Estimate ordinal regression's parameters:

In this section, there are four algorithms, SGD, SGD with gamma decay, ASGD with gamma decay and Batch. These can be help us finding out the parameters. In order to choose the best one, will simulate some dataset and estimate its parameters by above method identically and calculate the error, the smaller one will be the best choice.

After generating data and estimate parameters, we use RMSE to compare performance and summarized the overall RMSE as Table.1.1 and the training time as Table.1.2.

Table.1.1 Overall RMSE with each method in different Epoch

Epoch	SGD	SGD with gamma decay	ASGD with gamma decay	Batch SGD	Batch SGD with gamma decay	Batch ASGD with gamma decay
1	0.536017	0.04172	0.046478	0.042584	0.23638	0.34591
2	0.247521	0.035421	0.024703	0.025134	0.218038	0.326721
5	0.444694	0.033234	0.017606	0.030741	0.080138	0.169311
10	0.265318	0.030453	0.013888	0.043584	0.057672	0.123812
20	0.306517	0.030858	0.011511	0.047609	0.01691	0.0698
30	0.404615	0.024136	0.010792	0.034093	0.010392	0.058852
50	0.458559	0.030856	0.010022	0.033026	0.0066	0.030975

Table.1.2 Training time with each method in different Epoch

Epoch	SGD	SGD with gamma decay	ASGD with gamma decay	Batch SGD	Batch SGD with gamma decay	Batch ASGD with gamma decay
1	9.333325	8.690249	8.690249	0.141679	0.154905	0.154905
2	19.6473	19.38494	19.38494	0.250332	0.268075	0.268075
5	50.16117	44.437	44.437	0.671879	0.664393	0.664393
10	98.45409	87.10141	87.10141	1.22089	1.373705	1.373705
20	196.0749	170.5268	170.5268	2.482988	2.75993	2.75993
30	284.3657	258.3534	258.3534	3.756515	4.242949	4.242949
50	452.7361	444.0151	444.0151	6.568452	6.713576	6.713576

Estimate regression parameters

Random variable $X_t \in R^N, Y_t \in R^1$ $and\ t = 1,2,...,$ then

$$Y_t = X_t^T\theta + \xi_t$$

$\theta \in R^N$ is unknown parameters and ξ is a random noise.

$$\theta_t = \theta_{t-1} + \gamma_t\varphi(Y_t - \theta_{t-1}^T X_t)X_t$$

$$\bar{\theta}_t = \frac{1}{t}\sum_{i=0}^{t}\theta_i$$

By compare SGD's gradient with MLS, we can find out that

$$\varphi(x) = x$$

Now, let N=5, and $X_t \sim N(5,1)\ \forall\ t$

$$B = EX_1X_1^T = \begin{pmatrix} 26 & 25 & 25 & 25 & 25 \\ 25 & 26 & 25 & 25 & 25 \\ 25 & 25 & 26 & 25 & 25 \\ 25 & 25 & 25 & 26 & 25 \\ 25 & 25 & 25 & 25 & 26 \end{pmatrix}$$

$$\psi(X) = E_\psi(X + \xi_1) = X$$

$$\chi(X) = E_{\psi^2}(X + \xi_1) = X^2 + 1$$

$\bar{\theta}_t \to \theta$ almost surely and $(\bar{\theta}_t - \theta)\sqrt{t} \overset{D}{\to} N(0, V)$ where

$$V = B^{-1}\frac{\chi(0)}{\psi'^2(0)} = B^{-1}$$

Now, we set t = 1000, 10000, 100000, 1000000 and 5000000. And simulated 100 $\bar{\theta}_t$ and transform them by $(\bar{\theta}_t - \theta)\sqrt{t}$. Then we can calculate the covariance matrix from these data to compare with V to verify the theorem.

In this part, we choose RMSE and MAPE/100 with index to help us displaying the difference between simulation and theoretical value. The result was summarized as Table2.

Table.2 simulation data and theoretical value compare

	t	SGD	ASGD	SGD Gamma Decay with $\alpha = 0.6$	ASGD Gamma Decay with $\alpha = 0.6$	MLE
RMSE	1000	6.028903	5.225439	105.8464	171.3997	0.0646
	10000	68.45575	1.167064	159.8135	604.8938	0.0796
	100000	713.9468	0.8764387	2.846579	180.8449	0.0773
	1000000	6707.197	0.7993826	0.9150568	19.26127	
	5000000	36515.53	0.9089417	2.009199	3.970454	
MAPE	1000	7.167466	12.79372	255.7983	414.6533	0.0674
	10000	93.38666	2.860472	391.3354	1486.948	0.1179
	100000	1070.111	2.158123	6.239338	441.373	0.0310
	1000000	7901.849	1.975134	1.471857	46.55031	
	5000000	42894.66	2.22445	2.594442	9.77927	

In past simulation, we found out that gamma decay would be affect the estimation a lot. The gamma decay's equation $\gamma_t = \gamma_1 t^{-\alpha}$ means that learning rate will be decayed with t increase. We try some different γ_1 & α with t = 1, …, 100000, and simulated 100 $\overline{\theta}_t$ and transform them by $(\overline{\theta}_t - \theta)\sqrt{t}$. Then calculate its covariance matrix and compute RMSE with theoretical value V. The result is as Table.3 showed. We found each learning rate and alpha have a converge point. With the distance of the converge point be larger, the bias would be larger too.

Table.3 Different α and γ comparison

		Learning rate			
		0.05	0.02	0.01	0.005
Alpha	0	NA	NA	0.892287	0.341187
	0.1	NA	0.945486	0.217326	0.67272
	0.2	8.46E+87	0.304418	0.348152	2.023657
	0.3	5.43E+18	0.320356	1.474426	8.051103
	0.4	32845543	1.125227	8.033412	81.99313
	0.5	33925.53	10.08667	106.9581	1195.616
	0.6	8081.722	147.7876	1525.642	7712.765
	0.7	20740.67	2533.464	8564.414	19849.92
	0.8	73522.4	11862.09	19848.76	27141.01
	0.9	258502.9	20082.29	29008.73	34089.86
	1	17164.9	32768.95	43554.11	39853.12

This part, try some different γ_1 & α with t = 1, …, 10000, and simulated 100 $\overline{\theta}_t$. Then calculate the RMSE & MAPE as Table.4 and Table.5.

Table.4 Different α and γ RMSE

		Learning rate		
		0.02	0.01	0.005
Alpha	0	NaN	0.007	0.0075
	0.1	0.0191	0.0057	0.0109
	0.2	0.0073	0.0067	0.0181
	0.3	0.0085	0.011	0.0401
	0.4	0.0152	0.0271	0.1213
	0.5	0.043	0.0951	0.368
	0.6	0.1631	0.302	0.7452
	0.7	0.477	0.5312	1.0274
	0.8	0.8503	0.6863	1.2084
	0.9	1.1077	0.8069	1.2854
	1	1.2076	0.9363	1.2957

Table.5 Different α and γ MAPE

		Learning rate		
		0.02	0.01	0.005
Alpha	0	NaN	0.0376	0.007
	0.1	0.0295	0.0307	0.0103
	0.2	0.0159	0.0357	0.0166
	0.3	0.0192	0.0609	0.037
	0.4	0.0329	0.154	0.1118
	0.5	0.0927	0.5047	0.3449
	0.6	0.3109	1.6721	0.7062
	0.7	1.0868	2.9954	0.9645
	0.8	1.7623	3.7728	1.1085
	0.9	2.3422	4.6266	1.2355
	1	2.5994	5.1107	1.2376