Missing Feature Estimation-Reinforced Online Sparse Streaming Feature Selection

Ruiyang Xu, Di Wu, *Member, IEEE,* and Xin Luo, *Fellow, IEEE*

I. INTRODUCTION

This is the supplementary file for the paper entitled "*Missing Feature Estimation-Reinforced Online Sparse Streaming Feature Selection*". It mainly contains the figures of experimental results.

II. SUPPLEMENTARY TABLES

TABLE S(I) MEAN NUMBER OF SELECTED FEATURES VARYING WITH DIFFERENT ALGORITHMS, Ψ=0.1.

Models/Datasets	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	Average
G1+M1	7.40	12.70	14.70	18.30	23.60	8.70	36.10	44.90	43.00	78.30	54.80	11.70	29.52
G2+M1	6.50	2.00	6.00	4.60	6.00	4.00	4.00	4.70	6.00	6.40	4.10	1.00	4.61
M1	26.30	4.00	8.00	22.50	9.00	7.40	10.20	7.70	6.00	11.60	8.00	1.00	10.14
G1+M2	10.90	10.20	12.40	16.90	20.10	8.20	32.50	36.70	37.20	64.00	53.33	90.67	32.76
G2+M2	11.20	2.00	6.00	4.00	4.10	3.00	8.00	12.00	10.00	6.00	5.00	6.67	6.50
M2	75.50	3.00	31.00	74.90	16.00	18.50	16.00	37.00	24.80	54.20	61.00	47.00	38.24
G1+M3	8.20	12.70	13.80	22.50	22.20	8.67	37.00	44.30	45.60	87.20	56.67	89.67	37.38
G2+M3	197.30	29.00	22.50	15.00	27.00	1.00	1569.70	14.70	50.70	39.20	69.67	198.67	186.20
M3	102.20	16.00	1.00	1.00	1.00	27.30	125.40	1.00	1.00	1.00	1.00	1.00	23.24
G1+M4	10.90	29.10	10.70	39.00	13.70	29.80	54.20	10.60	24.20	34.70	40.00	25.33	26.85
G2+M4	11.20	27.40	9.80	20.30	15.50	33.80	49.50	10.40	30.50	39.90	37.40	34.33	26.67
M4	75.50	32.20	17.80	54.90	25.20	45.30	50.70	13.30	22.80	31.10	30.80	29.67	35.77
G1+M5	25.60	14.50	15.00	65.80	45.50	44.30	19.50	16.10	39.60	89.70	53.00	33.00	38.47
G2+M5	26.70	42.60	14.10	67.60	62.10	39.60	24.20	19.50	32.70	86.80	51.50	21.33	40.73
M5	34.40	6.50	5.20	8.80	8.20	21.40	8.10	16.20	20.90	137.30	8.90	46.70	26.88
G1+M6	28.90	7.80	4.10	6.80	5.00	10.70	11.20	3.10	3.10	5.40	9.50	14.00	9.13
G2+M6	26.10	8.00	4.60	7.30	6.00	9.90	9.00	3.00	4.00	4.40	9.40	12.33	8.67
M6	33.30	11.40	8.90	12.50	1.00	9.00	9.00	7.00	4.30	7.90	6.00	8.00	9.86

TABLE S(II) USING THE SELECTED FEATURES (RECORDED IN TABLE S(I)) TO TRAIN A CLASSIFIER FIRST AND THEN TESTING ITS ACCURACY (%), Ψ=0.1.

Models /Datasets	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	Average
G1+M1	**87.03**±2.82	**82.01**±3.97	**82.07**±7.73	**87.06**±2.13	**74.20**±4.17	**73.33**±4.20	**87.25**±2.81	**95.05**±1.56	**93.96**±2.60	**97.58**±0.88	**82.36**±5.51	**81.10**±5.16	**85.50**±3.57
G2+M1	84.87±2.63	80.45±2.59	78.58±2.23	84.79±2.77	67.56±3.28	72.30±2.40	86.64±1.61	94.95±2.22	92.87±1.71	97.41±0.70	81.01±3.03	67.02±2.76	82.37±2.33
M1	71.08±2.22	78.88±2.59	75.09±3.84	84.40±2.43	56.67±2.79	71.41±3.88	87.05±2.46	94.97±2.26	92.47±2.23	97.00±0.89	79.29±3.24	63.29±3.53	79.30±2.70
G1+M2	**93.70**±1.04	**82.18**±4.47	**78.42**±5.66	**88.06**±1.89	**71.78**±4.39	**73.41**±5.03	**87.54**±2.61	**94.76**±2.47	**95.16**±2.23	**97.64**±0.51	**82.88**±5.63	**83.14**±1.11	**85.72**±3.09
G2+M2	93.62±0.91	80.11±2.57	76.97±4.45	84.47±3.11	61.25±4.13	72.00±2.64	86.57±1.89	94.28±1.11	94.53±1.61	97.39±0.83	81.78±5.89	80.78±7.76	83.65±3.08
M2	88.59±0.49	77.65±3.19	76.51±3.32	83.38±2.32	61.05±5.16	71.37±3.17	86.23±2.80	93.53±1.61	93.14±2.18	96.10±0.84	80.15±3.83	82.75±1.66	82.54±2.55
G1+M3	**85.40**±1.65	**80.77**±3.50	**79.72**±3.97	**87.76**±2.34	**74.66**±3.91	**72.59**±4.86	**86.13**±2.04	**89.57**±3.58	**90.18**±3.41	**97.20**±0.96	**78.10**±7.87	**77.25**±2.72	**83.28**±3.40
G2+M3	83.61±4.78	79.86±2.59	63.78±18.08	85.78±2.00	58.57±15.01	61.48±6.51	83.32±2.81	73.31±12.43	78.79±14.10	95.02±2.86	77.57±14.83	72.68±6.36	76.15±8.53
M3	80.56±3.48	78.63±2.66	50.13±4.49	62.49±2.96	45.57±1.06	60.59±5.96	83.26±7.68	65.95±2.97	62.45±5.53	83.21±2.00	47.53±3.83	62.22±4.98	65.22±3.97
G1+M4	**93.70**±1.04	**78.18**±5.27	**88.45**±4.03	**93.07**±1.29	**77.34**±5.38	**68.41**±3.77	**82.61**±3.24	**91.54**±3.17	**88.41**±5.85	**96.63**±1.08	**79.98**±6.30	**79.22**±4.14	**84.80**±3.71
G2+M4	93.62±0.91	75.97±5.36	85.57±5.35	90.76±2.11	75.74±4.28	68.11±4.01	81.71±3.64	90.09±4.40	85.87±4.79	96.06±1.43	78.41±5.63	76.73±5.50	83.22±3.95
M4	88.59±0.49	74.07±5.52	84.50±5.74	85.47±3.11	71.03±5.22	68.30±4.10	79.81±5.46	88.53±4.23	83.00±5.81	92.99±2.29	75.89±6.80	73.86±7.60	80.50±4.70
G1+M5	**94.47**±0.71	**84.52**±4.78	**91.99**±3.57	**93.10**±1.28	**82.78**±3.36	**81.04**±3.14	**89.05**±2.32	**95.96**±1.96	**94.89**±2.24	**97.51**±0.73	**85.80**±4.39	**84.05**±4.22	**89.60**±2.73
G2+M5	94.19±0.80	80.38±4.82	89.47±4.24	92.61±1.66	81.75±2.92	79.52±3.76	87.67±2.84	94.29±2.06	93.11±2.34	97.23±1.04	85.77±4.17	83.66±4.49	88.30±2.93
M5	81.17±1.85	79.66±5.19	81.58±6.76	83.84±2.52	69.06±5.31	78.30±3.06	88.94±1.81	93.43±4.01	88.22±3.54	96.03±1.29	84.41±3.31	83.01±4.00	83.97±3.55
G1+M6	**95.00**±0.62	**83.66**±4.30	**91.91**±3.16	**92.99**±1.48	**78.79**±3.51	70.30±4.40	**88.08**±2.51	**95.44**±1.87	**95.98**±1.98	**97.93**±0.72	**77.73**±7.39	75.03±3.65	**86.90**±2.97
G2+M6	92.23±4.39	81.67±3.62	88.42±4.49	91.62±1.94	71.99±3.02	68.93±5.10	86.83±1.78	94.44±1.65	94.18±2.00	96.26±1.17	75.90±6.51	72.16±7.73	84.55±3.62
M6	81.38±3.67	79.82±2.68	87.10±3.47	86.78±3.00	49.16±1.78	**71.22**±3.80	85.32±3.61	94.08±1.68	93.47±2.96	96.13±1.18	77.37±4.52	**83.13**±0.55	82.28±2.59

TABLE S(III) THE RANK SUM OF THE WILCOXON SIGNED-RANKS

Models/ψ	0.3		0.5		0.7		0.9	
	*R^+	*R^-	*R^+	*R^-	*R^+	*R^-	*R^+	*R^-
G1+M1	-		-		-		-	
G2+M1	78	0	78	0	59	19	77	1
M1	78	0	78	0	71	7	67	11
G1+M2	-		-		-		-	
G2+M2	78	0	78	0	74	4	77	1
M2	78	0	78	0	71	7	60	18
G1+M3	-		-		-		-	
G2+M3	78	0	78	0	67	11	78	0
M3	78	0	78	0	78	0	78	0
G1+M4	-		-		-		-	
G2+M4	78	0	78	0	77	1	78	0
M4	76	2	78	0	77	1	69	9
G1+M5	-		-		-		-	
G2+M5	78	0	78	0	70	8	78	0
M5	78	0	78	0	78	0	67	11
G1+M6	-		-		-		-	
G2+M6	78	0	78	0	78	0	75	3
M6	70	8	67	11	62	16	69	9

* If min{R^+, R^-}>18, the null hypothesis will be taken.

TABLE S(IV) THE AVERAGE ACCURACY OF SELECTED FEATURES VARYING WITH DIFFERENT PARAMETERS OF THE MAPPING FUNCTION.

θ/Models	G1+M1	G1+M2	G1+M3	G1+M4	G1+M5	G1+M6	Average	^Rank
1.00	84.27±2.74	83.58±3.30	**83.55**±3.68	84.64±3.48	87.92±2.75	86.25±3.70	85.04±3.28	3.17
1.25	**85.49**±3.57	**85.72**±3.09	83.28±3.40	**84.80**±3.71	**89.60**±2.73	86.90±2.97	**85.97**±3.25	1.33
1.50	83.35±3.49	83.55±3.52	82.35±3.50	83.71±3.23	88.91±2.87	86.78±2.87	84.78±3.25	3.83
1.75	83.41±3.15	83.63±3.03	82.85±3.61	84.06±3.75	88.08±2.70	**87.31**±2.97	84.89±3.20	2.83
2.00	83.11±3.25	83.88±3.07	81.70±3.47	83.97±3.38	88.38±2.40	86.61±3.17	84.61±3.12	3.83

^ The Average rank.

TABLE S(V) THE AVERAGE ACCURACY OF SELECTED FEATURES VARYING WITH DIFFERENT CONTROLLING PARAMETERS.

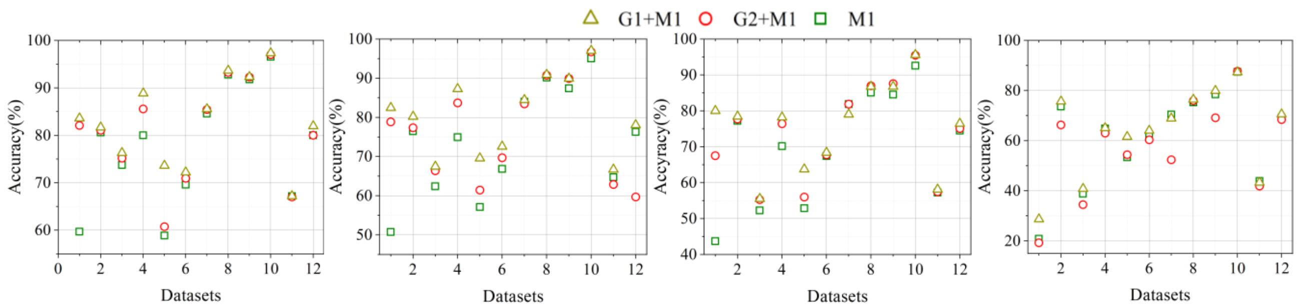
η/Models	G1+M1	G1+M2	G1+M3	G1+M4	G1+M5	G1+M6	Average	^Rank
0.10	84.35±3.22	84.07±3.13	82.65±3.34	83.55±3.56	88.97±2.62	**87.15**±3.06	85.12±3.16	2.83
0.25	83.82±3.42	83.47±2.84	82.44±3.43	84.19±3.46	88.85±2.43	85.42±3.16	84.70±3.12	4.33
0.50	**85.50**±3.57	**85.72**±3.09	**83.28**±3.40	**84.80**±3.71	**89.60**±2.73	86.90±2.97	**85.97**±3.25	1.17
0.75	84.25±3.56	83.28±3.43	82.22±3.92	84.23±3.82	88.95±2.45	85.88±3.53	84.80±3.45	3.83
1.00	84.52±3.17	83.51±3.72	82.77±3.16	84.18±3.70	89.14±2.45	85.64±2.72	84.96±3.15	2.83

^ The Average rank.

TABLE S(VI)THE AVERAGE ACCURACY OF SELECTED FEATURES VARYING WITH DIFFERENT *P*.

P/Models	G1+M1	G1+M2	G1+M3	G1+M4	G1+M5	G1+M6	Average	^Rank
100	83.61±3.82	84.13±3.18	82.35±2.93	82.34±3.44	87.04±2.97	85.29±3.51	84.13±3.31	3.00
200	**85.50**±3.57	**85.72**±3.09	**83.28**±3.40	84.80±3.71	89.60±2.73	86.90±2.97	**85.97**±3.25	1.83
300	82.43±3.12	83.83±3.33	81.24±3.44	**85.04**±3.27	89.81±2.63	86.95±3.59	84.88±3.23	2.33
400	80.69±3.21	81.38±4.27	79.53±4.30	84.61±3.31	**89.84**±2.59	**87.74**±3.59	83.97±3.55	2.83

III. SUPPLEMENTARY FIGURES



(a) ψ=0.3 (b) ψ=0.5 (c) ψ=0.7 (d) ψ=0.9

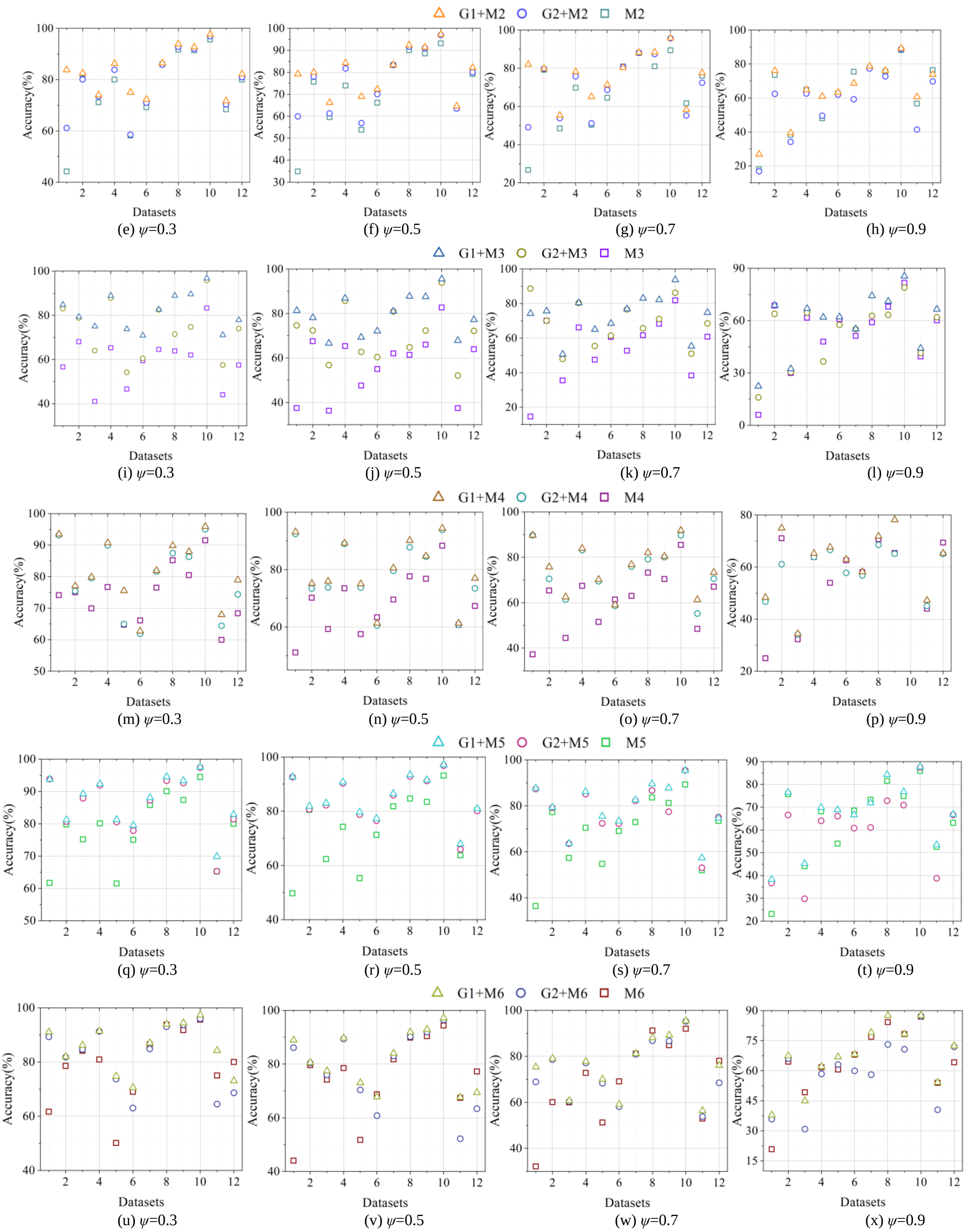


Fig. S1. The average accuracy comparison of both OS^2FS and OSFS model on each dataset.

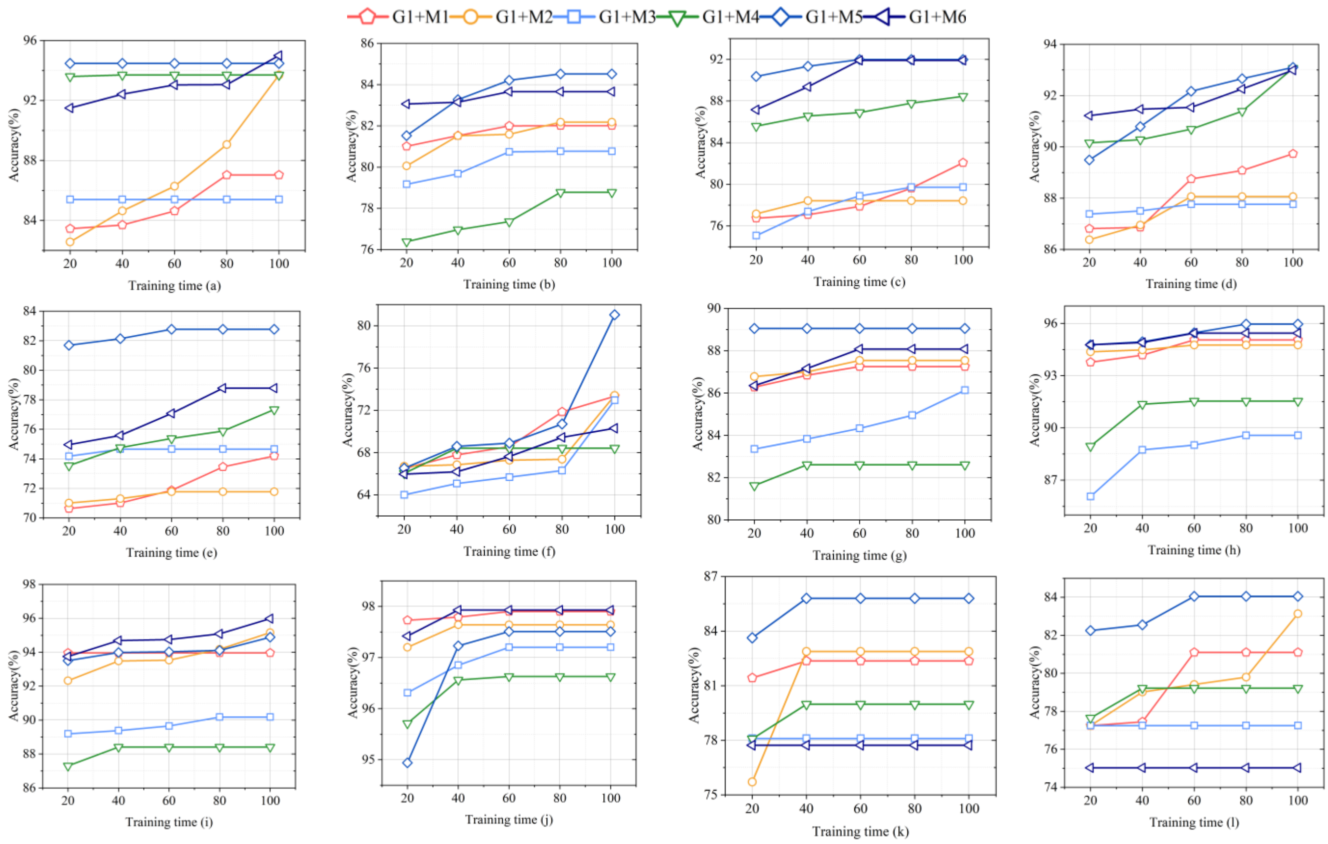


Fig. S2. Average accuracy as d increases from 20 to 100 on all the datasets at different layers. (a) D1. (b) D2. (c) D3. (d) D4. (e) D5. (f) D6. (g) D7. (h) D8. (i) D9. (j) D10. (k) D11. (l) D12.

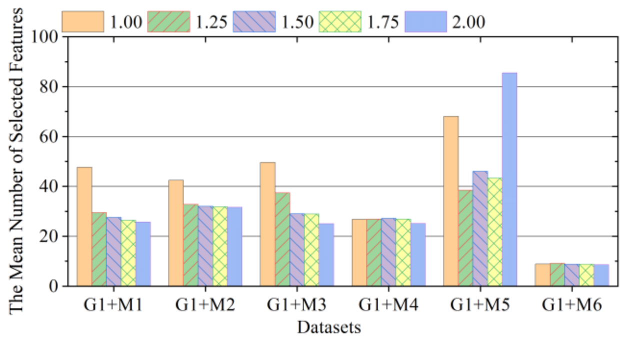


Fig. S3. Mean number of selected features varying with different θ.

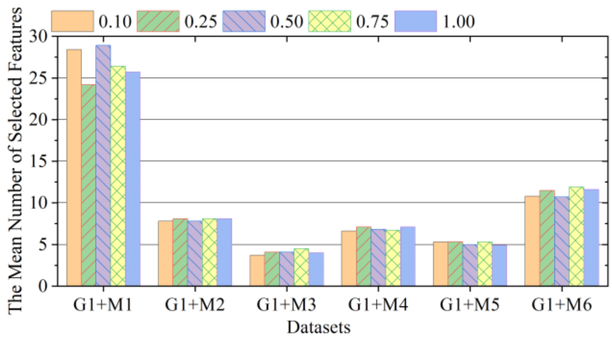


Fig. S4. Mean number of selected features varying with different η.

IV. Supplementary Appendix

First, let's give the definitions of L-smooth and strong convex functions as follows.

***Definition* 9** (**L-smooth function $f(\varphi)$**). If $f(\varphi)$ is L-smooth, satisfying the following condition:

$$\forall \varphi_1,\varphi_2 \in \mathbb{R}^d \ \ s.t. \ \left\|\nabla f\left(\varphi_1\right)-\nabla f\left(\varphi_2\right)\right\|_2 \le L\left\|\varphi_1-\varphi_2\right\|_2. \tag{12}$$

***Definition* 10** (**Strong convex $f(\varphi)$**). Given a strong convex function $f(\varphi)$ as δ>0 satisfying

$$\forall \varphi_1,\varphi_2 \in \mathbb{R}^d \ \ s.t. \ f\left(\varphi_1\right) \ge f\left(\varphi_2\right)+\nabla f\left(\varphi_2\right)\left(\varphi_1-\varphi_2\right)^{\mathrm{T}}+\frac{1}{2}\delta\left\|\varphi_1-\varphi_2\right\|_2^2. \tag{13}$$

***Lemma* 1**. The instant loss $\varepsilon_{h,j}$ is L-smooth, where L is the largest singular value of the matrix ($v_{j,\cdot}^{d\ \mathrm{T}} v_{j,\cdot}^{d} + \lambda E_l$) and E_l is $l\times l$ identity matrix.

Proof. Given two arbitrary and independent row vectors $u_{\alpha,\cdot}^{d}$.

and $u^d_{\beta,\cdot}$ of B_t, we have (14), as shown at the bottom of the page.

For simplifying the formula (14), set η=1, and derive the following formula.

$$\left\|\nabla\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)-\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\right\|_2=\left\|\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\left(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l\right)\right\|_2. \tag{15}$$

Based on the L_2-norm properties of matrix [1], it is further deduced as follows:

$$\left\|\nabla\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)-\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\right\|_2\le\left\|\left(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l\right)\right\|_2\left\|u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right\|_2, \tag{16}$$

where $\|\cdot\|_2$ calculates the maximum singular value. According to the above discussion, $L=\|(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l)\|_2$. Hence, *Lemma* 1 holds.

***Lemma* 2**. If δ is the smallest singular value of $(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l)$, the instantaneous loss $\varepsilon_{h,j}$ is strong-convexity.

Proof. Suppose there are two arbitrary vectors $u^d_{\alpha,\cdot}$ and $u^d_{\beta,\cdot}$, and follow the principle of Taylor series to study $\varepsilon_{m,j}$ on $u^d_{\alpha,\cdot}$:

$$\begin{aligned}&\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)\approx\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)+\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}\\&\qquad+\frac{1}{2}\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\nabla^2\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}.\\&\Rightarrow\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)-\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)=\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}\\&\qquad+\frac{1}{2}\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\nabla^2\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}.\end{aligned} \tag{17}$$

According to *Definition* 10, $\varepsilon_{h,j}$ is strong convex, which is expressed as follows:

$$\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)-\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\ge\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}+\frac{1}{2}\delta\left\|u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right\|_2^2. \tag{18}$$

Therefore, *Lemma* 2 is equal to choosing an appropriate δ value, and the following inequality holds:

$$\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\nabla^2\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}\ge\delta\left\|u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right\|_2^2. \tag{19}$$

From expression $\varepsilon_{h,j}$, it can be concluded that:

$$\nabla^2\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)=v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l. \tag{20}$$

By combining Eq. (19) and (20), and then we have

$$\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\left(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}\ge\delta\left\|u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right\|_2^2. \tag{21}$$

Furthermore, Eq. (21) can be expressed as

$$\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\left(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l-\delta E_l\right)\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)^{\mathrm{T}}\ge 0. \tag{22}$$

Eq. (22) is equivalent to proving that $(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l-\delta E_l)$ is a positive semi-definite matrix. As shown in reference [1], when δ is the smallest singular value of matrix $(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l)$, $(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l-\delta E_l)$ satisfies positive semi-definiteness. Hence, *Lemma* 2 holds.

At the timestamp t, considering the d-th iteration of CM-OS[2]FS and following the training rules of $u^d_{h,\cdot}$, for a single element $\forall\langle h,j\rangle\in K_t$ or O_t, there are:

$$u^{d,\tau}_{h,\cdot}\leftarrow u^{d,\tau-1}_{h,\cdot}-\zeta^{t-1}\cdot\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right). \tag{23}$$

Among them, the τ-th and (τ-1)-th entry updates in the t-th iteration of $u^d_{h,\cdot}$ express as $u^{d,\tau}_{h,\cdot}$ and $u^{d,\tau-1}_{h,\cdot}$ respectively, and $u^{d,*}_{h,\cdot}$ is the optimal update state of $u^d_{h,\cdot}$.

$$\begin{aligned}\left\|u^{d,\tau}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2&=\left\|u^{d,\tau-1}_{h,\cdot}-\eta^{t-1}\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)-u^{d,*}_{h,\cdot}\right\|_2^2\\&=\left\|u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2-2\eta^{t-1}\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\left(u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right)^{\mathrm{T}}\\&\quad+\left(\eta^{t-1}\right)^2\left\|\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\right\|_2^2.\end{aligned} \tag{24}$$

According to *Lemma* 2, the following formula is derived as:

$$\begin{aligned}&\varepsilon_{h,j}\left(u^{d,*}_{h,\cdot}\right)-\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\ge\\&\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\left(u^{d,*}_{h,\cdot}-u^{d,\tau-1}_{h,\cdot}\right)^{\mathrm{T}}+\frac{1}{2}\delta\left\|u^{d,*}_{h,\cdot}-u^{d,\tau-1}_{h,\cdot}\right\|_2^2.\end{aligned} \tag{25}$$

Based on the optimal state $u^d_{h,\cdot}$ of $u^{d,*}_{h,\cdot}$, we see that

$$\begin{cases}\nabla\varepsilon_{h,j}\left(u^{d,*}_{h,\cdot}\right)=0,\\ \varepsilon_{h,j}\left(u^{d,*}_{h,\cdot}\right)<\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right).\end{cases} \tag{26}$$

By combining (26) with (25), we have

$$\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\left(u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right)^{\mathrm{T}}\ge\frac{1}{2}\delta\left\|u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2. \tag{27}$$

Hence, based on (27), (24) is represented as

$$\left\|u^{\tau}_{m,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2\le\left(1-\zeta^{t-1}\delta\right)\left\|u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2+\left(\zeta^{t-1}\right)^2\left\|\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\right\|_2^2. \tag{28}$$

Next, take the expectation of Eq. (28), we have that

$$\begin{aligned}\mathrm{E}\left[\left\|u^{d,\tau}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2\right]&\le\left(1-\zeta^{t-1}\delta\right)\mathrm{E}\left[\left\|u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2\right]\\&\quad+\left(\zeta^{t-1}\right)^2\mathrm{E}\left[\left\|\nabla\varepsilon_{h,j}\left(u^{d,\tau-1}_{h,\cdot}\right)\right\|_2^2\right].\end{aligned} \tag{29}$$

According to the conclusion of reference [2], suppose there is a positive number z such that

$$\mathrm{E}\left[\left\|\nabla\varepsilon_{h,j}\left(u^{\tau-1}_{h,\cdot}\right)\right\|_2^2\right]\le z^2. \tag{30}$$

Hence, based on (30), (29) is given by:

$$\mathrm{E}\left[\left\|u^{d,\tau}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2\right]\le\left(1-\zeta^{t-1}\delta\right)\mathrm{E}\left[\left\|u^{d,\tau-1}_{h,\cdot}-u^{d,*}_{h,\cdot}\right\|_2^2\right]+\left(\zeta^{t-1}\right)^2z^2. \tag{31}$$

$$\begin{aligned}&\forall f'_{h,j}\in O_t: &&\begin{cases}\nabla\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)-\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\\=-\left(f'_{h,j}-u^d_{\alpha,\cdot}v^d_{j,\cdot}\right)v^d_{j,\cdot}+\lambda u^d_{\alpha,\cdot}+\left(f'_{h,j}-u^d_{\beta,\cdot}v^d_{j,\cdot}\right)v^d_{j,\cdot}-\lambda u^d_{\beta,\cdot}\\=\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\left(v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\lambda E_l\right).\end{cases}\\&\forall f'_{h,j}\notin O_t,\text{and } f'_{h,j}\in K_t: &&\begin{cases}\nabla\varepsilon_{h,j}\left(u^d_{\alpha,\cdot}\right)-\nabla\varepsilon_{h,j}\left(u^d_{\beta,\cdot}\right)\\=-\eta\left(f'_{h,j}-u^d_{\alpha,\cdot}v^d_{j,\cdot}\right)v^d_{j,\cdot}+\lambda u^d_{\alpha,\cdot}+\eta\left(f'_{h,j}-u^d_{\beta,\cdot}v^d_{j,\cdot}\right)v^d_{j,\cdot}-\lambda u^d_{\beta,\cdot}\\=\left(u^d_{\alpha,\cdot}-u^d_{\beta,\cdot}\right)\left(\eta v^{d\ \mathrm{T}}_{j,\cdot}v^d_{j,\cdot}+\zeta\lambda E_l\right).\end{cases}\end{aligned} \tag{14}$$

Assume that the learning rate $\zeta^{t-1}=\sigma/(\delta t)$ with $\sigma>1$, (31) can be computed by:

$$\mathrm{E}\left[\left\|u_{h,\cdot}^{d,\tau}-u_{h,\cdot}^{d,*}\right\|_2^2\right]\leq\left(1-\frac{\sigma}{t}\right)\mathrm{E}\left[\left\|u_{h,\cdot}^{d,\tau-1}-u_{h,\cdot}^{d,*}\right\|_2^2\right]+\frac{1}{t^2}\left(\frac{\sigma z}{\delta}\right)^2. \tag{32}$$

By extending formula (32), a boundary is represented as follows:

$$\mathrm{E}\left[\left\|u_{h,\cdot}^{d,\tau}-u_{h,\cdot}^{d,*}\right\|_2^2\right]\leq\frac{1}{t}\max\left\{\left\|u_{h,\cdot}^{d,1}-u_{h,\cdot}^{d,*}\right\|_2^2,\frac{\sigma^2 z^2}{\delta\sigma-1}\right\}, \tag{33}$$

where $u_{h,\cdot}^{d,1}$ represents the initial state of $u_{m,\cdot}$ at the t-th iteration.

Based on *Lemma* 1, $\varepsilon_{m,j}$ is L-smooth and we can deduce that

$$\varepsilon_{h,j}\left(u_{h,\cdot}^{d,\tau}\right)-\varepsilon_{h,j}\left(u_{h,\cdot}^{d,*}\right)\leq\frac{L}{2}\left\|u_{h,\cdot}^{d,\tau}-u_{h,\cdot}^{d,*}\right\|_2^2. \tag{34}$$

By extending formula (34), we can deduce that

$$\mathrm{E}\left[\varepsilon_{h,j}\left(u_{h,\cdot}^{d,\tau}\right)-\varepsilon_{h,j}\left(u_{h,\cdot}^{d,*}\right)\right]\leq\frac{L}{2}\mathrm{E}\left[\left\|u_{h,\cdot}^{d,\tau}-u_{h,\cdot}^{d,*}\right\|_2^2\right]. \tag{35}$$

By combining (33) with (35), the following conclusion is deduced that

$$\mathrm{E}\left[\varepsilon_{h,j}\left(u_{h,\cdot}^{d,\tau}\right)-\varepsilon_{h,j}\left(u_{h,\cdot}^{d,*}\right)\right]\leq\frac{L}{2t}\Theta(\sigma), \tag{36}$$

where we have

$$\Theta(\sigma)=\max\left\{\left\|u_{h,\cdot}^{1}-u_{h,\cdot}^{d,*}\right\|_2^2,\frac{\sigma^2 z^2}{\delta\sigma-1}\right\}. \tag{37}$$

Then, expand formula (36) on all the known entries of O_t and K_t :

$$\mathrm{E}\left[\sum_{(h,j)\in O_t\cup K_t}\left(\varepsilon_{h,j}\left(u_{h,\cdot}^{d,\tau}\right)-\varepsilon_{h,j}\left(u_{h,\cdot}^{d,*}\right)\right)\right]\leq\left(|O_t|+|K_t|\right)\frac{L}{2t}\Theta(\sigma). \tag{38}$$

where $\rho\rightarrow\infty$, $\left(|O_t|+|K_t|\right)\frac{L}{2t}\Theta(\sigma)\rightarrow 0$.

When $u_{h,\cdot}$ is considered a constant, we will encounter the same situation despite the fact that the learning objective (7) is non-convex, $u_{h,\cdot}$ and $v_{j,\cdot}$ can be updated alternately by SGD. Furthermore, as concluded in [3], the learning rate of SGD is $\zeta^{t-1}\leqslant 1/\delta t$ at the t-th iteration. Therefore, following *Lemma* 2, the learning rate satisfies $\zeta^{t-1}\leqslant 1/\delta t$ in the t-th iteration, where the minimum singular value of the $(v_{j,\cdot}^{d\,\mathrm{T}}v_{j,\cdot}^{d}+\lambda E_l)$ is δ. Additionally, it is derived that regularization does not affect convergence. Thus, *Theorem* 1 holds.

V. References

[1] X. D. Zhang, "Matrix analysis and applications," *Cambridge University Press*, 2017.

[2] A. S. Nemirovski, A. Juditsky, G. Lan, and A. A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *Siam Journal on Optimization*, vol. 19, no. 4, pp. 1574-1609, Jan. 2009.

[3] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," *In Proceedings of the International Conference on Machine Learning*, 2012, pp. 1571-1578.