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 $\mathrm{X_t} \in R^N, Y_t \in R^1 \; and \; t = 1,2, \ldots$, then

$$\mathrm{Y_t} = X_t^T \theta + \xi_\mathrm{t}$$

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

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

$$\overline{\theta_\mathrm{t}} = \frac{1}{t} \sum_{\mathrm{i=0}}^{\mathrm{t}} \theta_i$$

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

$$\varphi(\mathrm{x}) = x$$

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

$$\mathrm{B} = \mathrm{E} X_1 X_1^T = \begin{pmatrix} 26 & 25 & 25 & 25 & 25 \\ 25 & 26 & 25 & 25 & 25 \\ 25 & 25 & 26 & 25 & 25 \\ 25 & 25 & 25 & 26 & 25 \\ 25 & 25 & 25 & 25 & 26 \end{pmatrix}$$

$$\psi(\mathrm{X}) = E_\psi(X + \xi_1) = X$$

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

$\overline{\theta_\mathrm{t}} \rightarrow \theta$ almost surely and $(\overline{\theta_\mathrm{t}} - \theta)\sqrt{t} \overset{D}{\rightarrow} \mathrm{N}(0, \mathrm{V})$ where

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

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_GammaDecay	ASGD_GammaDecay
RMSE	1000	186.7961	559.5249	542.9791	778.2158
	10000	5.96061	155.4226	1003.565	2626.76
	100000	65.50717	19.69578	149.1597	2619.446
	1000000	696.5746	1.482268	2.454555	315.4811
	5000000	3817.504	0.651	4.635725	69.50927
MAPE	1000	117.0392	351.7522	341.2168	489.2872
	10000	0.577989	97.71905	630.7652	1651.574
	100000	16.37947	12.38328	93.50557	1646.982
	1000000	195.1138	0.930485	0.550106	198.3595
	5000000	499.0092	0.3471	0.463195	43.70374