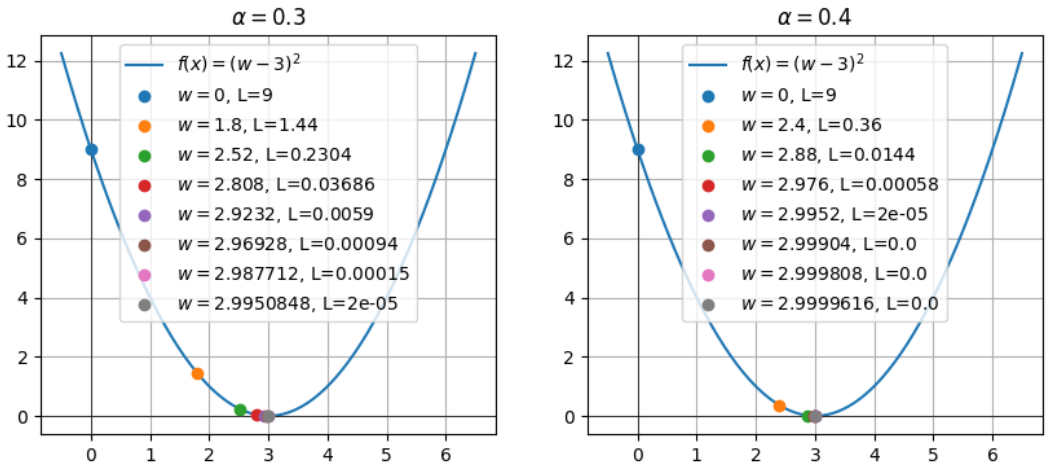
Ուսուցման արագություն: Պարամետրեր և հիպերպարամետրեր

Հայկ Կարապետյան

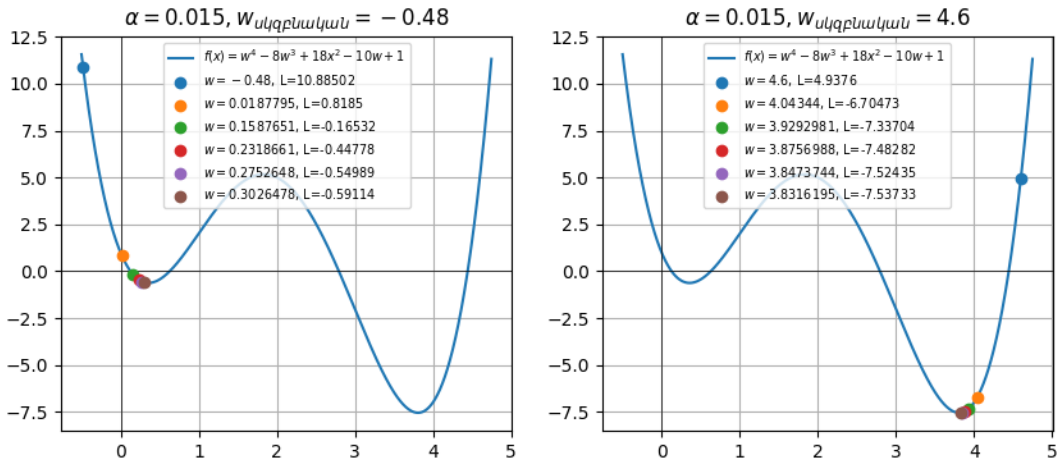
1 Ուսուցման արագություն

Ուսուցման արագությունը (learning rate, α) ցույց է տալիս, թե ինչքան արագ ֆունկցիան պետք է զուգամիտի մինիմումի կետին: Ուսուցման արագությանը մեծ արժեքը տալը ունի ինչպես լավ կողմեր, այնպես էլ վատ կողմեր: Լավ կողմերից է մինիմումի կետին արագ զուգամիտելու փաստը: Այսինքն ուսուցանվող պարամետրերի արժեքները ամեն քայլից հետո ավելի շատ կփոփոխվեն, քան ուսուցման փոքր արագության դեպքում (Նկար 1):



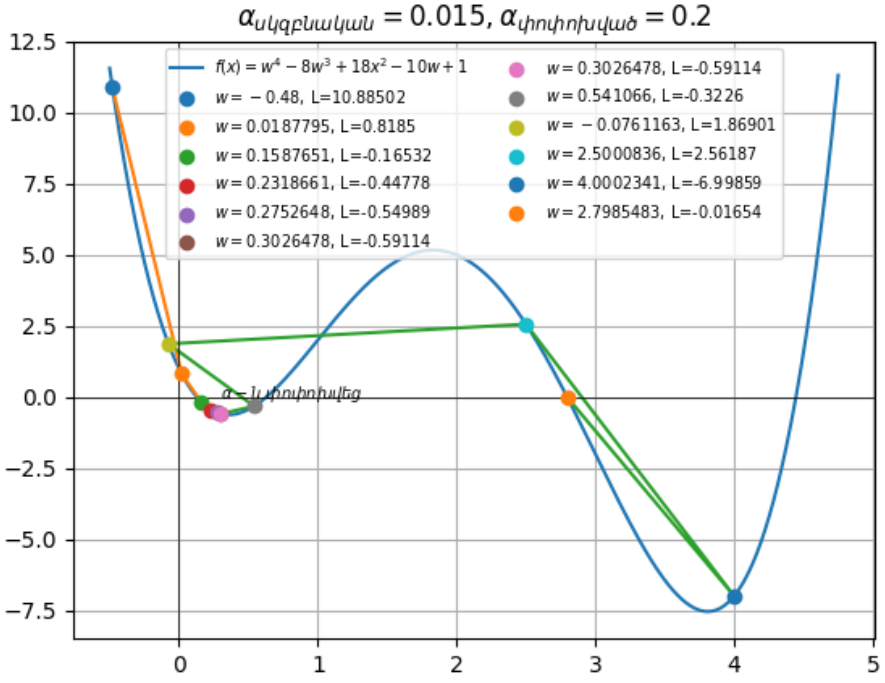
Նկար 1: Աջ կողմում կորուստը ավելի արագ է զուգամիտում մինիմում արժեք (w-ի փոփոխությունը ավելի կտրուկ է)

Երկրորդ լավ կողմը լոկալ մինիմումի կետից դուրս գալն է: Այսինքն ունենք կորստի ֆունկցիա, որը ունի մեկ լոկալ մինիմում և մեկ գլոբալ մինիմում: w-ի սկզբնարժեքավորումից կախված կատարելով գրադիենտային վայրեջք կամ կհայտնվենք գլոբալ մինիմումում կամ լոկալ (Նկար 2):



Նկար 2: w-ի սկզբնական արժեքից կախված, գրադիենտային վայրեջքի ալգորիթմը կարող է գտնել տարբեր մինիմումներ

Իսկ ինչ անել, երբ մեզ բավարար չէ ստացված մինիմում արժեքը (օրինակ՝ կորստի արժեքը հասել է -0.59 բայց մոդելի արդյունքը (ճշգրտություն) մեզ հերիք չէ): Կարող ենք ուսուցման արագությունը մեծացնենք, իսկ հետո անհրաժեշտության դեպքում փոքրացնենք (Նկար 3):



Նկար 3: Նարնջագույն գույնով պատկերված է մինչ α-ն փոփոխելը w-ի ընդունած արժեքների հերթականությունը, իսկ կանաչը ցույց է տալիս փոխելուց հետո ընդունած արժեքները:

Նկար 3-ից պարզ է դառնում, որ լոկալ մինիմումում գտնվելու դեպքում կարելի է մեծացնել α-ն և դուրս գալ դեպի գլոբալ մինիմում: Այս դեպքում անհրաժեշտություն է առաջանում փոքրացնել α-ն մեծացնելուց հետո, քանի որ 4 արժեք ընդունելուց հետո մեծ ուսուցման արագության պատճառով այն ընդունում է 2.8 արժեք, իսկ մեր մինիմումի կետը մոտավորապես 3.8-ն է: Սա էլ ուսուցման մեծ արագություն ընտրելու վատ կողմն է: Մեծ ուսուցման արագության ժամանակ հնարավոր է չհասնենք մինիմումի: Օրինակ՝ վերցնենք $L(w) = w^2$ ֆունկցիան և $w_{\text{սկզբ.}} = 4$: Կատարենք գրադիենտային վայրէջք երբ $\alpha = 1$:

$$\frac{\partial L(w)}{\partial w} = 2w$$

$$w_{step_1} = w_{\text{սկզբ.}} - \alpha\frac{\partial L(w)}{\partial w} = 4 - 1 \times 2 \times 4 = -4$$

$$w_{step_2} = w_{step_1} - \alpha\frac{\partial L(w)}{\partial w} = -4 - 1 \times 2 \times -4 = 4$$

$$w_{step_3} = 4$$

$\alpha = 1$ արժեքի դեպքում մենք երբեք չենք հասնի մինիմումի կետին: Նույնը կլինի, եթե α-ն լինի ավելի մեծ քան մեկը ($\alpha = 2$):

$$w_{step_1} = w_{\text{սկզբ.}} - \alpha\frac{\partial L(w)}{\partial w} = 4 - 2 \times 2 \times 4 = -12$$

$$w_{step_2} = w_{step_1} - \alpha\frac{\partial L(w)}{\partial w} = -12 - 2 \times 2 \times -12 = 36$$

$$w_{step_3} = -108, L(w_{step_3}) = (-108)^2 \text{ (Նկար 4)}$$

Նկար 4: α-ի մեծ արժեքի դեպքում կորստի ֆունկցիան կարող է կտրուկ մեծանալ:

Այդ պատճառով α-ի արժեքը ընտրելիս պետք է զգույշ լինել: Շատ փոքր լինելը կարող է հանգեցնել դանդաղ ուսուցման, շատ մեծ ընտրելը կարող է հանգեցնել մինիմումի կետ չհասնելուն, իսկ ժամանակ առ ժամանակ մեծացնելը և փոքրացնելը կօգնել դուրս գալ մինիմումի կետերից:

2 Պարամետրեր և հիպերպարամետրեր

Ներմուծենք երկու տերմին պարամետր (parameter) և հիպերպարամետր (hyperparameter): Պարամետրերը ցանցերում այն արժեքներն են, որոնք մոդելը փոփոխում է ուսուցման ընթացքում: Օրինակ՝ կշիռները, բիասները և տարբեր պարամետրիկ արժեքները (w, b, a[1]): Հիպերպարամետրերը այն արժեքներն են, որոնք ընտրվում են օգտագործելով կարգավորման տվյալները (validation data): Օրինակ՝ ակտիվացիոն ֆունկցիաների, ուսուցման արագության, շերտերի և նեյրոնների քանակի ընտրությունը կատարվում է կարգավորման տվյալների միջոցով:
Պարամետրերի փոփոխումը արդեն պարզ է, որ կատարվում է գրադիենտային վայրեջքի միջոցով: Իսկ ինչպե՞ս են փոփոխվում հիպերպարամետրերը:

1. Սկզբից կառուցում ենք նեյրոնային ցանցը ընտրելով հայտնի հիպերպարամետրեր (հայտնի ակտիվացիոն ֆունկցիաներ, հայտնի շերտերի և նեյրոնների քանակ և դասավորություն, ուսուցման արագություն):
2. Ուսուցանում ենք մոդելը:
3. Փորձարկում ենք մոդելը կարգավորման տվյալների վրա և ստանում ենք ինչ որ ճշգրտություն:
4. Փոփոխում ենք հիպերպարամետրերը և կատարում 2, 3 քայլերը N անգամ:
5. Փորձարկած հիպերպարամետրերից ընտրում ենք ամենամեծ ճշգրտություն ցուցաբերածը և մոդելը կիրառում ենք փորձարկման տվյալների (test data) վրա: Ստացված ճշգրտությունը ցույց կտա ինչքանով է մոդելը լավ աշխատում չտեսած տվյալների վրա:

3-րդ քայլը կարող է կատարվել նաև ուսուցման ընթացքում, ոչ թե ուսուցման վերջում: Օրինակ՝ ուսուցումը բաղկացած է 10000 գրադիենտային վայրեջքից: Փորձարկման վրա կարող ենք ստուգել մոդելի ճշգրտությունը ամեն 100 գրադիենտային վայրեջք կատարելուց հետո (լավ արդյունք չցուցաբերելու դեպքում կանգնեցնենք ուսուցումը): Գերուսուցումը և թերուսուցումը նույնպես հնարավոր է հասկանալ կարգավորման տվյալների միջոցով: Օրինակ՝ ուսուցման տվյալների վրա ճշգրտությունը 90% է, իսկ կրգավորման տվյալների վրա 60%, նշանակում է մոդելը գերուսուցված է:

1. Parametric ReLU: $PR(x) = \begin{cases} ax, \text{երբ } x < 0 \\ x, \text{երբ } x \geq 0 \end{cases}$