve constrained problems in an online fashion; see [3, 4, 26, 27, 16, 10] and references therein. For the purpose of statistical inference, [25] considered an Adaptive Inexact SSQP (AI-SSQP) scheme, where at each step

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

the method employs an iterative sketching solver to inexactly solve the SQP subproblem and selects a suitable random stepsize (inspired by stochastic line search) to update the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$. The authors showed that AI-SSQP iterates exhibit the asymptotic normality [25, Theorem 5.6]:

$$\sqrt{1/\bar{\alpha}_t} \cdot (\boldsymbol{x}_t - \boldsymbol{x}^\star, \boldsymbol{\lambda}_t - \boldsymbol{\lambda}^\star) \xrightarrow{d} \mathcal{N}(0, \tilde{\Xi}^\star), \tag{2}$$

where $\bar{\alpha}_t$ is the random stepsize and the limiting covariance matrix $\tilde{\Xi}^\star$ (given by (6)) is determined by the underlying sketching distribution. To construct confidence intervals, [25] introduced a *plug-in* covariance estimator for $\tilde{\Xi}^\star$. However, due to the difficulty of estimating sketching components ($C^\star$, $\tilde{C}^\star$) in $\tilde{\Xi}^\star$, their estimator set all such components to be zero. Consequently, the resulting covariance estimator is generally inconsistent, which significantly affects the performance of practical inference.

**Main contribution.** In this paper, we overcome these challenges by using a technique called *random scaling*. We first establish the asymptotic normality of the averaged AI-SSQP iterates, with the limiting covariance differing from and proved to be smaller than $\tilde{\Xi}^\star$ in (2). Furthermore, to perform online statistical inference, rather than pursuing direct covariance estimation, we propose a *random scaling* matrix that could be computed in an online fashion. As a result, we use the matrix to construct an *asymptotically pivotal* statistic, meaning that its limiting distribution is free of any unknown parameters. Our inference approach offers three key advantages. i) It is well-suited for online computation. ii) It enables the construction of asymptotically valid confidence intervals thanks to the pivotal statistic, hence resolving the inconsistency of the plug-in estimator in [25]. iii) Our pivotal statistic is matrix-free (i.e., no matrix inversion). Therefore, the memory and computational complexities of our inference procedure match those of first-order methods, i.e., $O((d+m)^2)$.

**Related literature on unconstrained online inference.** Online inference for unconstrained stochastic optimization serves a fundamental role in our study. Without the constraint $c(\boldsymbol{x}) = \boldsymbol{0}$, problem (1) reduces to the unconstrained optimization problem, where the SGD algorithm emerges as the natural approach. [31] established the almost sure convergence of SGD, and [29] showed the asymptotic normality of the averaged iterates. Based on the limiting distribution, [7, 33] designed online covariance estimators for the limiting covariance. To avoid computational and statistical challenges inherent in the covariance estimation, Lee et al. [22] designed pivotal statistics by a *random scaling* approach, which yields asymptotically valid confidence intervals. This technique has demonstrated promising performance across a wide range of optimization algorithms, including ROOT-SGD [24], stochastic approximation under Markovian data [23], and the Kiefer–Wolfowitz algorithm [8].

**Notation.** We let $\|\cdot\|$ denote the $\ell_2$ norm for vectors and spectral norm for matrices. We let $\lfloor\cdot\rfloor$ denote the floor function, which rounds down to the nearest integer. For a sequence of compatible matrices $\{A_i\}_i$, $\prod_{k=i}^{j} A_k = A_j A_{j-1} \cdots A_i$ when $j \geq i$ and $I$ when $j < i$. For two matrices $A$ and $B$, $A \succeq B$ means that $A - B$ is positive semidefinite. We let $G(\boldsymbol{x})$ denote the constraints Jacobian, that is, $G(\boldsymbol{x}) = \nabla c(\boldsymbol{x}) = (\nabla c_1(\boldsymbol{x}), \ldots, \nabla c_m(\boldsymbol{x}))^\top \in \mathbb{R}^{m\times d}$. Let $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top c(\boldsymbol{x})$ denote the Lagrangian function of (1), where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the dual vector associated with equality constraints. For simplicity, we let $f_t = f(\boldsymbol{x}_t)$, $c_t = c(\boldsymbol{x}_t)$, $\nabla\mathcal{L}_t = \nabla\mathcal{L}(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ (similar for $\nabla f_t$, $G_t$ etc.); and $f^\star = f(\boldsymbol{x}^\star)$, $c^\star = c(\boldsymbol{x}^\star)$, $\nabla\mathcal{L}^\star = \nabla\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ (similar for $\nabla f^\star$, $G^\star$ etc.). The bar notation $\bar{(\cdot)}$ denotes random quantities at each step, except for the averaged iterate $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{\lambda}}_t) \coloneqq \sum_{i=0}^{t-1}(\boldsymbol{x}_i, \boldsymbol{\lambda}_i)/t$. For example, $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t$ is the random estimate of $\nabla_{\boldsymbol{x}}\mathcal{L}$ at the step $t$.

## 2 Assumptions, Algorithm, and Asymptotic Normality

In this section, we review the AI-SSQP method for solving Problem (1). We refer to [25] for details.

• **Step 1: Estimate objective gradient and Hessian.** Given the current iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$, we draw a sample $\xi_t \sim \mathcal{P}$ and compute $\bar{g}_t = \nabla F(\boldsymbol{x}_t; \xi_t)$ and $\bar{H}_t = \nabla^2 F(\boldsymbol{x}_t; \xi_t)$. Recalling the Jacobian of the constraints $G_t = \nabla c_t$, we then compute $\bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t = \bar{g}_t + G_t^\top \boldsymbol{\lambda}_t$ and $\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_t = \bar{H}_t + \sum_{j=1}^{m}[\boldsymbol{\lambda}_t]_j \nabla^2 c_j(\boldsymbol{x}_t)$. Next, we define the regularized averaged Hessian $B_t$ as $B_t = \frac{1}{t}\sum_{i=0}^{t-1}\bar{\nabla}_{\boldsymbol{x}}^2\mathcal{L}_i + \Delta_t$, where $\Delta_t = \Delta(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ is a regularization term chosen such that $B_t$ is positive definite in $\{\boldsymbol{x} \in \mathbb{R}^d : G_t\boldsymbol{x} = \boldsymbol{0}\}$.

• **Step 2: Inexactly solve the SQP subproblem.** We aim to solve the following linear system:

$$\underbrace{\begin{pmatrix} B_t & G_t^\top \\ G_t & \boldsymbol{0} \end{pmatrix}}_{K_t} \underbrace{\begin{pmatrix} \tilde{\Delta}\boldsymbol{x}_t \\ \tilde{\Delta}\boldsymbol{\lambda}_t \end{pmatrix}}_{\tilde{\boldsymbol{z}}_t} = -\underbrace{\begin{pmatrix} \bar{\nabla}_{\boldsymbol{x}}\mathcal{L}_t \\ c_t \end{pmatrix}}_{\bar{\nabla}\mathcal{L}_t}. \tag{3}$$

We employ a sketching solver to derive an approximate solution. For each $t$, we perform $\tau$ sketching steps. At each sketching step $j$, we independently generate a random sketching matrix/vector $S_{t,j} \in \mathbb{R}^{(d+m)\times q}$ for some $q \geq 1$ from a distribution $S$, and aim to solve $S_{t,j}^\top K_t \boldsymbol{z} = -S_{t,j}^\top \bar{\nabla}\mathcal{L}_t$. We choose the solution that is closest to the current solution approximation $\boldsymbol{z}_{t,j}$, that is ($\boldsymbol{z}_{t,0} = \boldsymbol{0}$),

$$\boldsymbol{z}_{t,j+1} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z} - \boldsymbol{z}_{t,j}\|^2 \qquad \text{s.t.} \quad S_{t,j}^\top K_t \boldsymbol{z} = -S_{t,j}^\top \bar{\nabla}\mathcal{L}_t. \tag{4}$$

The closed form solution to (4) is given by $\boldsymbol{z}_{t,j+1} = \boldsymbol{z}_{t,j} - K_t S_{t,j}(S_{t,j}^\top K_t^2 S_{t,j})^\dagger S_{t,j}^\top (K_t \boldsymbol{z}_{t,j} + \bar{\nabla}\mathcal{L}_t)$ for $0 \leq j \leq \tau - 1$, where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse.

• **Step 3: Adaptive stepsize**. We define $(\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t) \coloneqq \boldsymbol{z}_{t,\tau}$ as our approximate Newton direction. Then we update the iterate $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ with an adaptive random stepsize $\bar{\alpha}_t$: $(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\lambda}_t) + \bar{\alpha}_t \cdot (\bar{\Delta}\boldsymbol{x}_t, \bar{\Delta}\boldsymbol{\lambda}_t)$, where $0 < \beta_t \leq \bar{\alpha}_t \leq \eta_t$ with $\eta_t = \beta_t + \chi_t$ and $\{\beta_t, \eta_t\}$ be two positive sequences.

We now state basic assumptions, which are identical to (in fact, slightly weaker than) those in [25]. The first assumption below is common in constrained optimization literature [28, 5].

**Assumption 1** *We assume that all iterates $\{(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)\}_t$ are contained in a compact, convex set $\mathcal{X} \times \Lambda$, such that $f(\boldsymbol{x})$ and $c(\boldsymbol{x})$ are twice continuously differentiable over $\mathcal{X}$, and the Lagrangian Hessian $\nabla^2\mathcal{L}$ is $\Upsilon_L$-Lipschitz continuous over $\mathcal{X} \times \Lambda$, i.e., $\|\nabla^2\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) - \nabla^2\mathcal{L}(\boldsymbol{x}', \boldsymbol{\lambda}')\| \leq \Upsilon_L \|(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{\lambda} - \boldsymbol{\lambda}')\|$, for $\forall(\boldsymbol{x}, \boldsymbol{\lambda}), (\boldsymbol{x}', \boldsymbol{\lambda}') \in \mathcal{X} \times \Lambda$. In addition, we assume that the constraint Jacobian $G_t$ has full row rank satisfying $G_t G_t^\top \succeq \gamma_G I$ for some $\gamma_G > 0$. The regularization term $\Delta_t$ ensures that $B_t$ satisfies $\|B_t\| \leq \Upsilon_B$ and $\boldsymbol{x}^\top B_t \boldsymbol{x} \geq \gamma_{RH}\|\boldsymbol{x}\|^2$, $\forall \boldsymbol{x} \in$ Kernel$(G_t)$ for some $\Upsilon_B, \gamma_{RH} > 0$.*

**Assumption 2** *We assume the gradient and Hessian estimates are unbiased: $\mathbb{E}[\nabla F(\boldsymbol{x}_t; \xi_t) \mid \boldsymbol{x}_t] = \nabla f_t$ and $\mathbb{E}[\nabla^2 F(\boldsymbol{x}_t; \xi_t) \mid \boldsymbol{x}_t] = \nabla^2 f_t$, $\forall t \geq 0$; and we assume the following moment conditions: $\mathbb{E}[\|\nabla F(\boldsymbol{x}_t; \xi_t) - \nabla f_t\|^{2+\delta} \mid \boldsymbol{x}_t] \leq \Upsilon_m$ and $\mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla^2 F(\boldsymbol{x}; \xi)\|^2] \leq \Upsilon_m$, for some $\delta, \Upsilon_m > 0$.*

We highlight that we only requires a bounded $(2+\delta)$-moment, which is weaker than existing inference study that require *at least* a bounded 4th-order moment [7, 33, 25, 12, 17, 21]. The next assumption characterizes the sketching distribution, which ensures the linear convergence of the sketching solver.

**Assumption 3** *We assume the sketching matrix $S \in \mathbb{R}^{(d+m)\times q}$ satisfies $\mathbb{E}[\|S\|\|S^\dagger\|] \leq \Upsilon_S$ and $\mathbb{E}[K_t S(S^\top K_t^2 S)^\dagger S^\top K_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}_t] \succeq \gamma_S I$, for any $t \geq 0$ and some constants $\gamma_S, \Upsilon_S > 0$.*

With the above assumptions, we now review the almost sure convergence of AI-SSQP. In nonlinear optimization, global convergence refers to converging to a stationary point from any initialization.

**Theorem 1 ([25], Theorem 4.8)** *Consider the AI-SSQP updates in Step 3 under Assumptions 1, 2, 3. There exists a threshold $\tau^\star$ such that for any sketching steps $\tau \geq \tau^\star$ and any stepsize control sequences $\{\beta_t, \eta_t = \beta_t + \chi_t\}$ satisfying $\sum_{t=0}^\infty \beta_t = \infty$, $\sum_{t=0}^\infty \beta_t^2 < \infty$, $\sum_{t=0}^\infty \chi_t < \infty$, we have $\|\nabla\mathcal{L}_t\| \to 0$ and $\|(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)\| \to 0$ as $t \to \infty$ almost surely.*

To segue into the local analysis, we assume from now on that the iterates $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ converge to a *regular* local solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, that is, $G^\star = \nabla c^\star$ has full row rank and $\nabla_{\boldsymbol{x}}^2\mathcal{L}^\star$ is positive definite on Kernel$(G^\star)$. These regularity conditions are necessary in various statistics literature [32, 14, 12, 25]. We denote the averaged iterate as $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{\lambda}}_t) \coloneqq 1/t \cdot \sum_{i=0}^{t-1}(\boldsymbol{x}_i, \boldsymbol{\lambda}_i)$. We define the limiting covariance matrix as $\Xi^\star \coloneqq (\nabla^2\mathcal{L}^\star)^{-1}\text{cov}(\nabla\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star; \xi))(\nabla^2\mathcal{L}^\star)^{-1}$. By [14, 12], $\Xi^\star$ is locally asymptotically minimax optimal in the sense of Hájek and Le Cam. We then define the product of the projection matrices: $\widetilde{C}^\star \coloneqq \prod_{j=1}^\tau (I - K^\star S_j(S_j^\top (K^\star)^2 S_j)^\dagger S_j^\top K^\star)$, and $C^\star \coloneqq \mathbb{E}[\widetilde{C}^\star] = (I - \mathbb{E}[K^\star S(S^\top (K^\star)^2 S)^\dagger S^\top K^\star])^\tau$, where $S_1, \ldots, S_\tau \overset{iid}{\sim} S$.

**Theorem 2 (Asymptotic normality of averaged AI-SSQP)** *Under Assumptions 1, 2, 3 and suppose the sketching step $\tau \geq \tau^\star$ and the stepsize control sequences $\beta_t = c_\beta/(t+1)^\beta$ and $\chi_t = c_\chi/(t+1)^\chi$ satisfy $c_\beta, c_\chi > 0$, $\beta \in (0.5, 1)$, and $\chi > \beta + 0.5$. Then, we have*

$$\sqrt{t} \cdot (\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star, \bar{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}^\star) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \bar{\Xi}^\star), \tag{5}$$

*where $\bar{\Xi}^\star = (I - C^\star)^{-1}\mathbb{E}[(I - \widetilde{C}^\star)\Xi^\star(I - \widetilde{C}^\star)^\top](I - C^\star)^{-1}$.*

The above theorem establishes the asymptotic normality of AI-SSQP method. We also mention that under Assumption 3, $I - C^\star$ is positive definite and hence invertible. [25] has shown that the limiting covariance $\tilde{\Xi}^\star$ in (2) solves the following Lyapunov equation:

$$\left(\left\{1 - \frac{\mathbf{1}_{\{\beta=1\}}}{2c_\beta}\right\} I - C^\star\right) \tilde{\Xi}^\star + \tilde{\Xi}^\star \left(\left\{1 - \frac{\mathbf{1}_{\{\beta=1\}}}{2c_\beta}\right\} I - C^\star\right) = \mathbb{E}[(I - \widetilde{C}^\star)\Xi^\star(I - \widetilde{C}^\star)^\top]. \quad (6)$$

To compare under the same scaling as (5), we let $\beta_t = 1/(t+1)$, corresponding to $c_\beta = 1$ and $\beta = 1$. We now examine in the following proposition the relationship between $\tilde{\Xi}^\star$, $\bar{\Xi}^\star$ and $\Xi^\star$.

**Proposition 1** *Under the conditions of Theorem 2, we have: (a) Without the sketching solver (i.e., solving* (3) *exactly),* $\bar{\Xi}^\star = \Xi^\star$*. With the sketching solver, we have* $\bar{\Xi}^\star \succeq \Xi^\star$*. But the difference can be bounded by* $\|\bar{\Xi}^\star - \Xi^\star\| \leq \frac{\{1+(1-\gamma_S)^\tau\}\cdot(1-\gamma_S)^\tau}{\{1-(1-\gamma_S)^\tau\}^2}\|\Xi^\star\|$*. (b) Consider the Lyapunov equation* (6) *with* $c_\beta = 1$ *and* $\beta = 1$*. If* $(1-\gamma_S)^\tau < 0.5$*, then we have* $\bar{\Xi}^\star \preceq \tilde{\Xi}^\star$*.*

We make two remarks concerning the proposition. First, the statistical efficiency of AI-SSQP is indeed affected by the sketching solver. Without sketching, the method achieves optimal statistical efficiency. With sketching, the randomness caused by the sketching solver degrades statistical efficiency, leading to $\bar{\Xi}^\star \succeq \Xi^\star$. Fortunately, the degradation decays exponentially with the number of sketching steps. Second, the averaged iterate of AI-SSQP exhibits better statistical efficiency than the last iterate.

## 3 Online Statistical Inference

In this section, we present our online inference procedures for $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$. We begin to extend the asymptotic normality in Theorem 2 to the Functional Central Limit Theorem (FCLT).

**Theorem 3** *Under conditions of Theorem 2, we have* $1/\sqrt{t} \cdot \sum_{i=0}^{\lfloor rt \rfloor - 1}(\boldsymbol{x}_i - \boldsymbol{x}^\star, \boldsymbol{\lambda}_i - \boldsymbol{\lambda}^\star) \Longrightarrow (\bar{\Xi}^\star)^{1/2} W_{d+m}(r)$*, where* $r \in [0,1]$*,* $W_{d+m}(\cdot)$ *is the standard* $(d+m)$*-dimensional Brownian motion.*

We use "$\Longrightarrow$" in Theorem 3 to denote convergence in distribution in a function space. Next, we provide a studentized test statistic based on Theorem 3. We define the *random scaling* matrix as $V_t := \frac{1}{t^2}\sum_{i=1}^t i^2(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{\lambda}}_i - \bar{\boldsymbol{\lambda}}_t)(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_t, \bar{\boldsymbol{\lambda}}_i - \bar{\boldsymbol{\lambda}}_t)^\top$. The following theorem establishes that the test statistic is asymptotically pivotal, meaning its limiting distribution is free of unknown parameters.

**Theorem 4** *Under the conditions of Theorem 2 and assuming cov*$(\nabla F(\boldsymbol{x}^\star; \xi)) \succ 0$*, for any vector* $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{\lambda}}) \in \mathbb{R}^{d+m}$ *with* $\boldsymbol{w} \notin$ *Span*$((G^\star)^\top) \otimes \mathbf{0}_m$*, we have*

$$\frac{\sqrt{t}\, \boldsymbol{w}^\top(\bar{\boldsymbol{x}}_t - \boldsymbol{x}^\star, \bar{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}^\star)}{\sqrt{\boldsymbol{w}^\top V_t \boldsymbol{w}}} \xrightarrow{d} \frac{W_1(1)}{\sqrt{\int_0^1 (W_1(r) - rW_1(1))^2\, dr}}. \quad (7)$$

The condition $\boldsymbol{w} \notin \text{Span}((G^\star)^\top) \otimes \mathbf{0}_m$ ensures that the inference direction possesses a non-trivial component in the tangential direction of the constraint function, i.e. Kernel$(G^\star)$. Furthermore, the asymptotic distribution on the RHS of (7) has been extensively studied in the econometrics and statistics literature [18, 1, 19, 2]. Table 1 presents its quantiles adapted from [1]. Using these quantiles and Theorem 4, one can immediately construct asymptotically pivotal confidence intervals.

| $p$ | 90% | 95% | 97.5% | 99% |
|---|---|---|---|---|
| Quantile($p$) | 3.875 | 5.323 | 6.747 | 8.613 |

Table 1: Quantile table of the distribution $W_1(1)/\{\int_0^1 (W_1(r) - rW_1(1))^2\, dr\}^{1/2}$.

We conclude this section by highlighting the online fashion of our inference procedure. First, $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{\lambda}}_t)$ are computed recursively via $(\bar{\boldsymbol{x}}_{t+1}, \bar{\boldsymbol{\lambda}}_{t+1}) = (\boldsymbol{x}_t, \boldsymbol{\lambda}_t)/(t+1) + t \cdot (\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{\lambda}}_t)/(t+1)$. Second, $V_t$ can be updated through:

$$V_t = \frac{1}{t^2}\sum_{i=1}^t i^2 \begin{pmatrix}\bar{\boldsymbol{x}}_i \\ \bar{\boldsymbol{\lambda}}_i\end{pmatrix}\begin{pmatrix}\bar{\boldsymbol{x}}_i \\ \bar{\boldsymbol{\lambda}}_i\end{pmatrix}^\top - \frac{1}{t^2}\begin{pmatrix}\sum_{i=1}^t i^2 \bar{\boldsymbol{x}}_i \\ \sum_{i=1}^t i^2 \bar{\boldsymbol{\lambda}}_i\end{pmatrix}\begin{pmatrix}\bar{\boldsymbol{x}}_t \\ \bar{\boldsymbol{\lambda}}_t\end{pmatrix}^\top - \frac{1}{t^2}\begin{pmatrix}\bar{\boldsymbol{x}}_t \\ \bar{\boldsymbol{\lambda}}_t\end{pmatrix}\begin{pmatrix}\sum_{i=1}^t i^2 \bar{\boldsymbol{x}}_i \\ \sum_{i=1}^t i^2 \bar{\boldsymbol{\lambda}}_i\end{pmatrix}^\top + \frac{\sum_{i=1}^t i^2}{t^2}\begin{pmatrix}\bar{\boldsymbol{x}}_t \\ \bar{\boldsymbol{\lambda}}_t\end{pmatrix}\begin{pmatrix}\bar{\boldsymbol{x}}_t \\ \bar{\boldsymbol{\lambda}}_t\end{pmatrix}^\top,$$

where each of the four components can be easily computed recursively. Therefore, the computation of $V_t$ admits a fully online implementation.

## References

[1] K. M. Abadir and P. Paruolo. Two mixed normal densities from cointegration analysis. *Econometrica*, 65(3):671, May 1997. ISSN 0012-9682. doi: 10.2307/2171758.

[2] K. M. Abadir and P. Paruolo. Simple robust testing of regression hypotheses: A comment. *Econometrica*, 70(5):2097–2099, Sept. 2002. ISSN 1468-0262. doi: 10.1111/1468-0262.00367.

[3] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2): 1352–1379, Jan. 2021. ISSN 1095-7189. doi: 10.1137/20m1354556.

[4] A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient jacobians. *Mathematics of Operations Research*, 49(4):2212–2248, Nov. 2024. ISSN 1526-5471. doi: 10.1287/moor.2021.0154.

[5] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[6] C. Chen, F. Tung, N. Vedula, and G. Mori. Constraint-aware deep neural network compression. In *Computer Vision – ECCV 2018*, pages 409–424. Springer International Publishing, September 2018. doi: 10.1007/978-3-030-01237-3_25. URL https://doi.org/10.1007/978-3-030-01237-3_25.

[7] X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1), Feb. 2020. ISSN 0090-5364. doi: 10.1214/18-aos1801.

[8] X. Chen, Z. Lai, H. Li, and Y. Zhang. Online statistical inference for stochastic optimization via kiefer-wolfowitz methods. *Journal of the American Statistical Association*, 119(548): 2972–2982, Jan. 2024. ISSN 1537-274X. doi: 10.1080/01621459.2023.2296703.

[9] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, July 2022. ISSN 1573-7691. doi: 10.1007/s10915-022-01939-z.

[10] F. E. Curtis, D. P. Robinson, and B. Zhou. A stochastic inexact sequential quadratic optimization algorithm for nonlinear equality-constrained optimization. *INFORMS Journal on Optimization*, 6(3–4):173–195, July 2024. ISSN 2575-1492. doi: 10.1287/ijoo.2022.0008.

[11] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, Jan. 2019. ISSN 1615-3383. doi: 10.1007/s10208-018-09409-5.

[12] D. Davis, D. Drusvyatskiy, and L. Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *The Annals of Statistics*, 52(4), Aug. 2024. ISSN 0090-5364. doi: 10.1214/24-aos2401.

[13] J.-H. Du, Y. Guo, and X. Wang. High-dimensional portfolio selection with cardinality constraints. *Journal of the American Statistical Association*, 118(542):779–791, Nov. 2022. ISSN 1537-274X. doi: 10.1080/01621459.2022.2133718.

[14] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1), Feb. 2021. ISSN 0090-5364. doi: 10.1214/19-aos1831.

[15] J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, jun 2012. doi: 10.1080/01621459.2012.682825.

[16] Y. Fang, S. Na, M. W. Mahoney, and M. Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization*, 34(2):2007–2037, June 2024. ISSN 1095-7189. doi: 10.1137/22m1537862.

[17] L. Jiang, A. Roy, K. Balasubramanian, D. Davis, D. Drusvyatskiy, and S. Na. Online covariance estimation in nonsmooth stochastic approximation. *arXiv preprint arXiv:2502.05305*, 2025.

[18] S. Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):1551, Nov. 1991. ISSN 0012-9682. doi: 10.2307/2938278.

[19] N. M. Kiefer, T. J. Vogelsang, and H. Bunzel. Simple robust testing of regression hypotheses. *Econometrica*, 68(3):695–714, May 2000. ISSN 1468-0262. doi: 10.1111/1468-0262.00128.

[20] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. Inexact objective function evaluations in a trust-region algorithm for pde-constrained optimization under uncertainty. *SIAM Journal on Scientific Computing*, 36(6):A3011–A3029, Jan. 2014. ISSN 1095-7197. doi: 10.1137/140955665.

[21] W. Kuang, M. Anitescu, and S. Na. Online covariance matrix estimation in sketched newton methods. *arXiv preprint arXiv:2502.07114*, 2025.

[22] S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022.

[23] X. Li, J. Liang, and Z. Zhang. Online statistical inference for nonlinear stochastic approximation with markovian data. *arXiv preprint arXiv:2302.07690*, 2023.

[24] Y. Luo, X. Huo, and Y. Mei. Covariance estimators for the root-sgd algorithm in online learning. *arXiv preprint arXiv:2212.01259*, 2022.

[25] S. Na and M. Mahoney. Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *Journal of Machine Learning Research*, 26(33):1–75, 2025.

[26] S. Na, M. Anitescu, and M. Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, 199(1–2):721–791, June 2022. ISSN 1436-4646. doi: 10.1007/s10107-022-01846-z.

[27] S. Na, M. Anitescu, and M. Kolar. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. *Mathematical Programming*, mar 2023. doi: 10.1007/s10107-023-01935-7.

[28] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer New York, 2nd edition, 2006. ISBN 9780387303031. doi: 10.1007/978-0-387-40065-5.

[29] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992. ISSN 1095-7138. doi: 10.1137/0330046.

[30] T. Rees, H. S. Dollar, and A. J. Wathen. Optimal solvers for pde-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, Jan. 2010. ISSN 1095-7197. doi: 10.1137/080727154.

[31] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, Sept. 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586.

[32] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Jan. 2014. ISBN 9781611973433. doi: 10.1137/1.9781611973433.

[33] W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, July 2021. ISSN 1537-274X. doi: 10.1080/01621459.2021.1933498.