Online Learning with Semiparametric Stochastic Approximation

Mingli Chen
University of Warwick

Kengo Kato
Cornell University

Ye Luo
University of Hong Kong

Preliminary, May 2025

Motivation

- **Online Learning**
  - Rapid growth driven by real-world applications and the digital economy.
  - Processes data incrementally without storing the full dataset—ideal for large-scale problems.
  - Real-time decision, enables immediate responses to new information as it arrives.
- **Challenges in Estimation and Inference with Streaming Data**
  - Existing work is largely restricted to parametric models.
  - Many problems involve:
    - Low-dimensional parameters of interest.
    - High-dimensional/nonparametric nuisance component.
  - Semiparametric approaches remain underexplored.
  - Inference: all intermediate estimates must be retained for variance calculation..

(Stochastic) Gradient Descent

A stochastic approximation method (Robbins and Monro [1951]), such as Stochastic Gradient Descent (SGD), is a salable algorithm for parameter estimation.

Let each observation at time t be $U_t = (y_t, x_t)$, where y_t is the response, $x_t \in \mathbb{R}^{d_1}$ is a low-dimensional covariate vector.

- Traditional SGD goal: find $\theta^* = \arg\min \mathbb{E}[f(\theta; U)]$
- $f(\cdot)$: (unknown) function, may corresponds to a squared loss function
- Iterative updating rule:

$$\theta_t = \theta_{t-1} - \eta_t \underbrace{\nabla f(\theta_{t-1}; y_t, x_t)}_{:=G(\theta_{t-1}; U_t)}, \tag{1}$$

recursively updates the estimate upon the arrival of each data point x_t, for $t = 1, \ldots, T$.

SGD (Cont.)

Especially relevant for online learning

- η_t: learning rate at time t
- ∇f: gradient of $f(\cdot)$

Iterative algorithm converges to θ^* with high probability.
Adaptive learning in macroeconomics, $\eta_t = 1/t$.

This Paper

Propose Semi-SGD as a one-pass algorithm: low space and time complexity, requiring only the current data and the previous estimate.

Case 1. Consider the following model:

$$F^{-1}_{y_t|x_t,v_t;\theta_\tau}(\tau) = x_t^\intercal \theta_{\tau,1} + \underbrace{P^{k_t}(v_t)^\intercal \theta_{\tau,2} + r_{\tau,t}}_{:=\lambda_\tau(v_t)} \tag{2}$$

This Paper

Propose Semi-SGD as a one-pass algorithm: low space and time complexity, requiring only the current data and the previous estimate.

Case 1. Consider the following model:

$$F^{-1}_{y_t|x_t,v_t;\theta_\tau}(\tau) = x_t^\intercal \theta_{\tau,1} + \underbrace{P^{k_t}(v_t)^\intercal \theta_{\tau,2} + r_{\tau,t}}_{:=\lambda_\tau(v_t)} \tag{2}$$

- Approximation $\lambda_\tau(v_t)$ using a sieve basis expansion $P^{k_t}(v_t)$.
- Denote $\theta_\tau = (\theta_{\tau,1}, \theta_{\tau,2})$, and $P^{k_t}(w_t) = \big(x_t, P^{k_t}(v_t)\big)$ is the Sieve to use at time t.
- k_t can be pre-specified as a function of T, the terminal value of t. Abbreviate k_t as κ.

As a result, we can write down a semi-parametric stochastic approximation process as:

$$\theta_t = \theta_{t-1} - \eta_t P^\kappa(w_t)(\tau - 1(P^\kappa(w_t)^\intercal \theta_{t-1} - y_t < 0)), \tag{3}$$

where we focus on $\eta_t = \eta_0 t^{-\alpha}$ as the learning rate, with $\eta_0 > 0$, and $\alpha \in (1/2, 1]$.

This Paper

Propose Semi-SGD as a one-pass algorithm: low space and time complexity, requiring only the current data and the previous estimate.

Case 2. Consider the following model:

$$y_t = x_t^\intercal \theta_1 + \underbrace{P^{k_t}(v_t)^\intercal \theta_2 + r_t}_{:=\lambda(v_t)} + \varepsilon_t \tag{4}$$

- Approximation $\lambda(v_t)$ using a sieve basis expansion $P^{k_t}(v_t)$.
- Denote $\theta = (\theta_1, \theta_2)$, and $P^{k_t}(w_t) = \big(x_t, P^{k_t}(v_t)\big)$ is the Sieve to use at time t.
- k_t can be pre-specified as a function of T, the terminal value of t. Abbreviate k_t as κ.

This Paper

Propose Semi-SGD as a one-pass algorithm: low space and time complexity, requiring only the current data and the previous estimate.

Case 2. Consider the following model:

$$y_t = x_t^\intercal \theta_1 + \underbrace{P^{k_t}(v_t)^\intercal \theta_2 + r_t}_{:=\lambda(v_t)} + \varepsilon_t \tag{4}$$

- Approximation $\lambda(v_t)$ using a sieve basis expansion $P^{k_t}(v_t)$.
- Denote $\theta = (\theta_1, \theta_2)$, and $P^{k_t}(w_t) = \big(x_t, P^{k_t}(v_t)\big)$ is the Sieve to use at time t.
- k_t can be pre-specified as a function of T, the terminal value of t. Abbreviate k_t as κ.

As a result, we can write down a semi-parametric stochastic approximation process as:

$$\theta_t = \theta_{t-1} - \eta_t P^\kappa(w_t)(P^\kappa(w_t)^\intercal \theta_{t-1} - y_t), \tag{5}$$

where we focus on $\eta_t = \eta_0 t^{-\alpha}$ as the learning rate, with $\eta_0 > 0$, and $\alpha \in (1/2, 1]$.

This Paper (Cont.)

Case 1. Given τ, κ, define $U_t = (y_t, w_t)$, and

$$G(\theta_{t-1}; U_t) = P^\kappa(w_t)^\intercal(\tau - 1(P_\kappa(w_t)^\intercal\theta_{t-1} - y_t < 0)). \tag{6}$$

Also, define $\theta^* := (\theta^*_{\tau,1}, \theta^*_{\tau,\kappa,2})$ where

$$\theta^*_{\tau,\kappa,2} := \arg\min_{\theta_2} \|\lambda_\tau(v_t) - P^\kappa(v_t)^\intercal\theta_2\|_d \tag{7}$$

for some $\theta_2 \in \mathbb{R}^\kappa$.

This Paper (Cont.)

Case 2. Given κ, define $U_t = (y_t, w_t)$, and

$$G(\theta_{t-1}; U_t) = P^{\kappa}(w_t)^{\intercal}(P_{\kappa}(w_t)^{\intercal}\theta_{t-1} - y_t). \tag{8}$$

Also, define $\theta^* := (\theta_1^*, \theta_{\kappa,2}^*)$ where

$$\theta_{\kappa,2}^* := \arg\min_{\theta_2} \|\lambda_{\tau}(v_t) - P^{\kappa}(v_t)^{\intercal}\theta_2\|_d \tag{9}$$

for some $\theta_2 \in \mathbb{R}^{\kappa}$.

Illustrative Example of Motivation

Applicable to general control function approach in various models, e.g. Lee [2007]

$$y = x\beta_\tau + z_1^\intercal \gamma_\tau + u, \tag{10}$$

$$x = \mu(\alpha) + z^\intercal \pi(\alpha) + v, \tag{11}$$

and

$$Q_{u|x,z}(\tau) = \lambda_\tau(v). \tag{12}$$

The parameter of interest is the quantile parameter $\beta_\tau \in \mathbb{R}^p$ and $\gamma_\tau \in \mathbb{R}^{d_{z1}}$ for a specific value of quantile τ, with vector of exogenous explanatory variables $z \in \mathbb{R}^q$.

With a conditional independence condition that $u|v, z = u|v$, it can be shown that

$$F^{-1}_{y|x,z;\theta}(\tau) = x\beta_\tau + z_1'\gamma_\tau + \lambda_\tau(v) \tag{13}$$

where $\lambda_\tau(v)$ is a unknown non-parametric function of v. Here, $\theta_{\tau,1} = (\beta_\tau, \gamma_\tau)$

Relation to Literature (very selective)

- Traditional stochastic gradient descent algorithm
  - Robbins and Monro [1951], Kiefer and Wolfowitz [1952], Ruppert [1988], Polyak and Juditsky [1992].
- Online learning
  - Bottou et al. [1998], Mairal, Bach, Ponce, and Sapiro [2010], Hoffman, Bach, and Blei [2010].
- Recent work focusing on inference
  - Chen, Liu, and Zhang [2021], Li, Liu, Kyrillidis, and Caramanis [2018], Forneron [2022], Lee, Liao, Seo, and Shin [2022], Fang, Xu, and Yang [2018].

Algorithm 1 Semi-SGD for SQR as in (5)

Input : Function

Initialization: Set θ, κ, B and T

for $t = 1, \ldots, T$ *and* $b = 1, .., B$ **do**

for any positive integer κ, construct $P^{\kappa}(w_t) = [x_t, p_{1\kappa}(v_t), ..., p_{\kappa\kappa}(v_t)]$ and update θ (and θ^b) via

$$\theta_t = \theta_{t-1} - \eta_t \cdot P^{\kappa}(\omega_t)(\tau - \mathbf{1}(P^{\kappa}(\omega_t)^{\intercal}\theta_{t-1} - y_t < 0)),$$
$$\theta_t^b = \theta_{t-1}^b - \eta_t \cdot W_{t,b} \cdot P^{\kappa}(\omega_t)(\tau - \mathbf{1}(P^{\kappa}(\omega_t)^{\intercal}\theta_{t-1} - y_t < 0)),$$

where η_t and $W_{t,b}$ are the step sizes (learning rates) and bootstrap weights of the t-th update respectively.

end

Output : obtain (1-α)-confidence interval estimator of $\bar{\theta}_t$: $\bar{\theta}_t \pm z_{\alpha/2}\tilde{\sigma}_B$, where $\tilde{\sigma}_B$ obtained from the bootstrap procedure.

Algorithm 2 Semi-SGD Control Function Approach for endogenous QR as in (13), given τ

Input : Function

Initialization : Set θ, T_1, κ, B, and T

Step 1 (the offline CF-QR) : **for** $t = 1, \ldots, \mathtt{T}_1$ **do**

Observe $(y_{1:T_1}, x_{1:T_1}, z_{1,1:T_1}, z_{2,1:T_1})$ Step 1a: Run QR of $x_{1:T_1}$ on $(1, z_{1,1:T_1}, z_{2,1:T_1})$ get

$$\pi_{T_1} \quad \text{and} \quad v_{1:T_1} \quad \text{from} \quad \text{eq (11)}$$

Step 1b: Given $\hat{v}_{1:T_1}$ as estimates of $v_{1:T_1}$, consider a series regression with $w_{1:T_1} = (x_{1:T_1}, z_{1,1:T_1}, P^{\kappa}(\hat{v}_{1:T_1}))$ as covariates, where $P^{\kappa}(\hat{v}_{1:T_1})$ is a Sieve of $\hat{v}_{1:T_1}$; Run QR again of $y_{1:T_1}$ on $w_{1:T_1}$, and obtain estimates of

$$(\hat{\beta}_{\tau,T_1},\ \hat{\gamma}_{\tau,T_1})$$

end

Step 2 (the online semi-SGD part): for $t = \mathtt{T}_1 + 1, \ldots, T$ **do**

Step 2a: Given quantile index α, update π and v

$$\pi_t = \pi_{t-1} - \eta_{1t} \cdot z_t^{\intercal}(\alpha - \mathbf{1}(z_t^{\intercal}\pi_{t-1} - x_t < 0));$$
$$v_t = x_t - z_t^{\intercal}\pi_t$$

Step 2b: for any positive integer κ, construct $P^{\kappa}(w_t) = [x_t, z_{1t}, p_1(v_t), ..., p_{\kappa}(v_t)]$; and update θ

$$\theta_t = \theta_{t-1} - \eta_{2t} \cdot P_{\kappa}(w_t)(\tau - \mathbf{1}(P_{\kappa}(w_t)^{\intercal}\theta_{t-1} - y_t < 0));$$

end

η_{1t} and η_{2t} are the step sizes (learning rates) of the t-th update for Step 1 and Step 2 respectively.

Comments

- Allow for κ increase at each step
- For the Initial Step in Algorithm 2, we are using T_1 observations to get good initial estimates for $\lambda_\tau(v)$
- For asymptotic results, discuss two cases, $\alpha \in (1/2, 1)$, and $\alpha = 1$
- stochastic (sub)gradient descent

Assumptions

1 (a) Data $U_t = \{(y_t, x_t, v_t), t = 1, ..., T\}$ are independently distributed. x_t, v_t has bounded and compact support $\mathcal{X} \times \mathcal{V}$; (b) $\lambda(v)$ is r-times continuously differentiable on $\mathcal{V}$.

2 Denote

$$A_\kappa := -\nabla \bar{G}(\theta_\kappa) = \mathbb{E}\big[P^\kappa(w_t) P^\kappa(w_t)^\intercal\big], \tag{14}$$

the Jacobian (Hadamard derivative) of the population gradient at θ_κ. Assume all eigenvalues of A_κ being positively bounded away from 0. Denote the lower bound is ψ.

3 For power series $\kappa = C_1 t^{v_1}$ for some constants C_1 satisfying $0 < C_1 < \infty$ and some v_1 satisfying $1/(2r) < v_1 < 1/8$, and for splines $\kappa = C_2 t^{v_2}$ for some constants C_2 satisfying $0 < C_2 < \infty$ and some v_2 satisfying $1/(2r) < v_2 < 1/5$.

4 (i) There exists a sequence $\zeta_0(\kappa)$ such that $\sup_{v \in \mathcal{V}} \|P^\kappa(v)\| \leq \zeta_0(\kappa)$, with $\zeta_0(\kappa)^2 \kappa / t^{\alpha/2} \to 0$. (ii) $\|\lambda^*(\cdot) - P^\kappa(\cdot)^\intercal \theta^*_{\kappa,2}\| \leq C\kappa^{-\zeta}$ for some fixed constant $C > 0$ and $\zeta > 0$.

- Consider $\bar{G}(\theta) := \mathbb{E}[G(\theta; U_t)]$. It can be shown that

$$\bar{G}(\theta^*) = \mathbb{E}\big[P^\kappa(w_t)^\intercal \big(y_t - P^\kappa(w_t)^\intercal \theta^*\big)\big] \tag{15}$$

$$= \mathbb{E}\big[P^\kappa(w_t)^\intercal \big(x_t\theta_1^* + \lambda(v_t) - P^\kappa(w_t)^\intercal \theta_{\kappa,2}^*\big)\big] \tag{16}$$

$$= \mathbb{E}\big[P^\kappa(w_t)^\intercal \big(\lambda(v_t) - P^\kappa(v_t)^\intercal \theta_{\kappa,2}^*\big)\big] + O\big(\|\lambda(v_t) - P^\kappa(v_t)^\intercal \theta_{\kappa,2}^*\|^2\big), \tag{17}$$

under some regularity conditions, we can conclude that for each component of $\bar{G}(\cdot)$, we have that: $|\bar{G}_j(\theta^*)| \leq C't^{-\nu'}$, $j = 1, 2, ..., d_w$ for some fixed constant $C' > 0$.

Sketch of Proofs

Decomposition for fixed κ

$$\theta_t - \theta^* = \theta_{t-1} - \theta^* - \eta_t \bar{G}(\theta_{t-1}) - \eta_t (G(\theta_{t-1}; U_t) - \bar{G}(\theta_{t-1})) \tag{18}$$

$$= (I - \eta_t A_\kappa)(\theta_{t-1} - \theta^*) - \eta_t r(\theta_{t-1} - \theta^*) - \eta_t (G(\theta_{t-1}; U_t) - \bar{G}(\theta_{t-1})), \tag{19}$$

with $r(\theta - \theta^*) = \bar{G}(\theta) - \bar{G}(\theta^*) - A_\kappa(\theta - \theta^*)$ high order residual.
Define $Q_{s,t} = \Pi_{l=s}^{t-1}(I - \eta_l A_\kappa)$ which is a matrix discount factor. The updating condition can be written as:

$$\theta_t - \theta^* = \underbrace{Q_{0,t}(\theta_0 - \theta^*)}_{\Psi_1} - \underbrace{\sum_{s=1}^{t-1} \eta_s Q_{s,t} r(\theta_s - \theta^*)}_{\Psi_2} - \underbrace{\sum_{s=1}^{t-1} \eta_s Q_{s,t}(G(\theta_s, U_s) - \bar{G}(\theta_s))}_{\Psi_3}, \tag{20}$$

By construction, we have that $Q_{s,t} \leq \exp(-\frac{\eta_0 \psi}{1-\alpha}(t^{1-\alpha} - s^{1-\alpha}))$ for $\alpha \in (0,1)$, and $Q_{s,t} \leq \left(\frac{s}{t}\right)^{\eta_0 \psi}$ when $\alpha = 1$.

Lemma 1 [Risk Bound]

Define a metric, $\|\theta_t - \theta_\kappa\|_d^2 := \|\theta_{t,1} - \theta_1\|^2 + \mathbb{E}_{w_s}[|P(w_s)^\intercal(\theta_{t,2} - \theta_{\kappa,2})|^2]$.

When the learning rate is $\eta_t = \eta_0 t^{-\alpha}$, for t large enough, we have:

$$\mathbb{E}[\|\theta_t - \theta_\kappa\|_d^2] \leq \begin{cases} C_1 t^{-\alpha} \ln t, & \text{if } \alpha \in \left(\frac{1}{2}, 1\right), \\ C_2 t^{-1}, & \text{if } \alpha = 1 \text{ and } 2\psi\eta_0 > 1, \end{cases}$$

for some fixed positive constants $\eta_0, C_1, C_2 > 0$.

Asymptotics

Define $L_\kappa = \frac{1}{t}\sum_{s=1}^{t} P^\kappa(w_s)P^\kappa(w_s)^\intercal$. $A_\kappa^* := -\nabla \bar{G}(\theta_\kappa^*)$
When $\alpha < 1$,

$$\Sigma_\kappa := \eta_0 \int_0^\infty \exp(-uA_\kappa^*) L_\kappa \exp(-uA_\kappa^*)^\intercal du, \tag{21}$$

and when $\alpha = 1$,

$$\Sigma_\kappa := \eta_0 \int_0^\infty \exp(-u/\eta_0) \exp(-uA_\kappa^*) L_\kappa \exp(-uA_\kappa^*)^\intercal du. \tag{22}$$

The additional term $\exp(u/\eta_0)$ accounts for the linearly decaying step size.
Consider any C^1 functional $g(\theta_1^*, \lambda(\cdot))$ with bounded derivative that is approximated by $g(\theta_{1,t}, P^\kappa(\cdot)^\intercal \theta_{2,t})$. Denote $\Omega_\kappa := \begin{pmatrix} \frac{\partial b}{\partial \theta_{1,t}} \\ \frac{\partial b}{\partial \lambda} P^\kappa(\cdot) \end{pmatrix}$. That said, g is Hadamard differentiable with respect to λ, e.g.,

$$g(\theta_{1,t}, P^\kappa(v)^\intercal \theta_{2,t}) = a^\intercal \theta_{1,t} + \int_v P^\kappa(v)^\intercal \theta_{2,t} \mu(v) \tag{23}$$

for some probability measure $\mu(v)$.

Theorem 1. If all the assumptions above hold and: $\kappa^{-\zeta} t^{\alpha/2} \to 0$, $\zeta_0^2(\kappa)\kappa / t^{\alpha/2} \to 0$.

$$\sqrt{\eta_0 t^\alpha} \left(\Omega_\kappa^\intercal \Sigma_\kappa \Omega_\kappa\right)^{-\frac{1}{2}} \left(g(\theta_{1,t}, P_\kappa(\cdot)^\intercal \theta_{2,t}) - g(\theta_1^*, \lambda(\cdot))\right) \rightsquigarrow N(0,1). \tag{24}$$

Design

- DGP1: $Y = \mu + X_1\beta + X_2\gamma_1 + X_3 + X_4^2 + X_5^3 + X_6^4 + X_7^5 + U * (X_1\beta + X_2\gamma_1)$
- DGP2 (cf. Lee [2007]):

$$Y_i = X_i\beta + Z_{1i}\gamma + U_i, \quad U_i = V_i + \phi(V_i) + 0.5[\tilde{U}_i - F_{\tilde{U}}^{-1}(\tau)],$$
$$X_i = \mu + Z_{1i}\pi_1 + Z_{2i}\pi_2 + V_i, \quad V_i = \exp(Z_{2i}/2)\tilde{V}_i, \quad i = 1, ..., n$$

where $Z_{1i}, Z_{2i}, \tilde{V}_i$ and $\tilde{U}_i$ are independently drawn from the standard normal distribution, $\phi(v) = 4\exp[-(v-1)^2]$, and $F_{\tilde{U}}$ is the CDF of $\tilde{U}$. The function $\phi(v)$ has a bell-shaped hump around one and represents a nonlinear component of $\lambda_\tau(v) = v + \phi(v)$. We set the parameter values $(\beta, \gamma, \mu, \pi_1, \pi_2) = (1, 1, 1, 3, 1)$. In all experiments $\tau = 0.9$ and $\alpha = 0.5$.

Simulation Results

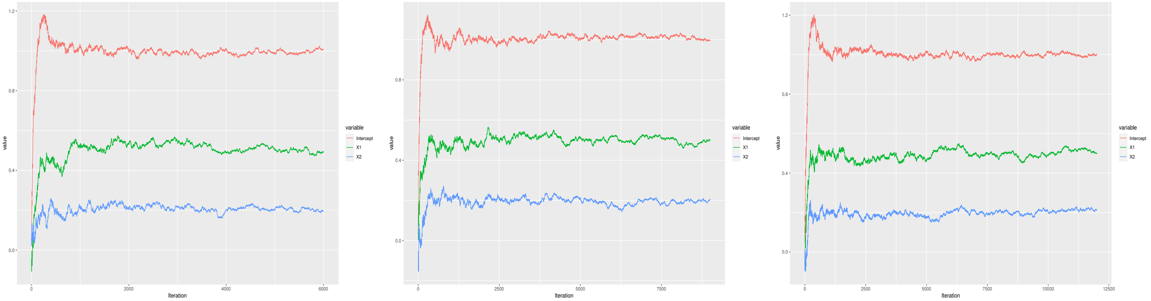


Figure 1: The simulation paths for SQR1 for $n = \{6000, 9000, 12,000\}$, $k = 3$, and coefficients $\{1, 0.5, 0.2\}$

Table 1: Coverage Probabilities of 95% Confidence Intervals for Semiparametric QR

(N, k, τ)			
	$(12000, 3, 0.5)$		
Bias	-0.003	0.0004	-0.0005
SE	0.004	0.008	0.005
CP	0.98	0.95	0.97
	$(12000, 4, 0.5)$		
Bias	-0.0005	0.0001	0
SE	0.004	0.008	0.005
CP	0.99	0.96	0.97
	$(12000, 5, 0.5)$		
Bias	-0.015	0.003	0.001
SE	0.01	0.01	0.01
CP	0.94	0.96	0.97

Based on 500 simulations, with $(\mu, \beta_1, \gamma_1) = (1, 0.5, 0.2)$

Figure 2: The simulation paths for SQR1, $n = 12,000$, $k \in \{3, 4, 5\}$, and coefficients $\{1, 0.5, 0.2\}$

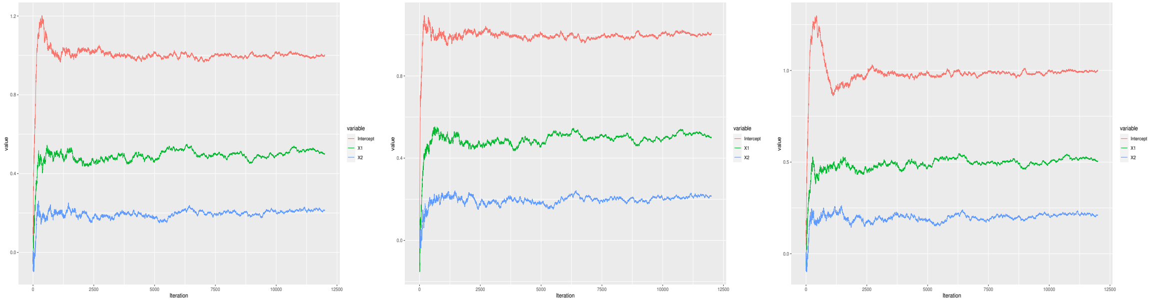
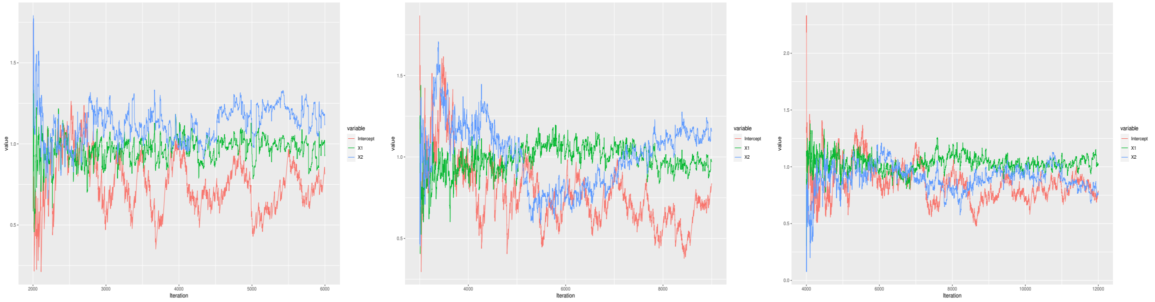


Figure 3: The simulation paths for SQR2, $n = \{6000, 9000, 12,000\}$, $k = 7$, and coefficients $\{1, 1, 1\}$



L. Bottou et al. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

X. Chen, W. Liu, and Y. Zhang. First-order newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–17, 2021.

Y. Fang, J. Xu, and L. Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 2018.

J.-J. Forneron. Estimation and inference by stochastic optimization. *arXiv preprint arXiv:2205.03254*, 2022.

M. Hoffman, F. Bach, and D. Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

S. Lee. Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics*, 141(2):1131–1158, 2007.

S. Lee, Y. Liao, M. H. Seo, and Y. Shin. Fast inference for quantile regression with tens of millions of observations. *Available at SSRN 4263158*, 2022.

T. Li, L. Liu, A. Kyrillidis, and C. Caramanis. Statistical inference using sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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