

АНАЛІЗ ФУНКЦІЇ ПОХИБКИ ЗА УМОВИ ЗАСТОСУВАННЯ АЛГОРИТМУ ОПТИМІЗАЦІЇ AMSGrad

С. Свелеба¹, І. Катеринчук¹, І. Куньо¹, О. Семотюк², Я. Шмигельський¹,
С. Вельгош¹, А. Копач¹, В. Стахура¹

¹ Львівський національний університет імені Івана Франка,
вул. Ген. Тарнавського, 107, 79017 Львів, Україна
incomlviv@gmail.com

²Українська Академія Друкарства,
вул. Під Голоском, 19, 79020 Львів, Україна

В роботі з допомогою логістичної функції, яка описує процес подвоєння, та Фур'є спектрів функції похибки було проведено тестування стохастичного методу оптимізації AMSGrad.

Реалізація алгоритму оптимізації градієнтного спуску за допомогою AMSGrad було здійснено для простої двовимірної функції, яка зводить у квадрат вхідні дані кожного виміру та визначає діапазон допустимих вхідних даних від -1,0 до 1,0.

Встановлено, що процес перенавчання супроводжується зміною швидкості цільової функції похибки, а Фур'є спектрам притаманна поява гармоніки. Показано, що при незначному наборі вхідних даних, коли значення β_2 близьке до 1, а $\beta_1 = 0.9$ спостерігається нестабільність в навчанні, яка зумовлена процесом перенавчання.

Ключові слова: оптимізаційні методи, функція похибки, AMSGrad, швидкість навчання, діаграми розгалуження.

Градієнтний спуск є одним із найпопулярніших алгоритмів для оптимізації та найпоширенішим способом оптимізації нейронних мереж. Водночас кожна найсучасніша бібліотека Deep Learning містить реалізації різноманітних алгоритмів для оптимізації градієнтного спуску (наприклад, документація lasagne, caffe та keras) [1]. Однак ці алгоритми, як зазначається в роботі [1], часто використовуються, як оптимізатори чорної скриньки, оскільки важко знайти практичне пояснення їхніх сильних і слабких сторін.

Обмеження градієнтного спуску полягає в тому, що один розмір кроку (швидкість навчання) використовується для всіх вхідних змінних. Розширення градієнтного спуску, як-от алгоритм Adaptive Movement Estimation (Adam), використовують окремий розмір кроку для кожної вхідної змінної, але в результаті розмір кроку може швидко зменшуватися до дуже малих значень [2].

AMSGrad є розширеною версією оптимізаційного методу градієнтного спуску Adam, який намагається покращити властивості конвергенції алгоритму, уникаючи великих різких змін у швидкості навчання для кожної вхідної змінної.

Алгоритм AMSGrad був описаний у статті [3]. Як правило, метод Adam автоматично адаптує окремий розмір кроку (швидкість навчання) для кожного параметра в задачі

оптимізації. Обмеженням методу Adam є те, що він може одночасно зменшувати розмір кроку, наближаючись до глобального мінімуму, це добре, але в деяких випадках він також збільшує розмір кроку, це погано. Метод AdaGrad усуває це. AdaGrad — це розширення для Adam, яке підтримує максимум вектора другого моменту $v(t)$ та використовує його для корекції зміщення градієнта, який використовується для оновлення параметра, замість самого вектора моменту. Це допомагає зупинити надто швидке сповільнення оптимізації (наприклад, передчасну конвергенцію) [3].

Ключова відмінність AMSGrad від Adam полягає в тому, що він підтримує максимум усіх $v(t)$ до поточного кроку часу та використовує це максимальне значення для нормалізації поточного середнього градієнта [4].

На початку обчислюються градієнти (часткові похідні) для поточного кроку за часом:

$$g(t) = f'(x(t-1)), \text{ де } f(x(t)) - \text{цільова функція.}$$

Далі вектор першого моменту $m(t)$ оновлюється за допомогою градієнта та гіперпараметра β_1 :

$$m(t) = \beta_1 m(t-1) + (1 - \beta_1) \cdot g(t).$$

Гіперпараметр β_1 можна підтримувати постійним або експоненціально зменшувати під час пошуку, наприклад: $\beta_1(t) = (\beta_1)^t$.

$$\text{Або, по черзі: } \beta_1(t) = \beta_1 / t$$

Вектор другого моменту оновлюється за допомогою квадрата градієнта та гіперпараметра β_2 : $v(t) = \beta_2 \cdot v(t-1) + (1 - \beta_2) \cdot (g(t))^2$

Далі оновлюється максимум для вектора другого моменту:

$$\hat{v}(t) = \max(\hat{v}(t-1), v(t))$$

Значення параметра w можна оновити за допомогою обчислених умов і гіперпараметра α - розміру кроку: $w(t) = w(t-1) - \alpha \cdot m(t) / \sqrt{\hat{v}(t)}$

Розмір кроку також можна підтримувати постійним або зменшувати за експонентою.

Отже для огляду є три гіперпараметри для алгоритму [2, 3]:

α : початковий розмір кроку (швидкість навчання), типове значення становить 0,002.

β_1 : коефіцієнт розпаду для першого імпульсу, типове значення становить 0,9.

β_2 : коефіцієнт розпаду для нескінченної норми, типове значення становить 0,999.

Розглянемо, як реалізувати алгоритм оптимізації градієнтного спуску за допомогою AMSGrad Momentum на прикладі задачі із роботи [2]. Тобто ми будемо використовувати просту двовимірну функцію, яка зводить у квадрат вхідні дані кожного виміру та визначає діапазон допустимих вхідних даних від -1,0 до 1,0, як це було проведено в роботі [2].

Для даної системи було проведено дослідження діаграми розгалуження з використанням функції відображення виду:

$$y_{n+1} = \eta - y_n - y_n^2,$$

де n - індекс ітерацій, $\eta = \alpha$ - параметр, який визначає швидкість навчання.

Її нерухомі точки:

$$y_{1,2}^* = -1 \pm \sqrt{\eta + 1},$$

власні значення, яких можна обчислити наступним чином:

$$\rho_{1,2}^* = 1 \mp 2\sqrt{\eta+1}.$$

Вибір даного логістичного відображення обумовлено тим, що воно описує процес подвоєння частоти коливань [5]. В нашому випадку цей процес зумовлений виникненням локальних мінімумів при підході до глобального мінімуму. Також було проведено дослідження Фур'є спектрів функції похибки в залежності від параметрів β_1 , β_2 та α .

На рис. 1 наведено діаграму розгалуження від α , Фур'є спектри та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad за умови 10 ітерацій.

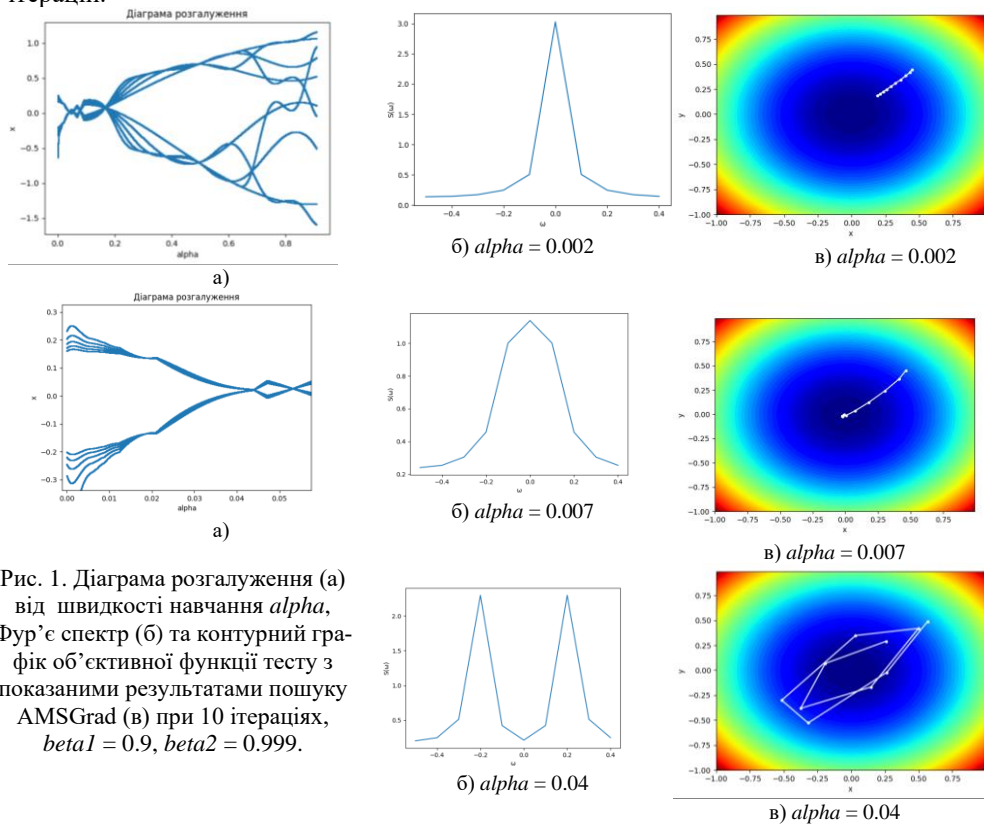


Рис. 1. Діаграма розгалуження (а) від швидкості навчання α , Фур'є спектр (б) та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad (в) при 10 ітераціях, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Згідно з рис.1 в залежності від швидкості навчання може проходити процес недонавчання ($\alpha \leq 0.002$), перенавчання ($0.02 > \alpha > 0.005$) та практично відсутність навчання ($\alpha > 0.03$). Тобто недонавчання системи можна охарактеризувати як недосягнення мінімуму функції похибки (рис. 1, в). Процес перенавчання супроводжується появою гармоніки у Фур'є спектрах (розширенням смуги Фур'є спектру рис. 1, б), що зумовлене неодноразовим переходом функції похибки через глобальний мінімум. Відсутність процесу навчання супроводжується існуванням гармоніки функції похибки, за умови перескоку через глобальний мінімуму (рис. 1, в). Подальше збільшення швидкості нав-

чання відразу не приводить до появи хаотичного стану. Виходячи із діаграми розгалуження (рис. 1, а) його слід було б очікувати при $\alpha > 0.5$. Розглядаючи діапазон зміни α при якому відбувається навчання ($\alpha < 0.01$), можна зазначити, що поведінка логістичної функції, згідно діаграми розгалуження є нелінійною. А процес відсутності навчання супроводжується існуванням каскаду аномальних змін на діаграмі розгалуження (рис. 1 при $\alpha = 0.04$).

Збільшення кількості ітерацій до 100, згідно рис. 2 в залежності від швидкості навчання, може проходити процес недонавчання ($\alpha \leq 0.001$), навчання з незначною похибкою навчання (менше 0.01%, $0.0055 > \alpha > 0.002$), перенавчання ($0.0175 > \alpha > 0.006$) та практично відсутність навчання ($\alpha > 0.02$). Подальше збільшення швидкості навчання, як і при 10 епохах, не приводить одразу до появи хаотичного стану.

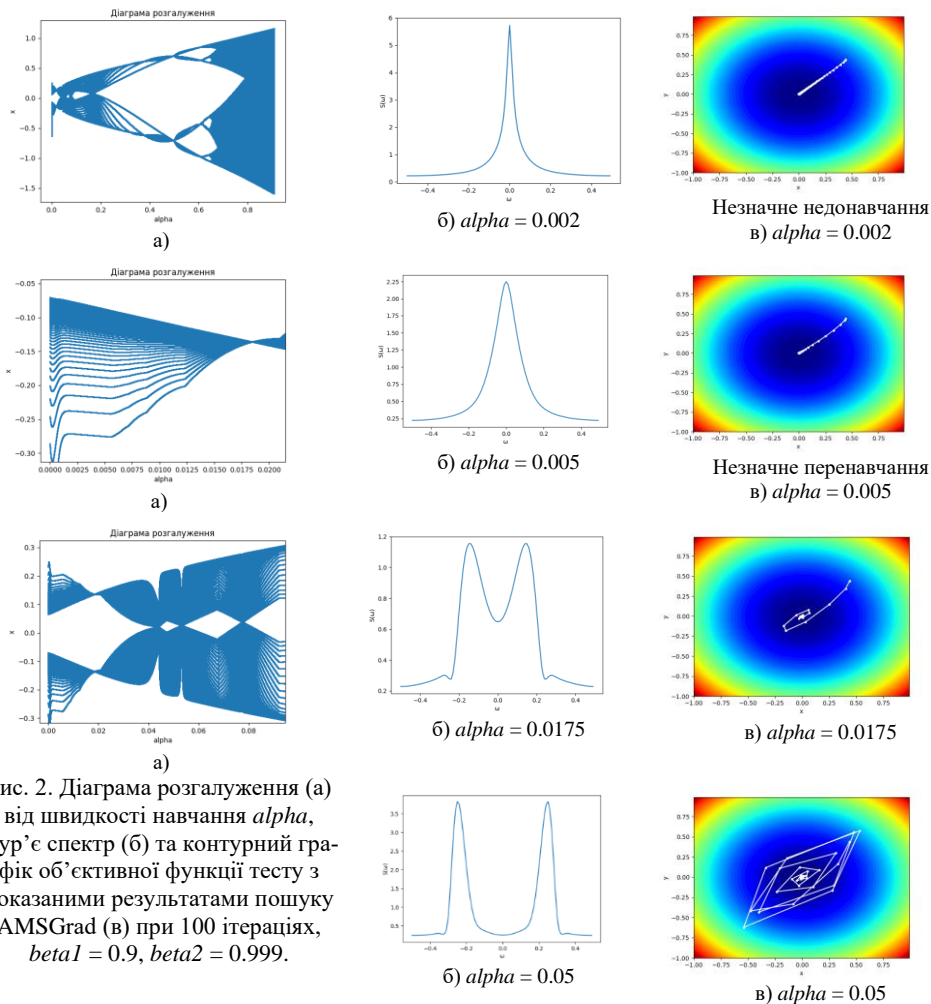


Рис. 2. Діаграма розгалуження (а) від швидкості навчання α , Фур'є спектр (б) та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad (в) при 100 ітераціях, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Виходячи із діаграми розгалуження (рис. 2, а) його слід було б очікувати при $alpha > 0.5$, як і для 10 ітерацій. Розглядаючи діапазон зміни $alpha$ при якому відбувається навчання ($alpha < 0.01$) поведінка логістичної функції згідно діаграми розгалуження також є нелінійною.

Процес перенавчання супроводжується існуванням каскаду аномальних змін на діаграмі розгалуження. Як і при 10 ітерацій, так і при 100 ітераціях характер залежності Фур'є спектрів так і результат пошуку на контурному графіку функції похибки є аналогічний. Процес навчання з мінімальною похибкою супроводжується наявністю мало змінною лінійною залежністю логістичної функції від швидкості навчання на діаграмі розгалуження (рис. 2, а діапазон $alpha = 0.0001 \div 0.0055$). На відміну від 10 ітерацій, при 100 ітераціях за умови, що $alpha > 0.03$ все ж проходить процес навчання і система досягає мінімуму, але функція похибки при цьому характеризується існуванням гармонік (рис. 2, б, в). Збільшення кількості ітерацій не приводить до кардинальної зміни поведінки діаграми розгалуження, Фур'є спектрів функції похибки і результату пошуку методу AMSGrad на контурному графіку об'єктивної функції тесту (рис. 3).

В даному оптимізаційному методі AMSGrad важливе значення відіграють гіперпараметри $beta1$ і $beta2$, тобто коефіцієнт розпаду для першого імпульсу і коефіцієнт розпаду для нескінченної норми. При цьому вектор першого моменту оновлюється за допомогою градієнта та гіперпараметра $beta1$, а другий вектор моменту оновлюється за допомогою квадрату градієнта та гіперпараметра $beta2$. Згідно роботи [2, 3] типове значення коефіцієнта розпаду для першого імпульсу становить 0.9, а значення для коефіцієнта розпаду для нескінченної норми, становить 0,999. На початку розглянемо вплив значення коефіцієнта розпаду на величину функції похибки навчання. На Рис.4 наведено діаграму розгалуження від величини гіперпараметра $beta2$, Фур'є спектри та контурний графік об'єктивної функції тесту з результатами пошуку AMSGrad, за умови 10 ітерацій при побудові діаграми розгалуження (рис. 4, а) і 1000 для Фур'є спектру (рис. 4, б), і результату пошуку оптимізаційного методу AMSGrad (рис. 4, в). Швидкість навчання вибиралась виходячи із вище наведених результатів, а саме: $alpha = 0.02$ - швидкість задовільного навчання; $alpha = 0.05$ – швидкість при якому проходить перенавчання системи. При швидкості найкращого навчання, за умови $beta2 > 0.1$ до $beta2 < 0.98$ проходить процес навчання з похибкою $\approx \pm 10\text{-}6\%$ (рис. 4, в). Починаючи з $beta2 > 0.98$ проходить процес перенавчання. Згідно результату пошуку оптимізаційного методу AMSGrad, процес перенавчання супроводжується появою спіральної залежності результату пошуку (рис. 4, в).

Згідно з отриманою діаграмою розгалуження залежності похибки навчання від параметра $beta2$ (рис. 4, а), перенавчання супроводжується збільшенням швидкості зміни похибки навчання від параметра $beta2$, а Фур'є спектри появою гармоніки (рис. 4, б). А саме розширення центрального піку на залежності потужності сигналу від частоти при $alpha = 0.02$, $B2=0.99$, та розщеплення його на кілька при $alpha = 0.02$, $B2=0.999$ (рис.4, а, б). Збільшення швидкості навчання до $alpha = 0.05$, за умови $beta2 > 0.1$ до $beta2 < 0.88$ проходить процес навчання з похибкою $\approx \pm 10\text{-}6\%$ (рис. 4, в). Починаючи з $beta2 > 0.9$ спостерігається процес перенавчання з появою спіральної залежності результату пошуку на контурному графіку об'єктивної функції тесту. Фур'є спектри функції похибки характеризуються наявністю гармоніки (рис. 4, б).

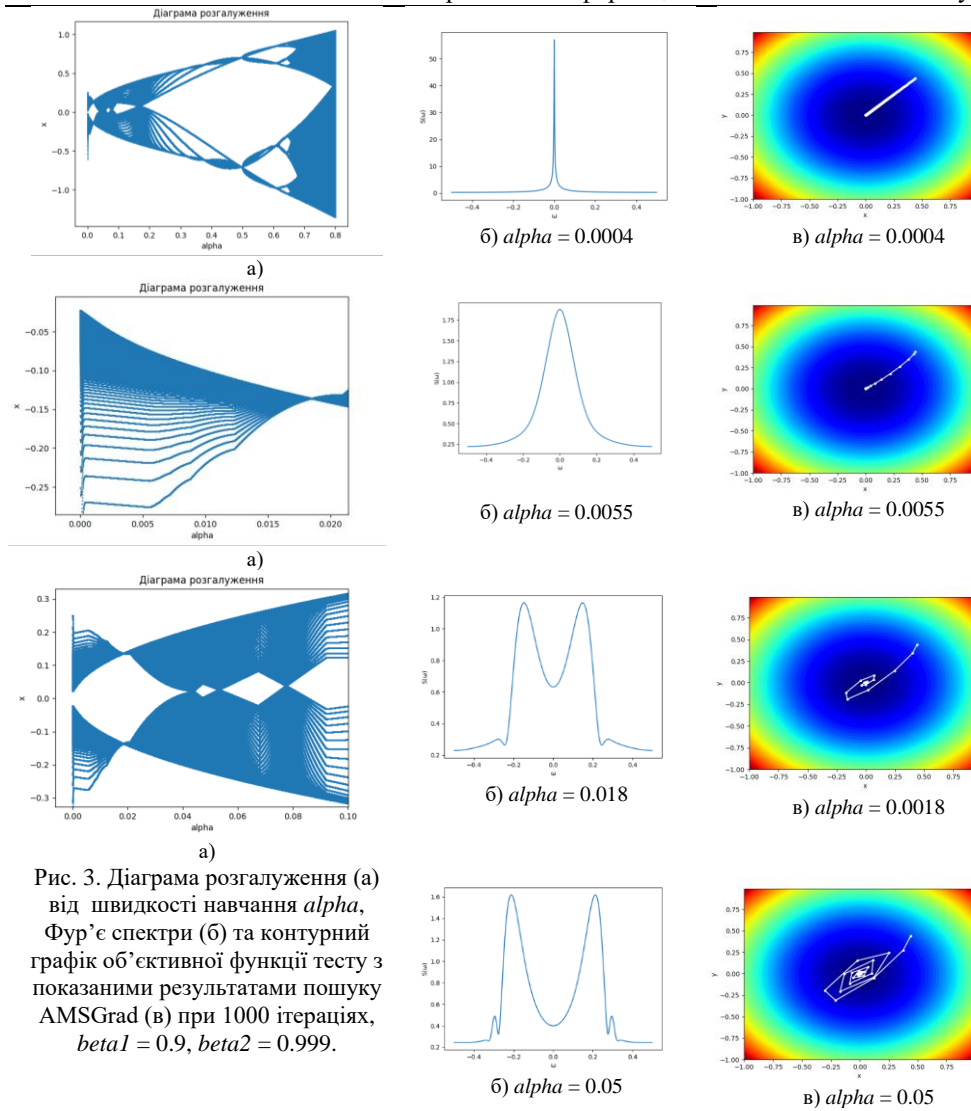


Рис. 3. Діаграма розгалуження (а) від швидкості навчання α , Фур'є спектри (б) та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad (в) при 1000 ітераціях, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Як і при $\alpha = 0.02$ так і за умови $\alpha = 0.05$ діаграм розгалуження притаманна практично лінійна залежність похибки навчання від величини β_2 . Збільшення швидкості зміни функції похибки від величини β_2 пов'язане з появою гармоніки функції похибки, тобто перенавчанням системи.

Отже в досліджуваній системі процес задовільного навчання проходить за умови $\beta_1 = 0.9$, $\alpha \leq 0.05$, 1000 ітерацій, при цьому зміна гіперпараметр β_2 знаходиться в інтервалі $0.1 \div 0.9$. Подальше збільшення величини β_2 сприяє появі процесу перенавчання системи.

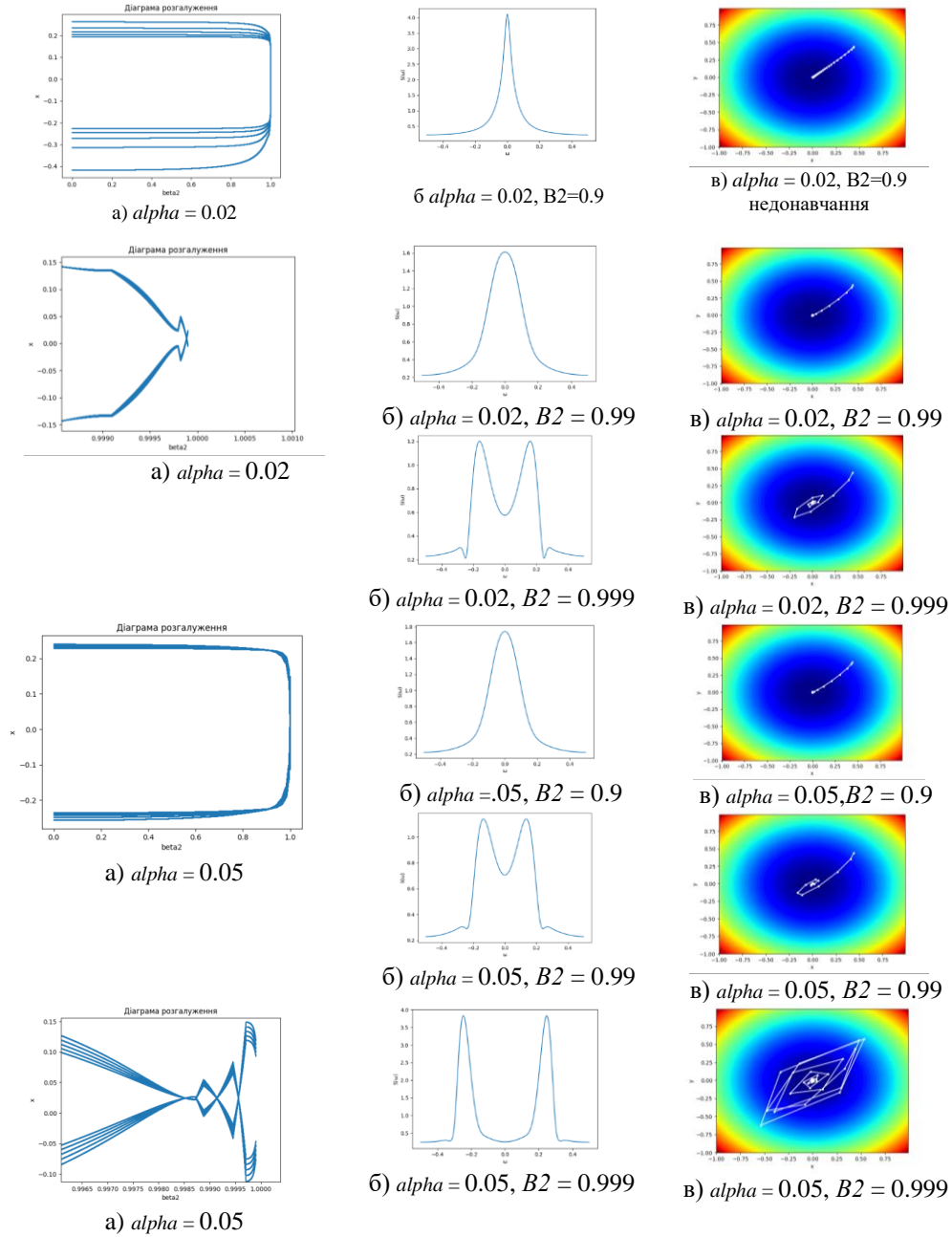


Рис. 4. Діаграма розгалуження (а) від величини гіперпараметра β_{n+2} , Фур'є спектри (б) та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad (в) при 10 ітераціях, $\beta_{n+1} = 0.9$.

Переходячи до розгляду впливу вектора першого моменту, який оновлюється за допомогою градієнта та гіперпараметра β_1 , слід зазначити, що за умови $\alpha > 0.01$ проходить перенавчання у всьому діапазоні зміни β_1 ($0.1 < \beta_1 < 0.999\dots$). Коли ж $\alpha = 0.002$ (як це зазначалось в праці [2, 3]), то проходить процес навчання практично у всьому діапазоні змін β_1 ($0.1 < \beta_1 < 0.999$). За умови $\beta_1 > 0.999$ спостерігається процес перенавчання.

На рис. 5 наведено діаграму розгалуження від величини гіперпараметра β_1 , Фур'є спектри та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad, за умови 10 ітерацій, для побудови діаграми розгалуження (рис. 5, а) і 1000 для Фур'є спектру (рис. 5, б), і результату пошуку оптимізаційного методу AMSGrad (Рис.5, в). Згідно з рис. 5 при $\alpha = 0.005$ і $\alpha = 0.01$ у всьому інтервалі зміни β_1 простежується процес перенавчання.

В праці [3] емпірично оцінено вплив членів корекції зміщення, при видаленні членів поправки на зміщення, які приводять до версії RMSProp [6] з імпульсом. В праці [3] під час навчання варіаційного автокодера (VAE) з такою ж архітектурою, як у [7] з одним прихованим шаром із 500 прихованими одиницями з нелінійністю `softplus` і 50-вимірною сферичною прихованою змінною Гауса, проводилось дослідження при зміні параметра β_1 і β_2 . Зміна параметрів проводилась в широкому діапазоні вибору гіперпараметрів, тобто β_1 [0, 0.9] і β_2 [0.99, 0.999, 0.9999], а також $\log_{10}(\alpha)$ [5, ..., 1]. Значення β_2 вибиралось близьким до 1, яке необхідне для стійкості для розріджених градієнтів, приводить до більшого зміщення ініціалізації, тому, що інтервал корекції зсуву є важливим у випадках повільного затухання, запобігаючи тим самим несприятливому впливу на оптимізацію. Значення β_2 , близькі до 1, справді призводять до нестабільності в навчанні, коли не було поправки на зміщення, особливо в перші кілька епох навчання. Найкращі результати були досягнуті з малими значеннями $(1 - \beta_2)$ [3] і корекцією зміщення. Це стає більш очевидним наприкінці оптимізації, коли градієнти мають тенденцію ставати рідшими, оскільки приховані блоки спеціалізуються на певних шаблонах. Тобто в праці [3] був представлений простий і обчислювально ефективний алгоритм для градієнтної оптимізації стохастичних цільових функцій. Цей метод спрямований на вирішення проблем машинного навчання при великих наборах даних та/або просторів параметрів великої розмірності. Метод поєднує в собі переваги двох нещодавно популярних методів оптимізації: здатність AdaGrad працювати з розрідженими градієнтами та здатність RMSProp працювати з нестационарними завданнями.

В нашому випадку при незначному наборі вхідних даних, коли значення β_2 , близьке до 1, теж приводить до нестабільності в навчанні. А саме до появи процесу перенавчання.

Розглянутий вище алгоритм оновлює експоненціальні ковзні середні градієнта ($m(t)$) і квадрата градієнта ($v(t)$), де гіперпараметри β_1 , β_2 [0, 1) контролюють експоненціальні швидкості спаду цих ковзних середніх [3]. Ковзні середні є оцінками 1-го моменту (середнє) і 2-го необробленого моменту (нецентрована дисперсія) градієнта. Однак ці ковзні середні ініціалізуються як (вектори) 0, що призводить до миттєвих оцінок, які зміщуються до нуля, особливо на початкових етапах часу, і особливо, коли швидкості спаду малі (тобто коли β_1 , β_2 близькі до 1). Гарна новина полягає в тому, що цьому зміщенню ініціалізації можна легко протидіяти, шляхом виправлення зміщення оцінки $m(t)$ і $v(t)$, як це описано в праці [3]. Отже метод AMSGrad використовує умови корекції зсуву ініціалізації.

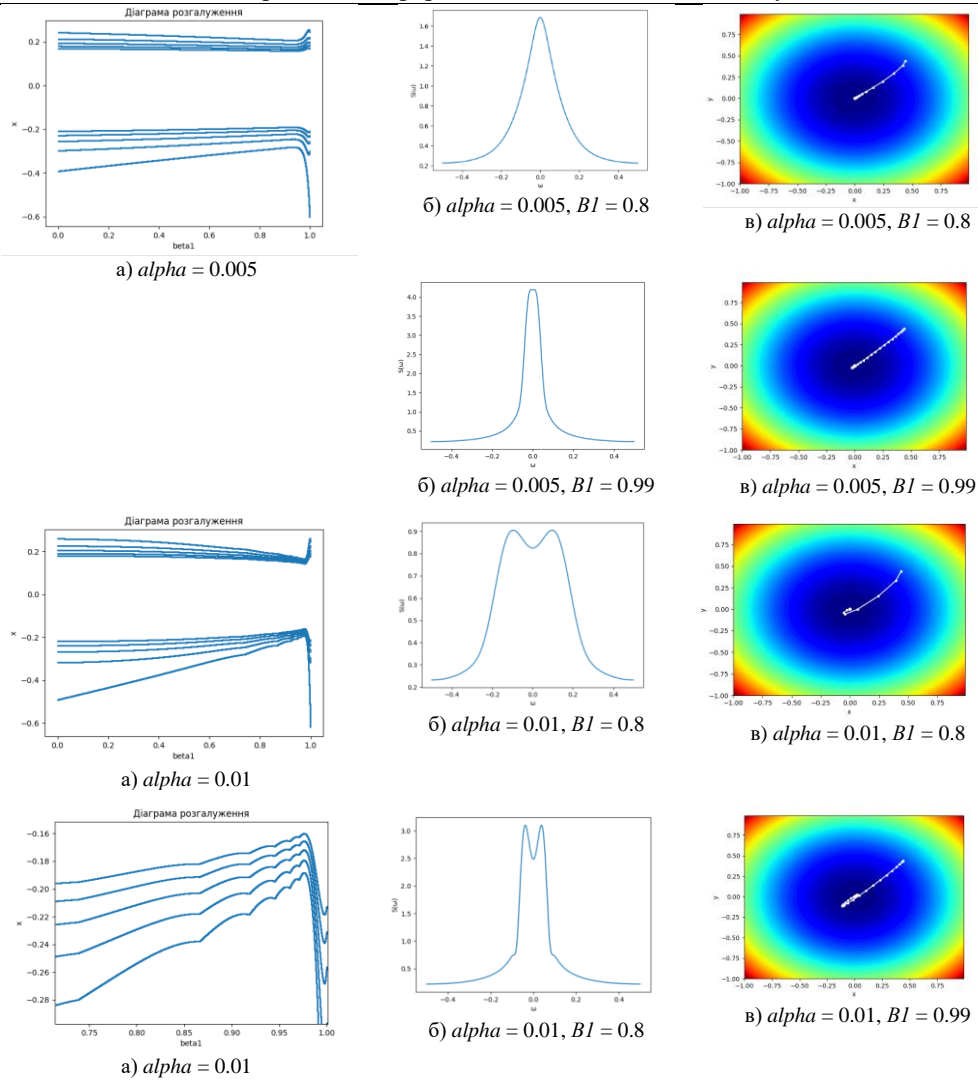


Рис. 5. Діаграма розгалуження (а) від величини гіперпараметра β_1 , Фур'є спектри (б) та контурний графік об'єктивної функції тесту з показаними результатами пошуку AMSGrad (в) при 10 ітераціях, $\beta_2 = 0.999$.

Розглянемо доданок для оцінки другого моменту[2, 3]. Нехай g буде градієнтом стохастичної цільової функції f , і ми хочемо оцінити його другий початковий момент (нецентровану дисперсію), використовуючи експоненційне ковзне середнє квадратичного градієнта зі швидкістю спаду β_2 . Нехай g_1, \dots, g_T — градієнти на наступних часових кроках, кожен з яких отримано з базового розподілу градієнта $g_t \sim p(g_t)$. Ініціалізуємо експоненціальне ковзне середнє як $v_0 = 0$ (вектор нулів). Оновлення експоненціального ковзного середнього на кроці часу t здійснюється за виразом

$$v_t = \text{beta2} \cdot v_{t-1} + (1 - \text{beta2}) \cdot g_t \cdot g_t$$

Експоненціальну швидкість згасання beta1 потрібно вибирати таким чином, щоб експоненціальна ковзна середня призначала малі ваги градієнтам, на попередньому кроці. Те, що залишилося, це член $(1 - \text{beta2}(t))$, який викликаний ініціалізацією поточного середнього. Тому в алгоритмі проходить ділення на цей член, щоб виправити зміщення ініціалізації. У випадку розріджених градієнтів, для надійної оцінки другого моменту потрібно усереднити багато градієнтів, вибравши мале значення beta2 ; однак це саме той випадок малого beta2 , де відсутність корекції зміщення ініціалізації призведе до початкових кроків, які є набагато більшими.

Проведені нами дослідження показують, що за умови $\text{beta1} = 0.9$, $\text{beta2} = 0.99$ при швидкості навчання $\alpha = 0.002$ і 100 ітераціях проходить недонавчання з похибкою навчання 6.6%; $\text{beta1} = 0.9$, $\text{beta2} = 0.999$ – недонавчання з похибкою навчання 0.016%; $\text{beta1} = 0.9$, $\text{beta2} = 0.9999$ – перенавчання з похибкою навчання 0.4%. За умови $\text{beta1} = 0.0$, $\text{beta2} = 0.99$ проходить недонавчання з похибкою навчання 5%; $\text{beta1} = 0.0$, $\text{beta2} = 0.999$ – недонавчання з похибкою навчання 10-5%; $\text{beta1} = 0.0$, $\text{beta2} = 0.9999$ – навчання з практично відсутньою похибкою навчання. Отже, отримані нами результати є дещо відмінними від отриманих у праці [3]. Така відмінність може бути пов'язана з малою вибіркою вхідних даних для навчання. Дійсно оптимізаційні методи, до яких відноситься метод AMSGrad, застосовують в згорткових нейромережах для швидкої обробки вхідних даних.

Отже, стохастичний метода оптимізації AMSGrad, який був описаний в роботі [3] і практично продемонстровано його роботу в [2], був протестований з допомогою логістичної функції, яка описує процес подвоєння, та Фур'є спектрів функції похибки. Встановлено що процес перенавчання супроводжується зміною швидкості цільової функції похибки, а Фур'є спектру притаманна поява гармонік. Показано, що при незначному наборі вхідних даних коли, значення beta2 , близьке до 1, а $\text{beta1} = 0.9$ спостерігається нестабільність в навчанні, яка зумовлена процесом перенавчання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] *Mykel J. Kochenderfer, Tim A. Wheeler Algorithms for Optimization/ Mykel J. Kochenderfer, Tim A. Wheeler – The MIT Press.– 2019. – 520p.*
- [2] *Jason Brownlee. Optimization for Machine Learning. Finding Function Optima with Python – The MIT Press. – 2021. – 403 p.*
- [3] *Diederik P. Adam: a method for stochastic optimization / Diederik P. Kingma, Jimmy Lei Ba – Published as a conference paper at ICLR 2015. – 2015. – P. 1-15. DOI: [10.48550/arXiv.1412.0092](https://arxiv.org/abs/1412.0092)*
- [4] *Sashank J. On the Convergence of Adam and Beyond /Sashank J. Reddi, Satyen Kale, Sanjiv Kumar .– Published as a conference paper at ICLR 2018. – 2019. – P. 1-23.*
- [5] *Yu. Taranenko Information entropy of chaos. URL: <https://habr.com/ru/post/447874/>*
- [6] *Tieleman, T. Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. / Tieleman, T. and Hinton, G. COURSERA: Neural Networks for Machine Learning. – 2012. – V.4. – P.26-31.*

- [7] Kingma, Diederik P Auto-Encoding Variational Bayes / Kingma, Diederik P, Welling, Max. – In The 2nd International Conference on Learning Representations (ICLR) – 2013. – V.11. – P. 1-14. DOI: [10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114)

ANALYSIS OF THE ERROR FUNCTION IN CASE OF APPLYING THE AMSGrad OPTIMIZATION ALGORITHM

S. Sveleba¹, I. Katerynychuk¹, I. Kuno¹, O. Semotyuk², Ya. Shmygelsky¹, S. Velgosh¹,
A. Kopach¹, V. Stakhura¹

¹ Lviv National University named after Ivan Franko,
107 Gen. Tarnavskoho St., 79017 Lviv, Ukraine
incomlviv@gmail.com

² Ukrainian Academy of Printing,
Pid Goloskom str.19, Lviv, 79020, Ukraine

In this paper, the AMSGrad stochastic optimization method was tested using the logistic function that describes the doubling process and the Fourier spectra of the error function.

The implementation of the gradient descent optimization algorithm using the AMSGrad method was carried out for a simple two-dimensional function that squares the input data of each measurement and determines the range of acceptable input data from -1.0 to 1.0. The program for minimizing the error function was implemented using Python programming language.

The influence of hyperparameters' values β_1 , β_2 , and the learning rate on optimizing the training of this system was analyzed. Also, branching diagrams from these parameters were constructed.

We found that the retraining process is accompanied by a change in the rate of the target error function, and the Fourier spectra are characterized by the harmonics' appearance. The instability in the learning process caused by the retraining process is shown to be observed in a small set of input data when the value of β_2 is close to 1 and $\beta_1 = 0.9$. Also, we found that a transition accompanies overtraining through the global minimum of the objective function, and multiple passages accompany the transition to chaos through the global minimum.

Keywords: optimization methods, error function, AMSGrad, learning rate, branching diagrams.

Стаття надійшла до редакції 11.06.2023
Прийнята до друку 15.09.2023