“Plus/minus” confidence intervals and thresholding

P. Zietkiewicz

April 25, 2024

Table of contents

Summary of Chee et al. (2023)

- $(Y, X) \in \mathbb{R}^d \times \mathbb{R}^p$ and $D_N = \{(Y_i, X_i) : i = 1, \dots, N\}$

$$\theta_* = \text{argmin}_{\theta \in \Theta} \mathbb{E}[\ell(\theta, Y, X)]$$

$$\hat{\theta}_N = \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{N} \ell(\theta, Y_i, X_i)$$

$$F_* = \mathbb{E}[\nabla \ell(\theta, Y, X) \nabla \ell(\theta, Y, X)^\top]$$

- SGD: $\theta_n = \theta_{n-1} - \gamma_n \nabla \ell(\theta_{n-1}; Y_i, X_i)$ for $i = 1, \dots, N$ and γ_n is the learning rate typically $\gamma_n = \gamma_1 / n$. Let θ_N be the one-pass estimator of θ_*.
- Advantages of one-pass over multi-pass: (1) Asymptotic covariance matrix is known in closed form (2) Covariance matrix can be bounded by a factor that depends only on the learning rate γ_1.

Summary of Chee et al. (2023)

- Propose the SGD-based CIs for each component $\theta_{*,j}$

$$\theta_{N,j} \pm 2\sqrt{\frac{\gamma_1^*}{N}} \text{ for } j = 1, \ldots, p.$$

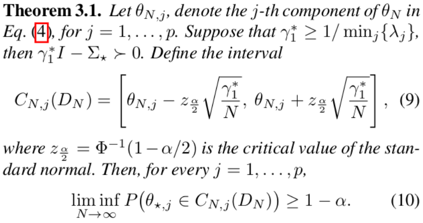
- Define $\Sigma_* = \gamma_1^2(2\gamma_1 F_* - I)^{-1} F_*$ where γ_1 is large enough such that $2\gamma_1 F_* - I \succ 0$. And has eigenvalues

$$\text{eigen}(\Sigma_*) = \{\frac{2\gamma_1^2\lambda_j}{2\gamma_1\lambda_j - 1} : j = 1, \ldots, p\}$$

where λ_j is the jth eigenvalue of F_*.

Summary of Chee et al. (2023)

Results:

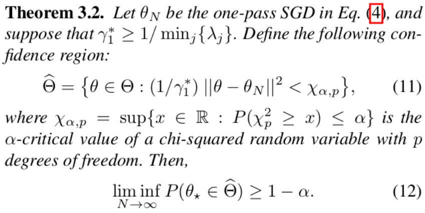


Theorem 3.1. *Let $\theta_{N,j}$, denote the j-th component of θ_N in Eq. (4), for $j = 1, \dots, p$. Suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$, then $\gamma_1^* I - \Sigma_\star \succ 0$. Define the interval*

$$C_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}},\ \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}\right], \tag{9}$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha/2)$ is the critical value of the standard normal. Then, for every $j = 1, \dots, p$,

$$\liminf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) \geq 1 - \alpha. \tag{10}$$



Theorem 3.2. *Let θ_N be the one-pass SGD in Eq. (4), and suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$. Define the following confidence region:*

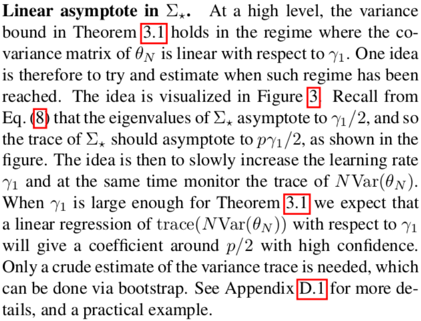
$$\widehat{\Theta} = \big\{\theta \in \Theta : (1/\gamma_1^*)\, ||\theta - \theta_N||^2 < \chi_{\alpha,p}\big\}, \tag{11}$$

where $\chi_{\alpha,p} = \sup\{x \in \mathbb{R} : P(\chi_p^2 \geq x) \leq \alpha\}$ is the α-critical value of a chi-squared random variable with p degrees of freedom. Then,

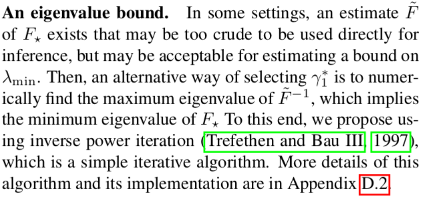
$$\liminf_{N\to\infty} P(\theta_\star \in \widehat{\Theta}) \geq 1 - \alpha. \tag{12}$$

Summary of Chee et al. (2023)

Selecting γ_1^*:



Linear asymptote in $\Sigma_\star$. At a high level, the variance bound in Theorem 3.1 holds in the regime where the covariance matrix of θ_N is linear with respect to γ_1. One idea is therefore to try and estimate when such regime has been reached. The idea is visualized in Figure 3. Recall from Eq. (8) that the eigenvalues of $\Sigma_\star$ asymptote to $\gamma_1/2$, and so the trace of $\Sigma_\star$ should asymptote to $p\gamma_1/2$, as shown in the figure. The idea is then to slowly increase the learning rate γ_1 and at the same time monitor the trace of $N\mathrm{Var}(\theta_N)$. When γ_1 is large enough for Theorem 3.1 we expect that a linear regression of $\mathrm{trace}(N\mathrm{Var}(\theta_N))$ with respect to γ_1 will give a coefficient around $p/2$ with high confidence. Only a crude estimate of the variance trace is needed, which can be done via bootstrap. See Appendix D.1 for more details, and a practical example.



An eigenvalue bound. In some settings, an estimate $\tilde{F}$ of $F_\star$ exists that may be too crude to be used directly for inference, but may be acceptable for estimating a bound on $\lambda_{\min}$. Then, an alternative way of selecting γ_1^* is to numerically find the maximum eigenvalue of $\tilde{F}^{-1}$, which implies the minimum eigenvalue of $F_\star$ To this end, we propose using inverse power iteration (Trefethen and Bau III, 1997), which is a simple iterative algorithm. More details of this algorithm and its implementation are in Appendix D.2.

Thresholding and SGD

- In the context of thresholding we define the pivots

$$\frac{\hat{\beta}_j}{\sqrt{\frac{\gamma_1^*}{N}}}$$

  where we have the usual behaviour for $\hat{\beta}_j$ and the same behaviour from $\sqrt{\frac{\gamma_1^*}{N}} = O(N^{-1/2})$.
- Seems to work.
- Next steps: implementing an iterative version so we can build confidence sets.

Binomial, $p = 10$, $s = 5$

Value	Proportion	Cumulative
(1,2,3,4,5)	0.9132	0.9132
(1,3,4,5)	0.0424	0.9556
()	0.0162	0.9718
(1,4,5)	0.0136	0.9854
(4,5)	0.0050	0.9904
(1)	0.0030	0.9934
(3)	0.0030	0.9964
(1,3,5)	0.0010	0.9974
(3,9)	0.0010	0.9984
(1,3)	0.0005	0.9989
(1,9)	0.0005	0.9994
(9)	0.0005	0.9999

Table: $n = 2000$, 95% CS: {(1,2,3,4,5), (1,3,4,5)}

Binomial, $p = 100$, $s = 5$

Value	Proportion	Cumulative
(1,2,3,4,5)	0.5786	0.5786
(1,2,3,4,5,29)	0.1571	0.7357
()	0.1016	0.8373
(1,2,3,4,5,28)	0.0861	0.9234
(1,2,3,4,5,29,61)	0.0311	0.9545
(1,3,4,5)	0.0205	0.9750
(1,4)	0.0061	0.9811
(1,2,3,4,5,61)	0.0039	0.9850
(5)	0.0039	0.9889
(1,2,3,4,5,7)	0.0028	0.9917
(1)	0.0028	0.9945
(4)	0.0022	0.9967
(1,3,4)	0.0017	0.9984
(1,3)	0.0011	0.9995
(1,3,5)	0.0006	1.0001

Table: $n = 2000$, 95% CS: {(1,2,3,4,5), (1,2,3,4,5,29), (), (1,2,3,4,5,28), (1,2,3,4,5,29,61)}

Binomial, $p = 40$, $s = 25$

Value	Proportion	Cumulative
(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)	0.2988	0.2988
()	0.1603	0.4591
(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25)	0.1046	0.5637
(2,4,14,22)	0.0229	0.5866
(2,4,5,13,14,22)	0.0193	0.6059
(3,14)	0.0187	0.6246
(1,2,4,5,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25)	0.0167	0.6413
(3,4,14)	0.0146	0.6559
(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,23,24,25)	0.0141	0.67
(1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25)	0.0125	0.6825
⋮	⋮	⋮

Table: $n = 2000$, 95% CS: { (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25), (),

(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25), (2,4,14,22), (2,4,5,13,14,22), (3,14),

(1,2,4,5,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25), (3,4,14),

(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,23,24,25), (1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25),

(1,2,4,6,7,8,9,10,11,12,13,14,15,16,17,19,20,21,22,23,24,25), (2,3,4,5,7,9,11,12,13,14,18,22,24,25),

(1,2,3,4,5,6,7,8,10,11,12,13,14,15,17,18,20,21,22,23,24,25), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,20,21,22,23,24,25),

(1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25), (2,3,14), (3),

(1,2,4,7,10,11,12,13,14,15,16,19,20,21,22,23,24,25), (2,14), (3,12,14), (2,4,14,19,22),

(1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,24,25),

(2,4,14,20,22), (3,9,12,14), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,21,22,23,24,25),

(1,2,4,5,7,10,11,12,13,14,15,16,19,20,21,22,23,24,25), (1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25)

References

Chee, J., H. Kim, and P. Toulis (2023, 25–27 Apr). “plus/minus the learning rate”: Easy and scalable statistical inference with sgd. In F. Ruiz, J. Dy, and J.-W. van de Meent (Eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Volume 206 of *Proceedings of Machine Learning Research*, pp. 2285–2309. PMLR.