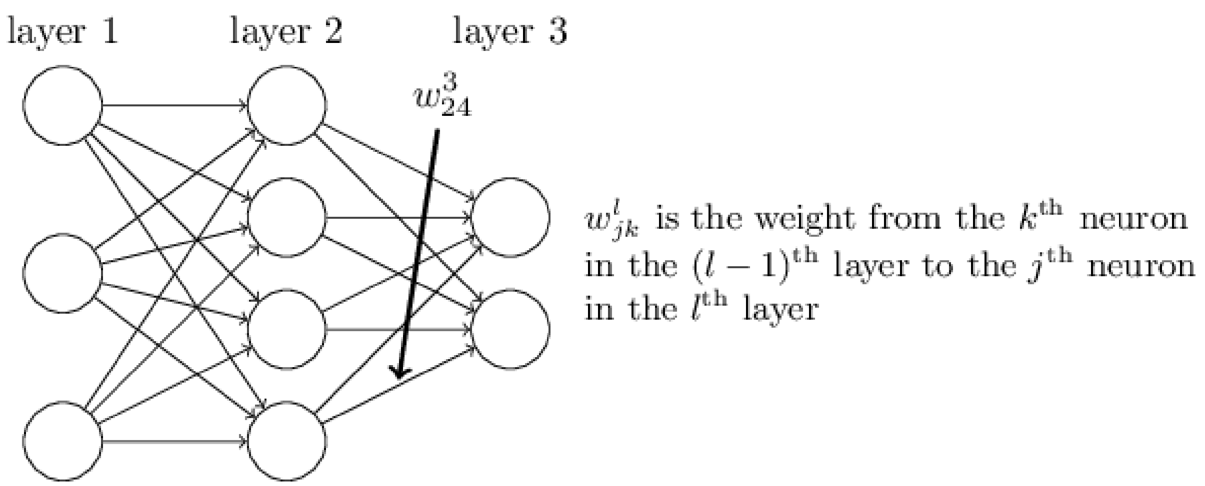
Հետադարձ տարածման

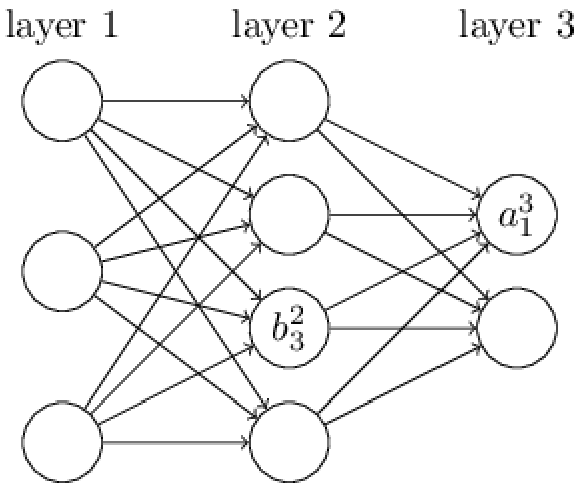
Հետադարձ տարածման հիմքում ընկած է C գնի ֆունկցիայի՝ նեյրոնային ցանցերի w կշիռների (կամ b շեղումների) նկատմամբ մասնակի ածանցյալի ∂C/∂w արտահայտությունը: Այն ցույց է տալիս, թե ինչ «արագությամբ» է գինը փոփոխվում՝ կախված կշիռների և շեղումների փոփոխությունից: Այս արտահայտությունը փոքր-ինչ բարդ տեսք ունի, սակայն այն միաժամանակ շատ գեղեցիկ է, որի յուրաքանչյուր մասնիկն ունի բնական, ինտուիտիվ բացատրություն: Եվ, այսպիսով, հետադարձ տարածումը միայն արագագործ ալգորիթմ չէ, որը «ստիպված» ենք սովորել: Իրականում այն տալիս է խորը ներըմբռնում (insight) այն մասին, թե ինչպես է կշիռների և շեղումների փոփոխությունն ազդում ցանցի վարքագծի վրա: Այդ իսկ պատճառով այս ալգորիթմն արժանի է մանրամասն ուսուցման:

Մարզանք. նեյրոնային ցանցի ելքային արժեքների հաշվման արագագործ, մատրիցային մոտեցում

Նախ ցանցի կշիռներին տանք այնպիսի նշանակումներ, որոնք կօգտագործենք այդ կշիռներին հստակորեն հղվելու նպատակով: Նշանակենք w_{jk}^{l}-ով ցանցի (l–1)-րդ շերտի k-րդ նեյրոնը l-րդ շերտի j-րդ նեյրոնին միացնող կապի կշիռը: Այսպիսով, օրինակ, ներքևի դիագրամում պատկերված կշիռը երկրորդ շերտի չորրորդ նեյրոնը միացնում է երրորդ շերտի երկրորդ նեյրոնի հետ.



Նմանատիպ նշանակումներ են օգտագործվում նաև ցանցի շեղումների և ակտիվացիաների համար։ Եվ այսպես, l-րդ շերտի j-րդ նեյրոնին համապատասխանող շեղումը կնշանակենք b_j^l-ով։ Ստորև ներկայացված դիագրամը ցույց է տալիս, թե ինչպես օգտագործել այս նշանակումները.



Ըստ այս նշանակումների, lրդ շերտի jրդ նեյրոնի a_j^l ակտիվացիան (l–1)-րդ շերտի նեյրոնների ակտիվացիաների հետ կապված է հետևյալ հավասարմամբ (1)

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

որտեղ գումարն ըստ (l–1)-րդ շերտի k նեյրոնների է։ Սահմանենք w^l *կշիռների մատրիցը (weight matrix)*: w^l կշիռների մատրիցը կազմված է l-րդ շերտին կապվող նեյրոնների կշիռներից, այսինքն, j-րդ տողի k-րդ սյունակի էլեմենտը w_{jk}^l կշիռն է։ Նույն ձևով, յուրաքանչյուր l շերտի համար սահմանենք b^l *շեղման վեկտոր (bias vector)*: Ակնհայտ է, որ շեղման վեկտորի անդամները պարզապես b_j^l արժեքներն են՝ l-րդ շերտի նեյրոնների շեղումները։ Եվ վերջապես, սահմանենք a^l ակտիվացիայի վեկտորը, որի անդամներն a_j^l ակտիվացիաներն են։

Հավասարումն գրենք մատրիցային տեսքով։ Մեր նպատակն է, որպեսի σ-ն կարողանանք կիրառել v վեկտորի բոլոր էլեմենտների վրա։ Ֆունկցիայի էլեմենտ-առ-էլեմենտ կիրառումը նշանակենք σ(v) արտահայտությամբ։ σ(v)-ի տարրերը պարզապես $\sigma(v)_j=\sigma(v_j)$ էլեմենտներն են։ Որպես օրինակ, դիտարկենք այն դեպքը, երբ ունենք $f(x)=x^2$ ֆունկցիան, ապա ֆունկցիայի վեկտորացված տեսքը կլինի՝

$$f\left(\begin{bmatrix}2\\3\end{bmatrix}\right)=\begin{bmatrix}f(2)\\f(3)\end{bmatrix}=\begin{bmatrix}4\\9\end{bmatrix}$$

այսինքն վեկտորացված f-ն ուղղակի վեկտորի յուաքանչյուր անդամը քառակուսի է բարձրացնում:

Հաշվի առնելով այս բոլոր նշանակումները, (1) հավասարումը կարելի է գրել այսպիսի կոմպակտ վեկտորացված տեսքով. (2)

$$a^l=\sigma(w^la^{l-1}+b^l)$$

Այս արտահայտությունը ընդհանուր առմամբ հնարավորություն է տալիս դիտարկել նախորդ շերտի ակտիվացիաների ազդեցությունը հաջորդ շերտի ակտիվացիաների վրա. ուղղակի կիրառում ենք կշիռների մատրիցը ակտիվացիաների վրա, այնուհետև ավելացնում ենք շեղման վեկտորը և վերջապես կիրառում σ ֆունկցիան :

Երկու ենթադրություններ գնային ֆունկցիայի վերաբերյալ

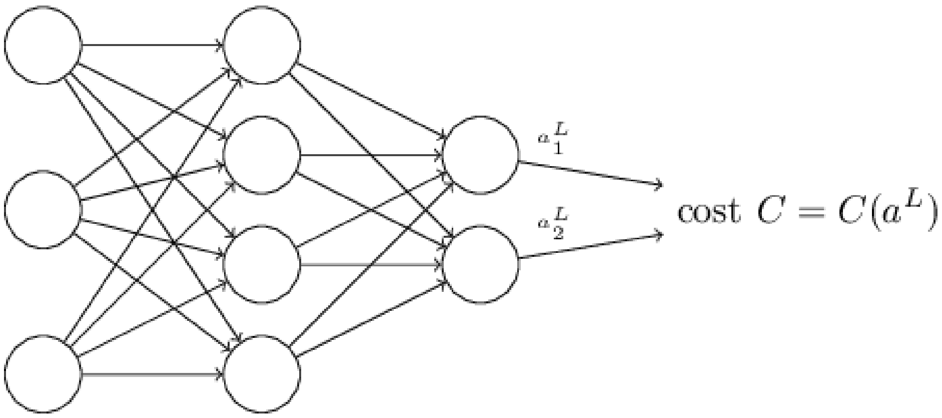
Հետադարձ տարածման նպատակն է հաշվել C գնային ֆունկցիայի ∂C/∂w և ∂C/∂b մասնական ածանցյալները w կշիռների և b շեղումների նկատմամբ: Կատարենք երկու ենթադրություններ հետադարձ տարածման համար: Սակայն, մինչ այդ դիտարկենք քառակուսային գնային ֆունկցիան, որը ունի հետևյալ տեսքը՝ (3)

$$C=\frac{1}{2n}\sum_x\|y(x)-a^L(x)\|^2$$

որտեղ n-ը մարզման օրինակների քանակն է, իսկ գումարն ըստ անհատական x մարզման օրինակների է, y=y(x) համապատասխան ցանկալի ելքային արժեքն է և $a^L=a^L(x)$ ցանցի ելքային վեկտորն է x մուտքային վեկտորի դեպքում:

Այսպիսով, ի՞նչ ենթադրություններ կարող ենք կատարել C գնային ֆունկցիայի վերաբերյալ, որպեսզի հետադարձ տարածումը հնարավոր լինի կիրառել: Առաջին ենթադրությունն այն է, որ գնային ֆունկցիան կարելի է արտահայտել $C = \frac{1}{n}\sum_x C_x$ հավասարմամբ որպես x մարզման օրինակներից կախված առանձին C_x գնային ֆունկցիաների հանրահաշվական միջին: Այս պնդումը ճիշտ է քառակուսային գնային ֆունկցիայի դեպքում, որտեղ $C_x=\frac{1}{2}\|y-a^L\|^2$ գնային ֆունկցիան է՝ կախված մեկ մարզման օրինակից: Այս ենթադրությունը ճիշտ կլինի նաև մնացած այլ գնային ֆունկցիաների դեպքում:

Գնային ֆունկցիայի մասին երկրորդ ենթադրությունը այն է, որ այն կարելի է արտահայտել որպես ֆունկցիա կախված նեյրոնային ցանցերի ելքային արժեքներից.



Օրինակ, քառակուսային գնային ֆունկցիան բավարարում է այս պնդմանը, քանի որ քառակուսային գնային ֆունկցիան տրված x մուտքային վեկտորի համար կարելի է արտահայտել որպես (4)

$$C = \frac{1}{2}\|y - a^L\|^2 = \frac{1}{2}\sum_j (y_j - a_j^L)^2$$

որը ֆունկցիա է՝ կախված ելքային ակտիվացիաներից: C-ն կարող ենք դիտարկել որպես a^L ելքային վեկտորներից կախված ֆունկցիա, որտեղ y պարզապես պարամետր է, որը մասնակցում է ֆունկցիայի սահմանմանը:

Հադամարի արտադրյալը՝ s⊙t

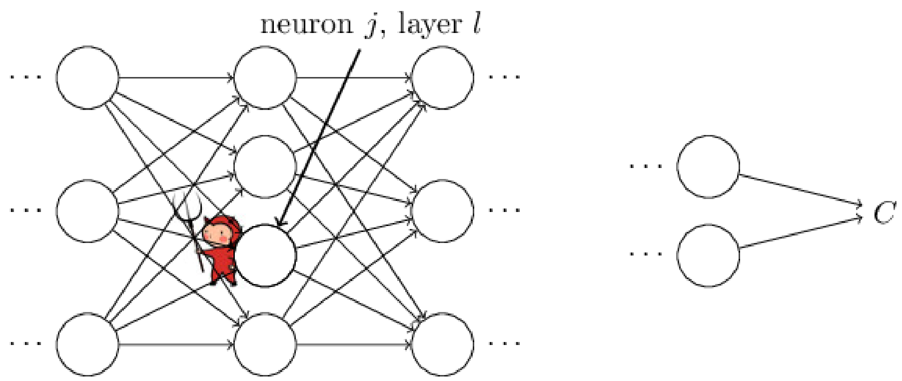
Հետադարձ տարածման ալգորիթմը հիմնված է որոշ գծային հանրահաշվի գործողությունների վրա՝ վեկտորների գումարում, վեկտորի բազմապատկում մատրիցով և այլն: Սակայն գործողություններից մեկն ավելի քիչ հաճախ է օգտագործվում քան մնացածները: Ենթադրենք, որ s և t միևնույն չափողականությամբ վեկտորներ են: Վեկտորների *էլեմենտ առ էլեմենտ* արտադրյալը նշանակենք s⊙t արտահայտությամբ: Հետևաբար s⊙t արտադրյալի տարրերը կլինեն $(s\odot t)_j = s_j t_j$ արժեքները: Դիտարկենք հետևյալ օրինակը.

$$\begin{bmatrix}1\\2\end{bmatrix} \odot \begin{bmatrix}3\\4\end{bmatrix} = \begin{bmatrix}1*3\\2*4\end{bmatrix} = \begin{bmatrix}3\\8\end{bmatrix}$$

Այս էլեմենտ առ էլեմենտ արտադրյալը այլ կերպ անվանում են *Հադամարի արտադրյալ* կամ *Շուրի արտադրյալ*:

Հետադարձ տարածման հիմքում ընկած չորս հիմնական հավասարումները

Հետադարձ տարածումն այն մասին է, թե ինչպես է ցանցի գնային ֆունկցիան փոփոխվում՝ կախված կշիռների և շեղումների փոփոխություններից: Սա իր հերթին նշանակում է, որ պետք է հաշվել $\partial C/\partial w^l_{jk}$ և $\partial C/\partial b^l_j$ մասնակի ածանցյալները: Նախ ներմուծենք δ^l_j միջանկյալ մեծությունը, որը կկոչենք l-րդ շերտի j-րդ նեյրոնի *սխալանք (error)*:



l-րդ շերտի j-րդ նեյրոնի δ_j^l սխալանքը սահմանենք որպես (5)

$$\delta_j^l \equiv \frac{\partial C}{\partial z_j^l}.$$

Ինչպես մնացած դեպքերում, δ^l -ով կնշանակենք l-րդ շերտի սխալանքների վեկտորը: Այնպիսի դասակարգման խնդիրներում, ինչպիսին է MNIST-ը, «սխալանք» տերմինը սովորաբար օգտագործվում է դասակարգման ձախողման գործակցի իմաստով: Օրինակ, երբ նեյրոնային ցանցը թվանշանների 96.0 տոկոսը ճիշտ է դասակարգում, ապա ձախողման գործակիցը 4.0 տոկոս է:

Հարձակման պլանը: Հետադարձ տարածումը հիմնված է 4 ֆունդամենտալ հավասարումների վրա: Այդ հավասարումները միասին մեզ հնարավորություն են տալիս հաշվել δ^l սխալանքն ու գնային ֆունկցիայի գրադիենտը:

Ելքային շերտի սխալանքի δ^L հավասարումը: δ^L հավասարման էլեմենտներն ունեն հետևյալ տեսքը՝ (BP1)

$$\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L).$$

Դիտարկենք այդ արտահայտությունը: Հավասարման աջ կողմի առաջին հատվածը՝ $\partial C/\partial a_j^l$ -ը պարզապես ցույց է տալիս, թե ինչ արագությամբ է գնային ֆունկցիան փոփոխվում՝ կախված j-րդ ելքային ակտիվացիայից: Աջակողմյան երկրորդ արտահայտությունը ցույց է տալիս σ ակտիվացիայի ֆունկցիայի փոփոխման արագությունը՝ կախված z_j^L -ից:

(BP1) հավասարումը սահմանում է δL վեկտորի անդամները: Հավասարումը չունի վեկտորական տեսք, ինչն անհրաժեշտ է հետադարձ տարածման համար, այնուամենայնիվ այն ճշգրիտ նկարագրում է ելքային շերտի սխալանքը: Այսպիսով, δL արտահայտությունը կարելի է գրել մատրիցային տեսքով հետևյալ կերպ՝ (BP1a)

$$\delta^L = \nabla_a C \odot \sigma'(z^L).$$

Որտեղ $\nabla_a C$ սահմանվում է որպես վեկտոր, որի անդամներն են $\partial C/\partial a_j^l$ մասնական ածանցյալները: $\nabla_a C$ -ն կարելի է ընկալել որպես C գնային ֆունկիայի փոփոխման գործակիցը՝ կախված ելքային ակտիվացիաներից: Հեշտ է նկատել, որ (BP1a) և (BP1) հավասարումները համարժեք են:

δ^l սխալանքի արտահայտումն ըստ հաջորդ շերտի δ^{l+1} սխալանքի: Դիտարկենք հետևյալ հավասարումը. (BP2)

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l),$$

որտեղ $(w^{l+1})^T$ կշիռների w^{l+1} տրանսպոնացված մատրիցն է՝ (l+1)-րդ շերտի համար :

Միավորելով (BP2) և (BP1) հավասարումները, կարող ենք δ^l սխալանքը հաշվել ցանցի կամայական շերտում: Սկսենք δ^L -ի հաշվումից՝ օգտագործելով (BP1) հավասարումը, այնուհետև կօգտագործենք (BP2) հավասարումը, որպեսզի հաշվենք δ^{L-1} սխալանքը, այնուհետև կկրկնենք (BP2) հավասարման կիրառումը, որպեսզի հաշվենք δ^{L-1} և այդպես շարունակ մինչև ցանցի «սկիզբը»

Գնային ֆունկցիայի փոփոխման գործակցի հավասարումը՝ կախված ցանցի շեղումներից: Դիտարկենք հետևյալ հավասարումը. (BP3)

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l.$$

Այն է, δ_j^l սխալանքը *ճիշտ նույնն* է, ինչ $\partial C/\partial \delta_j^l$ փոփոխման գործակիցը: Քանի որ (BP1) և (BP2) արդեն ցույց են տվել, թե ինչպես կարելի է հաշվել δ_j^l սխալանքը: Կարող ենք (BP3) հավասարումն գրել կարճ որպես

$$\frac{\partial C}{\partial b} = \delta,$$

որտեղ δ հաշվարկում ենք նույն նեյրոնի համար, ինչի համար դիտարկում էինք b շեղումը:

Գնային ֆունկցիայի փոփոխման գործակցի հավասարումը՝ կախված ցանցի կշիռներից: Դիտարկենք հետևյալ հավասարումը՝ (BP4)

$$\frac{\partial C}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j.$$

Այն ցույց է տալիս, թե ինչպես կարելի է հաշվել $\partial C/\partial w^l_{jk}$ մասնական ածանցյալներն ըստ δ^l և a^{l-1} մեծությունների: Հավասարումը կարելի է արտահայտել ավելի կարճ տեսքով հետևյալ կերպ.

$$\frac{\partial C}{\partial w} = a_{\rm in} \delta_{\rm out},$$

որտեղ a_{in} մեծությունը w կշիռների հետ միասին մուտք հանդիսացող նեյրոնի ակտիվացիան է և δ_{out}-ն նեյրոնի արդյունքի ելքային սխալանքն է w կշիռների դեպքում: Երբ ain փոքր է, $a_{in}\approx0$, ապա ∂C/∂w գրադիենտի արժեքը նույնպես կձգտի փոքր արժեքների: Այդ դեպքում կասենք, որ կշիռը *դանդաղ է սովորում*, ինչը նշանակում է, որ այն շատ չի փոփոխվում գրադիենտային վայրէջքի ժամանակ: Այլ կերպ ասած (BP4) հավասարման հետևանքներից մեկն այն է, որ թույլ ակտիվացիայով (low-activation) նեյրոններից դուրս եկող կշիռները դանդաղ են սովորում:

(BP1) - (BP4) . հավասարումներից կարելի է անել նաև այլ հետևություններ: Դիտարկենք $\sigma'(z^L_j)$ արտահայտությունը (BP1) . հավասարման մեջ: Վերջին շերտի կշիռները դանդաղ կսովորեն, եթե ելքային նեյրոնն ունի թույլ ակտիվացիա (≈0) կամ ուժեղ ակտիվացիա (≈1): Այս դեպքում ասում են, որ ելքային նեյրոնը *հագեցած (saturated)* է և արդյունքում կշիռը դադարել է սովորել (կամ սովորում է շատ դանդաղ): Նմանատիպ պնդումներ կարելի է անել նաև ելքային նեյրոնի շեղումների մասին:

Նմանատիպ հետևություններ կարող ենք կատարել նաև միջանկյալ շերտերի դեպքում: Դիտարկենք $\sigma'(z^l)$ արտահայտությունը (BP2) . հավասարման մեջ: Պարզ է, որ δ^l_j ավելի հավանական է, որ փոքր լինի, եթե նեյրոնը մոտ է հագեցմանը: Եվ սա, իր հերթին նշանակում է, որ յուրաքանչյուր կշիռների մուտք հագեցած նեյրոնին կսովորի դանդաղորեն:

Այսպիսով, մենք սովորեցինք, որ կշիռը կսովորի դանդաղորեն, եթե մուտքային նեյրոնը թույլ ակտիվացիայով է կամ երբ ելքային նեյրոնը հագեցած է, այն է ունի ուժեղ կամ թույլ ակտիվացիա:

Հետադարձ տարածման ալգորիթմը

Հետադարձ տարածման հավասարումները ցույց են տալիս, թե ինչպես կարելի է հաշվել գնային ֆունկցիայի գրադիենտը: Արտահայտենք դա ալգորիթմի տեսքով.

1. **Մուտք** x: a^1 ակտիվացիայի շերտը սկզբնավորենք մուտքային շերտով:
2. **Առաջաբերում (Feedforward):** Յուրաքանչյուր l=2,3,…,L շերտերի համար հաշվել z^l=$w^l a^{l-1} + b^l$ and a^l=σ(z^l).
3. **Ելքային սխալանք** δ^L**:** Հաշվել δ^L=$\nabla_a C \odot \sigma'(z^l)$ վեկտորը:
4. **Սխալանքի հետադարձ տարածումը (Backpropagate):** Յուրաքանչյուր l=L−1, L−2,…,2 շերտի համար հաշվել δ^l=$((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$.
5. **Ելք:** Գնային ֆունկցիայի գրադիենտը տրվում է $\frac{\partial C}{\partial w^l_{jk}}$= $a^{l-1}_k \delta^l_j$ և $\frac{\partial C}{\partial b^l_j}$=$\delta^l_j$ հավասարումներով:

Դիտարկելով ալգորիթմը, կարելի է տեսնել, թե ինչու է այն կոչվում *հետադարձ* տարածում: Սխալանքի վեկտորները հետադարձ հաշվարկվում են՝ սկսելով վերջին շերտից: Հետադարձ գործողությունը հետևանք է այն բանի, որ գնային ֆունկցիան կախված է ցանցի ելքային արժեքներից: Որպեսզի հասկանանք գնի փոփոխությունը կախված կշիռներից և շեղումներից, պետք է կրկնողաբար կիրառենք բարդ ֆունկցիայի դիֆերենցման կանոնը, հետ վերադառնալով շերտ առ շերտ, որպեսզի ստանանք համապատասխան արտահայտությունները:

Հետադարձ տարածման ալգորիթմը հաշվում է գնային ֆունկցիայի գրադիենտը տրված C=C_x մարզման օրինակի համար: Հաճախ հետադարձ տարածումը միավորվում է այնպիսի ուսուցման ալգորիթմի հետ, ինչպիսին է ստոկաստիկ գրադիենտային վայրէջքը, որտեղ գրադիենտը հաշվարկվում է բազմաթիվ օրինակների հիման վրա։ Տրված m մարզման օրինակների մինի-փաթեթի համար, հետևյալ ալգորիթմը կիրառում է գրադիենտային վայրէջքի ուսուցման քայլը՝ հիմնված այդ մինի-փաթեթի վրա։

1. **Մուտքագրեք մարզման օրինակների բազմությունը։**
2. **Յուրաքանչյուր** x **մարզման օրինակի համար.** Սկզբնավորենք համապատասխան $a^{x,1}$ մուտքային ակտիվացիան և կատարենք հետևյալ քայլերը.
   - **Առաջ բերում (Feedforward):** Յուրաքանչյուր l=2,3,…,L շերտի համար հաշվենք $z^{x,1}$ =$w^l a^{x,l-1} + b^l$ և $a^{x,l}$=σ($z^{x,l}$) արտահայտությունների արժեքները։
   - **Ելքային սխալանքը՝** $\delta^{x,L}$**:** Վեկտորի հաշվումը. $\delta^{x,L}$=$\nabla_a C_x$ ⊙ σ'($z^{x,l}$) .
   - **Սխալանքի հետադարձ տարածումը:** Յուրաքանչյուր l=L−1,L−2,…,2 հաշվել $\delta^{x,l}$=(($w^{l+1})^T$ $\delta^{x,l+1}$)⊙ σ'($z^{x,l}$).
3. **Գրադիենտային վայրէջք:** Յուրաքանչյուր l= L, L−1, …,2 թարմացնել կշիռները ըստ հետևյալ կանոնի՝ $w^l \rightarrow w^l - \frac{\eta}{m}\sum_x \delta^{x,l} (a^{l-1})^T$, և շեղումներն ըստ $b^l \rightarrow b^l - \frac{\eta}{m}\sum_x \delta^{x,l}$ կանոնի.

Ի՞նչ իմաստով է հետադարձ տարածումն արագագործ ալգորիթմ

Ի՞նչ իմաստով է հետադարձ տարածումն արագագործ ալգորիթմ։ Այս հարցին պատասխանելու համար դիտարկենք գրադիենտի հաշվման այլ մոտեցում։ Դիտարկենք գնի ֆունկցիան կախված միայն կշիռներից C=C(w) (շեղումներին դեռ կվերադառնանք)։ Այնուհետև համարակալում ենք կշիռները՝ $w_1, w_2, \ldots$ և հաշվում $\partial C/\partial w_j$ մասնական ածանցյալները տրված w_j կշռի համար։ Կարելի է օգտագործել հետևյալ մոտարկումը՝ (6)

$$\frac{\partial C}{\partial w_j} \approx \frac{C(w+\epsilon e_j) - C(w)}{\epsilon},$$

որտեղ $\epsilon>0$ փոքր դրական թիվ է և e_j միավոր վեկտորն է j-րդ ուղղության վրա: Այսպիսով, կարող ենք գնահատել $\partial C/\partial w_j$ մասնական ածանցյալները հաշվելով C գինը երկու տարբեր (իրար մոտ) w_j արժեքների համար և կիրառել (6) հավասարումը: Նույն ձևով կարող ենք հաշվել նաև $\partial C/\partial b$ շեղումների դեպքում:

Այս մոտեցումը խոստումնալից է թվում: Այն հիմնված է համեմատաբար պարզ կոնցեպտների վրա և բավականին դյուրին է իրականացնելը:

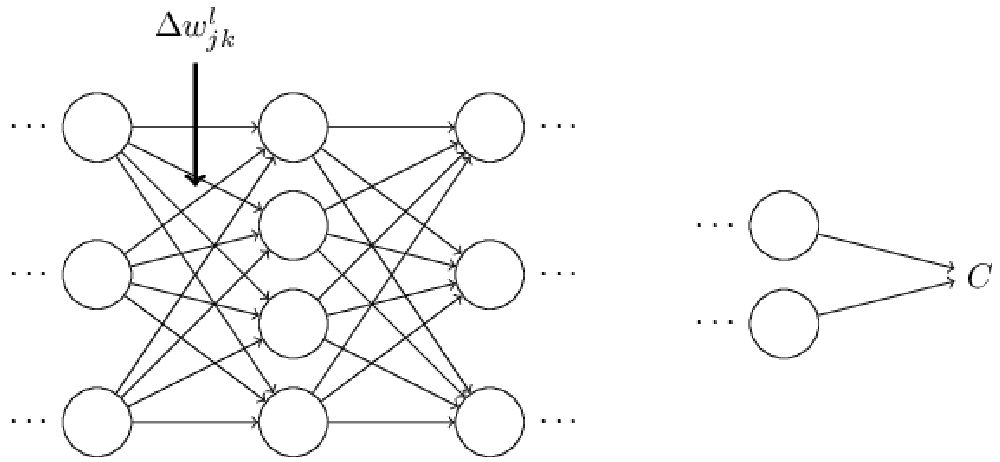
Դժբախտաբար, մինչդեռ այս մոտեցումը խոստումնալից է թվում, իրականացումը բավականին դանդաղագործ է: Որպեսզի հաշվենք գրադիենտը, պետք է գնային ֆունկցիայի արժեքը մի քանի անգամներ հաշվենք, ինչի հետևանքով մի քանի անգամ պետք է առաջ շարժվենք ցանցով: Դա չի ներառում C(w) արժեքի հաշվումը:

Հետադարձ տարածման առավելությունը կայանում է նրանում, որ այն միաժամանակ հաշվում է *բոլոր* $\partial C/\partial w_j$ մասնական ածանցյալները՝ օգտագործելով միայն մեկ առաջ տարածում և մեկ հետադարձ տարածում ցանցով: Հետադարձ առաջխաղացման հաշվարկային արագությունը (computational cost) համարյա նույնն է, ինչ առաջ տարածման : Հետևաբար հետադարձ տարածման ընդհանուր հաշվարկային գինը նույնն է, ինչ ցանցում երկու առաջաբեր գործողություն կատարելիս: Համեմատենք դա միլիոնից ավել առաջաբեր գործողությունների հետ, որ անհրաժեշտ կլիներ, եթե մենք որդեգրեինք ներքևում նշված հավասարման հիման վրա կառուցված մեթոդը: (6) Եվ այսպիսով, չնայած նրան, որ հետադարձ տարածումն ունի բավականին բարդ տեսք, քան (6) մոտեցումը, այն անհամեմատ ավելի արագագործ է:

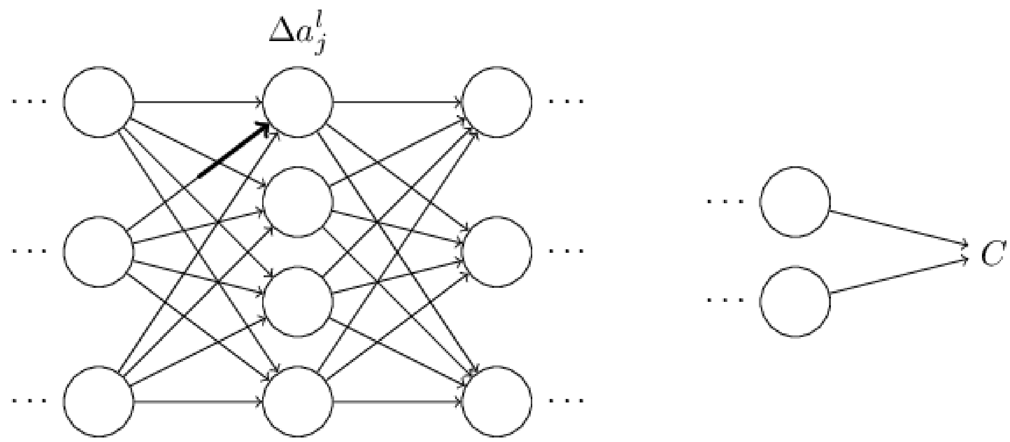
Հետադարձ տարածում. ամբողջական պատկերը

Այսպիսով, գոյություն ունեն առեղծվածներ կապված հետադարձ տարածման հետ: Առաջին. ի՞նչ է ըստ էության ալգորիթմն անում: Իհարկե, մենք ցույց տվեցինք, որ սխալանքը ելքից տարածվում է դեպի ետ, սակայն արդյո՞ք կարող ենք ավելի խորանալ և ձեռք բերել ինտուցիա այն մասին, թե ինչ է տեղի ունենում երբ մենք այսքան մատրիցներ և վեկտորներ իրար ենք բազմապատկում: Երկրորդ առեղծվածն այն է, թե ինչպե՞ս կարելի է հայտնագործել հետադարձ տարածման ալգորիթմը: Իհարկե կարելի է հետևել և հասկանալ ալգորիթմի քայլերն ու ապացույցը, սակայն արդյո՞ք խնդիրը հասկանում ենք այնքան խորությամբ, որ կարող էինք ալգորիթմն ինքնուրույն հայտնագործել: Արդյոք կա տրամաբանական շղթա, որին հետևելով կարող էինք հանգել հետադարձ տարածման ալգորիթմին: Այս հատվածում կդիտարկենք այդ երկու առեղծվածները:

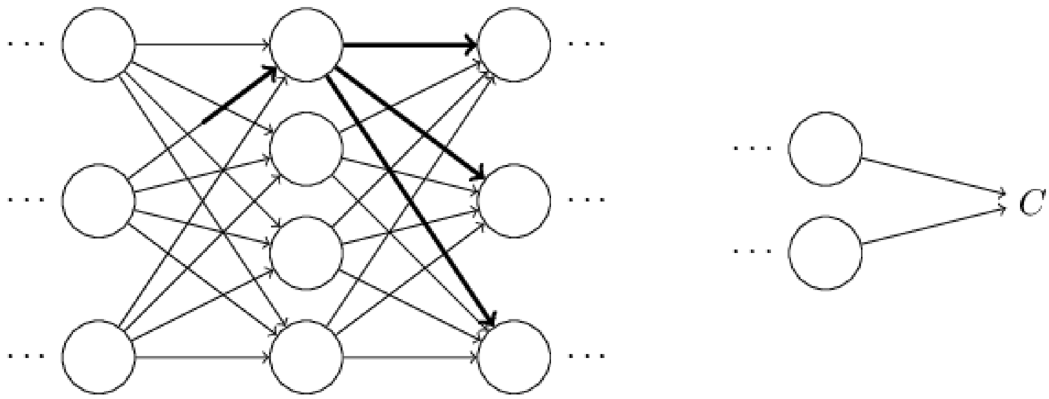
Որպեսզի մեր ինտուցիան այս ալգորիթմի վերաբերյալ բարելավենք, ենթադրենք, որ ցանցի որևիցե w_{jk}^l կշռի նկատմամբ կիրառել ենք Δw_{jk}^l փոփոխությունը.



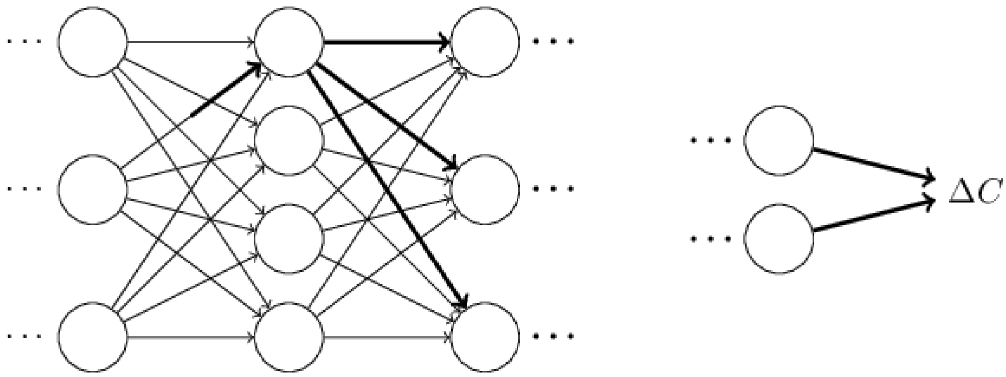
Այս փոփոխությունը կհանգեցնի համապատասխան փոփոխություն այդ նեյրոնին համապատասխանող ելքային ակտիվացիայում:



Որն իր հերթին կհանգեցնի հաջորդ շերտի *բոլոր* ակտիվացիաների փոփոխության:



Այդ փոփոխություննեն իրենց հերթին կառաջացնեն փոփոխութւուններ հաջորդ շերտում, այնուհետև հաջորդ և այդպես շարունակ մինչև վերջնական շերտը հանգեցնելով փոփոխության գնային ֆունկցիայում:



Գնային ֆունկցիայի ΔC փոփոխությունը կշռի Δw_{jk}^{l} փոփոխության հետ կապված է հետևյալ հավասարմամբ. (7)

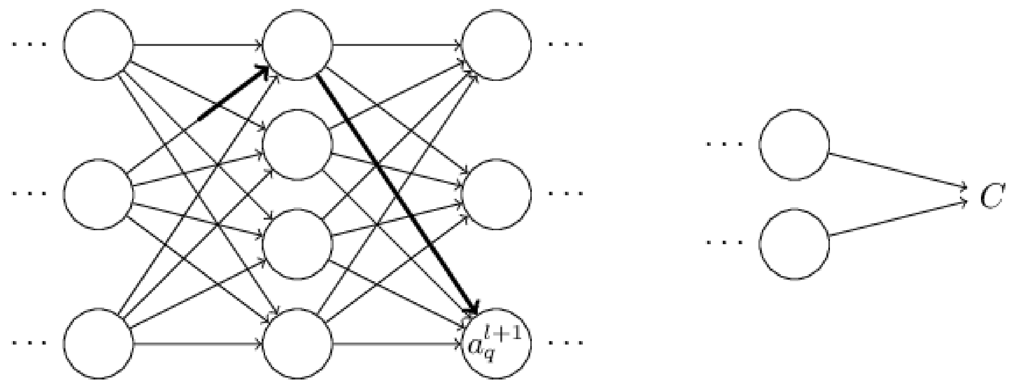
$$\Delta C \approx \frac{\partial C}{\partial w_{jk}^{l}} \Delta w_{jk}^{l}.$$

Սա ցույց է տալիս, որ $\frac{\partial C}{\partial w_{jk}^{l}}$ հաշվարկելու հնարավոր տարբերակներից մեկն այն է, որ հետևենք, թե ինչպես է w_{jk}^{l} տարծվում և փոքրիկ փոփոխություն առաջացնում C գնային ֆունկցիայում: Եթե դա ուշադիր կերպով իրականացնենք, հաշվելով բոլոր միջանկյալ արժեքները, ապա կարող ենք հաշվել $\frac{\partial C}{\partial w_{jk}^{l}}$ արժեքը…

Փորձենք դա իրականացնել: Δw_{jk}^{l} փոփոխությունն առաջացնում է Δa_{j}^{l} փոքրիկ փոփոխությունը l-րդ շերտի j-րդ ակտիվացիայում: Այդ փոփոխությունը կարելի է նկարագրել հետևյալ կերպ: (8)

$$\Delta a_j^l \approx \frac{\partial a_j^l}{\partial w_{jk}^l} \Delta w_{jk}^l.$$

Փոփոխությունը Δa_j^l ակտիվացիայում կառաջացնի փոփոխություններ հաջորդ *բոլոր* շերտերի ակտիվացիաներում, օրինակ, (l+1)-րդ շերտում։ Դիտարկենք, թե ինչպես է փոփոխվում այդ ակտիվացիաներից որևիցէ մեկը, օրինակ՝ a_q^{l+1} ։



Ըստ էության, հետևյալ փոփոխությունը տեղի կունենա. (9)

$$\Delta a_q^{l+1} \approx \frac{\partial a_q^{l+1}}{\partial a_j^l} \Delta a_j^l.$$

Կատարելով փոխարինում (8) , հավասարումով, կստանանք. (10)

$$\Delta a_q^{l+1} \approx \frac{\partial a_q^{l+1}}{\partial a_j^l} \frac{\partial a_j^l}{\partial w_{jk}^l} \Delta w_{jk}^l.$$

Իհարկե, Δa_q^{l+1} ակտիվացիայի փոփոխությունն իր հերթին կհանգեցնի փոփոխությունների հաջորդ շերտերում։ Կարող ենք պատկերացնել ցանցում այդպիսի փոփոխությունների ուղի՝ w_{jk}^l-ից դեպի C գնային ֆունկցիա, որտեղ յուրաքանչյուր փոփոխություն ակտիվացիայում հանգեցնում է հաջորդ շերտի ակտիվացիաների փոփոխության, վերջապես շղթայաբար հասնելով գնային ֆունկցիայի փոփոխության ելքում։ Եթե այդ ուղին անցնում է a_l^l , a_q^{l+1} ,…, a_n^{L-1} , a_m^L ակտիվացիաներով, ապա գնային ֆունկցիայի փոփոխությունը կարելի է նկարագրել հետևյալ կերպ. (11)

$$\Delta C \approx \frac{\partial C}{\partial a^L_m}\frac{\partial a^L_m}{\partial a^{L-1}_n}\frac{\partial a^{L-1}_n}{\partial a^{L-2}_p}\ldots\frac{\partial a^{l+1}_q}{\partial a^l_j}\frac{\partial a^l_j}{\partial w^l_{jk}}\Delta w^l_{jk},$$

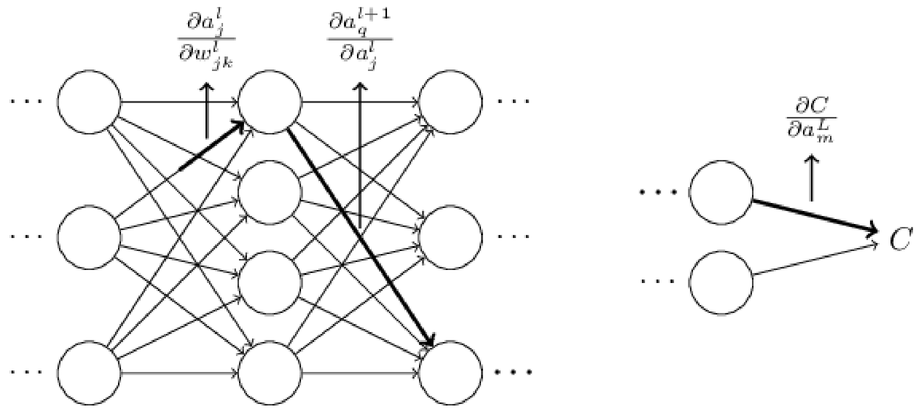
որտեղ յուրաքանչյուր փոփոխված նեյրոնի համար կիրառել ենք ∂a/∂a տիպի փոփոխություն նկարագրող արտահայտությունը՝ վերջում կիրառելով ∂C/∂a_m^L գնային ֆունկցիայի փոփոխությունը կախված ելքային նեյրոնից։ Այսպիսով, C-ի փոփոխությունը կարելի է ներկայացնել այս կերպ՝ հիմնված ակտիվացիաների շղթայական փոփոխությունների վրա ։ C գնային ֆունկցիայի ամբողջական փոփոխությունը հաշվելու համար պետք է գումարել փոփոխությունները տրված կշռից դեպի գնային ֆունկցիա բոլոր հնարավոր ուղիներով։ Օրինակ. (12)

$$\Delta C \approx \sum_{mnp\ldots q}\frac{\partial C}{\partial a^L_m}\frac{\partial a^L_m}{\partial a^{L-1}_n}\frac{\partial a^{L-1}_n}{\partial a^{L-2}_p}\ldots\frac{\partial a^{l+1}_q}{\partial a^l_j}\frac{\partial a^l_j}{\partial w^l_{jk}}\Delta w^l_{jk},$$

որտեղ գումարն ըստ բոլոր հնարավոր միջանկյալ նեյրոնների է։ Համեմատելով (7) հավասարման հետ, կտեսնենք, որ (13)

$$\frac{\partial C}{\partial w^l_{jk}} = \sum_{mnp\ldots q}\frac{\partial C}{\partial a^L_m}\frac{\partial a^L_m}{\partial a^{L-1}_n}\frac{\partial a^{L-1}_n}{\partial a^{L-2}_p}\ldots\frac{\partial a^{l+1}_q}{\partial a^l_j}\frac{\partial a^l_j}{\partial w^l_{jk}}.$$

Այժմ (13) հավասարումը բարդ տեսք ունի։ Սակայն այն ունի պարզ ինտուիտիվ մեկնաբանություն։ Մենք հաշվում ենք C փոփոխության գործակիցը կախված ցանցի կշռից։ Այնուամենայնիվ, այն ունի գեղեցիկ ինտուիտիվ մեկնաբանություն։ Մենք հաշվարկում ենք C-ի փոփոխության արագությունը ցանցում որոշակի քաշի նկատմամբ։ Հավասարումը մեզ ասում է, որ ցանցում երկու նեյրոնների միջև յուրաքանչյուր եզր կապված է արագության գործակցի հետ, որը մեկ նեյրոնի ակտիվացման մասնակի ածանցյալն է մյուս նեյրոնի ակտիվացման նկատմամբ։ Առաջին քաշից մինչև առաջին նեյրոնը ընկած եզրն ունի արագության գործակից $\partial a_j^l/\partial w_{jk}^l$։ Ճանապարհի արագության գործակիցը պարզապես ճանապարհի երկայնքով արագության գործոնների արտադրյալն է։ Եվ փոփոխության ընդհանուր արագությունը $\partial C/\partial w_{jk}^l$ պարզապես բոլոր ուղիների արագության գործոնների գումարն է սկզբնական քաշից մինչև վերջնական արժեքը: Այս ընթացակարգը պատկերազարդված է այստեղ՝ մեկ ճանապարհի համար:



Կարող ենք ստանալ հստակ արտահայտություններ (Է3) հավասարման բոլոր առանձին մասնակի ածանցյալների համար: Դա հեշտ է անել մի փոքր հաշվարկով: Դա անելուց հետո կարող ենք փորձել պարզել, թե ինչպես գրել ինդեքսների վրա բոլոր գումարները որպես մատրիցային բազմապատկումներ: Սա պահանջում է որոշակի համառություն: Այս ամենը անելուց և այնուհետև հնարավորինս պարզեցնելուց հետո, հայտնաբերում ենք, որ ի վերջո ստանում ենք հենց հետադարձ տարածման ալգորիթմը: Եվ այսպես, կարող ենք մտածել հետադարձ տարածման ալգորիթմի մասին որպես այս բոլոր ուղիների համար գումարի արագության գործակցի վրա հաշվարկման միջոց: Կամ, մի փոքր այլ կերպ ասած, հետադարձ տարածման ալգորիթմը խելացի միջոց է կշիռների (և շեղումների) փոքր տատանումները հետևելու համար, երբ դրանք տարածվում են ցանցով, հասնում ելքային արժեքին և այնուհետև ազդում արժեքի վրա: