A Nonlinear PID-Incorporated Adaptive Stochastic Gradient Descent Algorithm for Latent Factor Analysis: Supplementary File

Jinli Li, Xin Luo, *Senior Member*, *IEEE*, Ye Yuan, *Member*, *IEEE*, and Shangce Gao, *Senior Member*, *IEEE*

This is the supplementary file for the paper entitled *A Nonlinear PID-Incorporated Adaptive Stochastic Gradient Descent Algorithm for Latent Factor Analysis*. It mainly contains the convergence analysis of an ANPS-based LFA model.

SUPPLEMENTARY CONVERGENCE ANALYSIS

1) *Proof of Lemma* 1

By replacing x_m in $\varepsilon_{m,n}$ with arbitrary vectors x_g and x_h which are independent of each other, we have:

$$\begin{aligned}&\nabla\varepsilon_{m,n}(x_g)-\nabla\varepsilon_{m,n}(x_h)\\&=(x_g-x_h)(y_n^{\mathrm{T}}y_n+\lambda I_f),\\&\Rightarrow\left\|\nabla\varepsilon_{m,n}(x_g)-\nabla\varepsilon_{m,n}(x_h)\right\|_2\le\left\|(x_g-x_h)\right\|_2\left\|(y_n^{\mathrm{T}}y_n+\lambda I_f)\right\|_2.\end{aligned}\tag{S1}$$

Note that $\left\|y_n^{\mathrm{T}}y_n+\lambda I_f\right\|$ is $(y_n^{\mathrm{T}}y_n+\lambda I_f)$'s largest singular value according to L_2-norm's property. So ***Lemma* 1** holds.□

***Lemma* 2.** $\varepsilon_{m,n}$ is strongly convex if δ is the minimum singular value of $(y_n^{\mathrm{T}}y_n+\lambda I_f)$.

2) *Proof of Lemma* 2

Given arbitrary vectors x_g and x_h, we consider the state of $\varepsilon_{m,n}$ at x_h following the principle of Taylor-series as:

$$\begin{aligned}\varepsilon_{m,n}(x_g)&\approx\varepsilon_{m,n}(x_h)+\nabla\varepsilon_{m,n}(x_h)(x_g-x_h)^{\mathrm{T}}\\&\quad+\frac{1}{2}(x_g-x_h)\nabla^2\varepsilon_{m,n}(x_h)(x_g-x_h)^{\mathrm{T}},\\\Rightarrow\varepsilon_{m,n}(x_g)-\varepsilon_{m,n}(x_h)&=\nabla\varepsilon_{m,n}(x_h)(x_g-x_h)^{\mathrm{T}}\\&\quad+\frac{1}{2}(x_g-x_h)\nabla^2\varepsilon_{m,n}(x_h)(x_g-x_h)^{\mathrm{T}}.\end{aligned}\tag{S2}$$

Recall ***Definition* 4**, the instant objective $\varepsilon_{m,n}$'s strong-convexity satisfies the following condition:

$$\varepsilon_{m,n}(x_g)-\varepsilon_{m,n}(x_h)\ge\nabla\varepsilon_{m,n}(x_h)(x_g-x_h)^{\mathrm{T}}+\frac{1}{2}\delta\left\|x_g-x_h\right\|_2^2.\tag{S3}$$

Hence, δ is desired to make the following condition hold:

$$(x_g-x_h)\nabla^2\varepsilon_{m,n}(x_h)(x_g-x_h)^{\mathrm{T}}\ge\delta\left\|x_g-x_h\right\|_2^2.\tag{S4}$$

From the expression of $\varepsilon_{m,n}$, we achieve:

$$\nabla^2\varepsilon_{m,n}(x_h)=y_n^{T}y_n+\lambda I_f.\tag{S5}$$

Thus, we achieve the following equation:

$$\underbrace{(x_g-x_h)}_{a}\underbrace{(y_n^{\mathrm{T}}y_n+\lambda I_f-\delta I_f)}_{b}\underbrace{(x_g-x_h)^{\mathrm{T}}}_{c}\ge 0,\tag{S6}$$

which is equivalent to prove that part (b) of (S6) is a positive semi-definite matrix. As unveiled by [50, 51], part (b) is a positive semi-definite matrix if δ is the minimum singular value of matrix $(y_n^{\mathrm{T}}y_n+\lambda I_f)$. Thus, ***Lemma* 2** holds.□

3) *Proof of Theorem* 1

Considering the ι-th training iteration, $\forall r_{m,n}\in\Lambda$ and $j\in\{1,\ldots,J\}$, x_m is updated with the following scheme:

$$x_{(j)m}^{\kappa}\leftarrow x_{(j)m}^{\kappa-1}-\eta\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right),\tag{S7}$$

where $x_{(j)m}^{\kappa}$ denotes $x_{(j)m}$'s state after its κ-th update point. Assume that x_m^* is the local optimal state of x_m, then we have:

$$\begin{aligned}\left\|x_{(j)m}^{\kappa}-x_m^*\right\|_2^2&=\left\|x_{(j)m}^{\kappa-1}-\eta\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)-x_m^*\right\|_2^2\\&=\left\|x_{(j)m}^{\kappa-1}-x_m^*\right\|_2^2+\eta^2\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2^2\\&\quad-2\eta\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\left(x_{(j)m}^{\kappa-1}-x_m^*\right)^{\mathrm{T}}.\end{aligned}\tag{S8}$$

According to ***Definition* 4** and ***Lemma* 2**, we have:

$$\begin{aligned}\varepsilon_{m,n}(x_m^*)&\ge\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)+\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\left(x_m^*-x_{(j)m}^{\kappa-1}\right)^{\mathrm{T}}\\&\quad+\frac{1}{2}\delta\left\|x_m^*-x_{(j)m}^{\kappa-1}\right\|_2^2.\end{aligned}\tag{S9}$$

Since x_m^* is the optimal state of x_m, we achieve that:

$$\begin{cases}\varepsilon_{m,n}(x_m^*)<\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right),\\\nabla\varepsilon_{m,n}(x_m^*)=0.\end{cases}\tag{S10}$$

Based on (S9)-(S10), the following inequality is achieved:

$$\nabla\varepsilon_{m,n}\left(x_{(j)m}^{(\kappa-1)}\right)\left(x_{(j)m}^{(\kappa-1)}-x_m^*\right)^{\mathrm{T}}\ge\frac{1}{2}\delta\left\|x_{(j)m}^{(\kappa-1)}-x_m^*\right\|_2^2.\tag{S11}$$

By combining (S8) and (S11), we infer that:

$$\left\|x_{(j)m}^{\kappa}-x_m^*\right\|_2^2<(1-\eta\delta)\left\|x_{(j)m}^{\kappa-1}-x_m^*\right\|_2^2+\eta^2\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2^2.$$

(S12)

By taking the expectation of (S12), we further have:

$$\mathrm{E}\left[\left\|x_{(j)m}^{\kappa}-x_m^{*}\right\|_2^2\right]<(1-\eta\delta)\,\mathrm{E}\left[\left\|x_{(j)m}^{\kappa-1}-x_m^{*}\right\|_2^2\right]+\eta^2\mathrm{E}\left[\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2^2\right]. \tag{S13}$$

Following [49], assume that a positive number p fulfills:

$$\mathrm{E}\left[\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2^2\right]\le p^2, \tag{S14}$$

then (S13) is equivalent to:

$$\mathrm{E}\left[\left\|x_{(j)m}^{\kappa}-x_m^{*}\right\|_2^2\right]<(1-\eta\delta)\,\mathrm{E}\left[\left\|x_{(j)m}^{\kappa-1}-x_m^{*}\right\|_2^2\right]+(\eta p)^2. \tag{S15}$$

By making $\eta=\alpha/(\delta\iota)$ with α>1, (S15) is reformulated as:

$$\mathrm{E}\left[\left\|x_{(j)m}^{\kappa}-x_m^{*}\right\|_2^2\right]<\left(1-\frac{\alpha}{\iota}\right)\mathrm{E}\left[\left\|x_{(j)m}^{\kappa-1}-x_m^{*}\right\|_2^2\right]+\left(\frac{\alpha}{\delta\iota}p\right)^2. \tag{S16}$$

By expanding (S16), we achieve the following condition:

$$\mathrm{E}\left[\left\|x_{(j)m}^{\kappa}-x_m^{*}\right\|_2^2\right]\le\frac{1}{\iota}\max\left\{\left\|x_{(j)m}^{1}-x_m^{*}\right\|_2^2,\frac{\alpha^2p^2}{\delta\alpha-1}\right\}. \tag{S17}$$

where $x_{(j)m}^{1}$ is the initial state of $x_{(j)m}$ at the ι-th training iteration. Since ***Lemma* 1** indicates that $\varepsilon_{m,n}$ is L-smooth, we achieve the following inference:

$$\begin{aligned}&\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)-\varepsilon_{m,n}\left(x_m^{*}\right)\le\frac{L}{2}\left\|x_{(j)m}^{\kappa}-x_m^{*}\right\|_2^2\\&\Rightarrow\mathrm{E}\left[\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)-\varepsilon_{m,n}\left(x_m^{*}\right)\right]\le\frac{L}{2}\mathrm{E}\left[\left\|x_{(j)m}^{\kappa}-x_m^{*}\right\|_2^2\right].\end{aligned} \tag{S18}$$

By substituting (S17) into (S18), we achieve that:

$$\mathrm{E}\left[\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)-\varepsilon_{m,n}\left(x_m^{*}\right)\right]\le\frac{LQ(\alpha)}{2\iota}, \tag{S19}$$

which satisfies the following condition:

$$Q(\alpha)=\max\left\{\left\|x_{(j)m}^{1}-x_m^{*}\right\|_2^2,\frac{\alpha^2p^2}{\delta\alpha-1}\right\}. \tag{S20}$$

Let $\Lambda(m)$ be the subset of Λ related to $m\in M$, then (S19) can be expanded into:

$$\mathrm{E}\left[\sum_{r_{m,n}\in\Lambda(m)}\left(\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)-\varepsilon_{m,n}\left(x_m^{*}\right)\right)\right]\le\frac{|\Lambda(m)|LQ(\alpha)}{2\iota}. \tag{S21}$$

Note that we have $\sum_{m\in M}|\Lambda(m)|=|\Lambda|$. So (S21) can be further expanded into:

$$\begin{aligned}&\mathrm{E}\left[\sum_{m\in M}\sum_{r_{m,n}\in\Lambda(m)}\left(\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)-\varepsilon_{m,n}\left(x_m^{*}\right)\right)\right]\\&=\mathrm{E}\left[\sum_{r_{m,n}\in\Lambda}\left(\varepsilon_{m,n}\left(x_{(j)m}^{\kappa}\right)-\varepsilon_{m,n}\left(x_m^{*}\right)\right)\right]\le\frac{|\Lambda|LW(\alpha)}{2\iota}.\end{aligned} \tag{S22}$$

With (S22) we have $\frac{|\Lambda|LW(\alpha)}{2\iota}\to0$ when $\iota\to\infty$, indicating the convergence with x_m. Note that the convergence with y_n can be achieved in the same way. Thus, ***Theorem* 1** holds.□

4) *Proof of Theorem* 2

Considering (8), we infer the following expression.

$$\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)=\lambda\cdot x_{(j)m}^{\kappa-1}-e_{m,n}^{\iota}\cdot y_{(j)n}. \tag{S23}$$

According to the L_2-norm's properties, we have the following inequality:

$$\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2\le\lambda\left\|x_{(j)m}^{\kappa-1}\right\|_2+\left\|e_{m,n}^{\iota}\cdot y_{(j)n}\right\|_2. \tag{S24}$$

Since $y_{(j)n}$ is considered as a constant when $x_{(j)m}$ is the active optimization parameter, we can reformulate (S24) as:

$$\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2\le\lambda\left\|x_{(j)m}^{\kappa-1}\right\|_2+y_{(j)n}\left\|e_{m,n}^{\iota}\right\|_2. \tag{S25}$$

From (S25), we evidently infer that an ANPS-based LFA model converges with the SGD-based learning scheme if $\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2$ has an upper bound. Hence, to prove **Aspect 2** is equivalent to prove that $\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2$ has an upper bound after refining the estimation learning error by NPID.

Based on (S14) and (S25), we have the following inference

$$\left\|x_{(j)m}^{\kappa-1}\right\|_2\le\alpha,\ \left\|e_{m,n}^{\iota}\right\|_2\le\beta; \tag{S26}$$

where α and β are both finite positive numbers. Following (15), after refining the estimation learning error by an NPID controller, (S23) can be reformulated as:

$$\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)=\lambda\cdot x_{(j)m}^{\kappa-1}-\tilde{e}_{(j)m,n}^{\iota}\cdot y_{(j)n}. \tag{S27}$$

And based on the L_2-norm's properties, (S25) naturally satisfies the following condition:

$$\left\|\nabla\varepsilon_{m,n}\left(x_{(j)m}^{\kappa-1}\right)\right\|_2\le\lambda\cdot\left\|x_{(j)m}^{\kappa-1}\right\|_2+y_{(j)n}\cdot\left\|\tilde{e}_{(j)m,n}^{\iota}\right\|_2. \tag{S28}$$

On the other hand, for conveniently analyzing the convergence of an ANPS-based LFA model, we simplify the NPID-based learning error refinement (15) as:

$$\begin{aligned}\tilde{e}_{(j)m,n}^{\iota}=&K_{(j)P}\cdot e_{(j)m,n}^{\iota}+K_{(j)I}\cdot\sum_{k=0}^{\iota}e_{(j)m,n}^{k}\\&+K_{(j)D}\cdot\left(e_{(j)m,n}^{\iota}-e_{(j)m,n}^{\iota-1}\right),\end{aligned} \tag{S29}$$

$$\begin{cases}K_{(j)P}=K_{(j)p1}^{\iota-1}+K_{(j)p2}^{\iota-1}\left(1-\operatorname{sech}\left(K_{(j)p3}^{\iota-1}e_{(j)m,n}^{\iota}\right)\right),\\K_{(j)I}=K_{(j)i1}^{\iota-1}\operatorname{sech}\left(K_{(j)i2}^{\iota-1}e_{(j)m,n}^{\iota}\right),\\K_{(j)D}=\left(K_{(j)d1}^{\iota-1}+K_{(j)d2}^{\iota-1}\right)\Big/\left(1+K_{(j)d3}^{\iota-1}\exp\left(K_{(j)d4}^{\iota-1}e_{(j)m,n}^{\iota}\right)\right).\end{cases} \tag{S30}$$

Note that the value range of sech(·) lies in (0, 1). Hence, K_P, K_I, and K_D are all finite positive numbers. Considering the L_2-norm of (S29), we have:

$$\begin{aligned}\left\|\tilde{e}_{(j)m,n}^{\iota}\right\|_2=&\left\|K_{(j)p}\left(e_{(j)m,n}^{\iota}\right)\cdot e_{(j)m,n}^{\iota}\right\|_2+\left\|K_i\left(e_{(j)m,n}^{\iota}\right)\cdot\sum_{k=0}^{\iota}e_{(j)m,n}^{k}\right\|_2\\&+\left\|K_d\left(e_{(j)m,n}^{\iota}\right)\cdot\left(e_{(j)m,n}^{\iota}-e_{(j)m,n}^{\iota-1}\right)\right\|_2.\end{aligned} \tag{S31}$$

And based on the L_2-norm's properties, (S31) satisfies the following condition:

$$\begin{aligned}\left\|\tilde{e}_{(j)m,n}^{\iota}\right\|_2\le&K_{(j)p}\left(e_{(j)m,n}^{\iota}\right)\left\|e_{(j)m,n}^{\iota}\right\|_2+K_i\left(e_{(j)m,n}^{\iota}\right)\left\|\sum_{k=0}^{\iota}e_{(j)m,n}^{k}\right\|_2\\&+K_d\left(e_{(j)m,n}^{\iota}\right)\left\|e_{(j)m,n}^{\iota}-e_{(j)m,n}^{\iota-1}\right\|_2,\end{aligned}$$

$$ \tag{S32} $$

With (S26), we can reformulate (S32) into:

$$\begin{aligned} \left\|\tilde{e}^{\,\iota}_{j\ m,n}\right\|_2 \le & K_{\ j\ p}\ \ e^{\,\iota}_{j\ m,n}\ \ \cdot\beta+\iota\cdot K_i\ \ e^{\,\iota}_{j\ m,n}\ \ \cdot\beta \\ & +2\cdot K_d\ \ e^{\,\iota}_{j\ m,n}\ \ \cdot\beta. \end{aligned} \tag{S33}$$

Hence, $\left\|\tilde{e}^{\,\iota}_{j\ m,n}\right\|_2$ has an upper bound as:

$$\left\|\tilde{e}^{\,\iota}_{j\ m,n}\right\|_2 \le \chi, \tag{S34}$$

where χ is a finite positive number. By substituting (S26) and (S34) into (S28), we deduce that:

$$\left\|\nabla\varepsilon_{m,n}\ \ x^{\kappa-1}_{j\ m}\ \ \right\|_2 \le \lambda\cdot\alpha+y_{\ j\ n}\cdot\chi. \tag{S35}$$

Hence, there always exists a positive number θ to fulfill:

$$\left\|\nabla\varepsilon_{m,n}\ \ x^{\kappa-1}_{j\ m}\ \ \right\|_2 \le \theta. \tag{S36}$$

Thus, ***Theorem* 2** holds.□