

# Weight decay, L2 regularisation and AdamW

Ten problems on weight decay: why an L2 penalty and decoupled decay are the same update under plain SGD and what the shrink factor and half-life are, the ridge solution as the fixed point of both, how momentum amplifies an L2 decay by  $1/(1 - \beta)$ , how Adam normalises it away, the stationary points of AdamW and why they are not the ridge solution, the equilibrium norm and effective learning rate of a scale-invariant layer with and without decay, AdamW as an exponential moving average of updates, decay on a quadratic, and what scaling the loss does to each optimiser, with worked solutions and the mistakes that carry a  $\lambda$  between optimisers that do not share it.

## BEFORE YOU START

Weight decay is one line in every optimiser, and the line means something different in each. Under plain gradient descent, adding  $\frac{\lambda}{2} \|\theta\|^2$  to the loss and shrinking the weights by a factor  $1 - \eta\lambda$  each step are the same thing; under momentum the shrink is amplified; under Adam it is divided away entry by entry; and AdamW, which applies the shrink outside the adaptive step, no longer minimises any penalised loss at all. These ten problems work each case out: the shrink factor and its half-life, the ridge solution as a fixed point, the amplification by  $1/(1 - \beta)$ , the normalised decay inside Adam, AdamW's stationary points, what decay does to a weight matrix that is followed by a normalisation layer (where the loss cannot see the weight's norm at all, and decay turns into a learning rate), AdamW as a moving average of its own updates, the geometry of decay on a quadratic, and how each optimiser responds to the loss being multiplied by a constant. The five mistakes are each a  $\lambda$  carried from one optimiser to another: between SGD and momentum, between a loss and a normalised layer, between Adam and the ridge solution, between two learning rates, and between two loss scales.

- The parameters are  $\theta \in \mathbb{R}^d$ , the data loss is  $L(\theta)$ , and  $g_t = \nabla L(\theta_{t-1})$  is its gradient at step  $t$ ;  $\eta > 0$  is the learning rate and  $\lambda \geq 0$  the decay coefficient. Entrywise operations ( $|\cdot|$ , sign, squares, division) act entry by entry as on the optimiser page.
- **L2 penalty:** the optimiser is run on  $L_\lambda(\theta) = L(\theta) + \frac{\lambda}{2} \|\theta\|^2$ , whose gradient is  $g_t + \lambda\theta_{t-1}$ . This is PyTorch's `weight_decay=` in SGD and in Adam.
- **Decoupled weight decay:** the weights are multiplied by  $1 - \eta\lambda$  and the optimiser's step is computed from  $g_t$  alone:  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta u_t$ , where  $u_t$  is the optimiser's direction ( $g_t$  for SGD,  $v_t$  for momentum,  $\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$  for Adam). With Adam this is AdamW, PyTorch's `AdamW`, whose decay is  $\eta\lambda\theta$  per step with  $\lambda = 0.01$  by default.

- Momentum and Adam are as on the optimiser page: classic momentum  $v_t = \beta v_{t-1} +$  (gradient),  $\theta_t = \theta_{t-1} - \eta v_t$ ; Adam with  $\beta_1, \beta_2$ , bias-corrected  $\hat{m}_t, \hat{v}_t$ , and  $\epsilon$ . That page's Problem 8 showed that a constant gradient  $g$  gives  $\hat{m}_t = g$  and  $\hat{v}_t = g^2$ , so Adam's step is  $-\eta g / (|g| + \epsilon)$ .
- Ridge regression in this page's units:  $L(\theta) = \frac{1}{2N} \|X\theta - y\|^2$  with gradient  $\frac{1}{N} X^\top (X\theta - y)$ , so  $L_\lambda$  is minimised by  $\theta^* = (\frac{1}{N} X^\top X + \lambda I)^{-1} \frac{1}{N} X^\top y$  (the regression page, with  $\lambda$  scaled to match the  $\frac{\lambda}{2}$  here).
- A **scale-invariant** weight  $w$  is one the loss sees only through its direction,  $L(w) = f(w/\|w\|)$ : any weight matrix followed by batch norm or layer norm, and the embeddings of the contrastive page.  $G$  denotes  $\|\nabla L\|$  evaluated at  $\hat{w} = w/\|w\|$ , the gradient norm at unit scale.

Builds on: Momentum, RMSProp and Adam: optimiser updates by hand, Regression gradients: linear, logistic and softmax

## Problems

**1.** Show that SGD on  $L_\lambda$  is the decoupled update  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta g_t$ . difficulty ·  
With no data gradient, write  $\theta_T$  in terms of  $\theta_0$ , and find the number of steps that halves the weights when  $\eta = 0.1$  and  $\lambda = 10^{-4}$ .

**2.** Show that the L2 and decoupled SGD updates on the ridge loss share difficulty ·  
the fixed point  $\theta^*$  of Before you start, and that the fixed point does not depend on  $\eta$ .

**3.** Momentum with an L2 penalty is  $v_t = \beta v_{t-1} + g_t + \lambda \theta_{t-1}$ ,  $\theta_t =$  difficulty ···  
 $\theta_{t-1} - \eta v_t$ ; decoupled is  $v_t = \beta v_{t-1} + g_t$ ,  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta v_t$ . With no data gradient, show that the L2 version satisfies  $\theta_t = (1 + \beta - \eta\lambda)\theta_{t-1} - \beta\theta_{t-2}$ , find its asymptotic shrink factor per step to first order in  $\eta\lambda$ , and compare with the decoupled version.

**4.** Adam with an L2 penalty feeds  $g_t + \lambda \theta_{t-1}$  to the moment estimates. difficulty ··  
For a constant gradient vector  $g$ , and once the moment estimates have caught up with their input (so that  $\hat{m}_t \approx g + \lambda \theta_{t-1}$  and  $\hat{v}_t \approx (g + \lambda \theta_{t-1})^2$ ), write the step. Show that an entry with  $g_i = 0$  shrinks by about  $\eta$  per step whatever  $\lambda$  is, and that an entry with  $|g_i| \gg \lambda|\theta_i|$  receives a decay of about  $\eta\lambda\theta_i/|g_i|$ .

**5.** AdamW with  $\epsilon > 0$ :  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ . Show that a stationary point with gradient  $g = g(\theta)$  satisfies  $\lambda\theta_i(|g_i| + \epsilon) = -g_i$  for every entry. Deduce what happens for a constant gradient  $g$  with  $|g_i| \gg \epsilon$ , and, for a loss whose gradient vanishes at its minimiser, that the stationary point is the ridge solution with penalty coefficient  $\lambda\epsilon$ . difficulty ...

**6.** A scale-invariant weight:  $L(w) = f(w/\|w\|)$ . Show that  $w^\top \nabla L(w) = 0$  and  $\|\nabla L(w)\| = G/\|w\|$ . For decoupled SGD,  $w_t = (1 - \eta\lambda)w_{t-1} - \eta g_t$ , show that  $\|w_t\|^2 = (1 - \eta\lambda)^2\|w_{t-1}\|^2 + \eta^2\|g_t\|^2$ , find the equilibrium norm when  $\|g_t\| = G/\|w_{t-1}\|$  with  $G$  constant, and the resulting effective learning rate  $\eta/\|w\|^2$  on the direction of  $w$ . difficulty ...

**7.** Same layer with no decay ( $\lambda = 0$ ). Show that  $n_t = \|w_t\|^2$  satisfies  $n_t^2 \approx n_{t-1}^2 + 2\eta^2 G^2$ , hence  $\|w_T\| \approx (n_0^2 + 2\eta^2 G^2 T)^{1/4}$ , and describe how the effective learning rate behaves over training. difficulty ..

**8.** Write AdamW as  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta u_t$  with  $u_t$  the normalised Adam direction. Unroll it to a closed form for  $\theta_T$ , show that the coefficients of the  $u_t$  sum to  $1/\lambda$  as  $T \rightarrow \infty$ , and, for a learning-rate schedule  $\eta_t$ , show that the factor multiplying  $\theta_0$  after  $T$  steps is about  $e^{-\lambda \sum_t \eta_t}$ . Evaluate it for  $\eta = 10^{-3}$ ,  $\lambda = 0.1$  and  $T = 100,000$ . difficulty ..

**9.** Decoupled SGD on the quadratic  $L(\theta) = \frac{1}{2}\theta^\top A\theta - b^\top \theta$  with  $A$  symmetric positive definite. Show that  $\theta_t - \theta^* = (I - \eta(A + \lambda I))(\theta_{t-1} - \theta^*)$  with  $\theta^* = (A + \lambda I)^{-1}b$ , give the condition on  $\eta$  for convergence, and for the optimiser page's  $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$  with  $\lambda = 1$  find the range of  $\eta$ , the best  $\eta$ , its worst-case factor, and the condition number with and without the decay. difficulty ..

**10.** The loss is multiplied by a constant  $c > 0$  (a change of units, or a different reduction over the batch), so the gradient becomes  $cg_t$ . Show that SGD with an L2 penalty on  $cL$  with  $(\eta, \lambda)$  is SGD on  $L$  with  $(c\eta, \lambda/c)$ ; that AdamW with  $\epsilon \rightarrow 0$  produces the same iterates for  $cL$  as for  $L$ ; and that Adam with an L2 penalty on  $cL$  with  $\lambda$  equals Adam with an L2 penalty on  $L$  with  $\lambda/c$ . difficulty ..

## Answers

1. SGD with an L2 penalty is  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta g_t$ : the decay and the penalty are the same update; with no gradient,  $\theta_T = (1 - \eta\lambda)^T \theta_0$ , and  $\eta\lambda = 10^{-5}$  halves a weight in about 69,000 steps
2. Both forms have the fixed point  $g(\theta) + \lambda\theta = 0$ , which is the ridge solution  $\theta^* = (\frac{1}{N}X^\top X + \lambda I)^{-1} \frac{1}{N}X^\top y$ , for every  $\eta$
3. L2 + momentum shrinks by about  $1 - \eta\lambda/(1 - \beta)$  per step, ten times the decoupled  $1 - \eta\lambda$  at  $\beta = 0.9$ ; to match a decoupled  $\lambda$ , the L2 coefficient must be  $(1 - \beta)\lambda$
4. Adam + L2 step  $\approx -\eta(g + \lambda\theta)/(|g + \lambda\theta| + \epsilon)$ : a weight with no data gradient moves by  $\eta$  per step towards 0 regardless of  $\lambda$ , while a weight with a large gradient is decayed by only  $\eta\lambda\theta_i/|g_i|$  per step
5. AdamW's stationary points satisfy  $\lambda\theta_i(|g_i| + \epsilon) = -g_i$ ; under a persistent gradient the weight settles at  $|\theta_i| \approx 1/\lambda$  whatever  $|g_i|$ ; where the gradient vanishes at the optimum, the stationary point is the ridge solution with penalty  $\lambda\epsilon$ , essentially unregularised at  $\epsilon = 10^{-8}$
6.  $w^\top \nabla L = 0$  and  $\|\nabla L(w)\| = G/\|w\|$ ;  $\|w_t\|^2 = (1 - \eta\lambda)^2 \|w_{t-1}\|^2 + \eta^2 \|g_t\|^2$ ; equilibrium  $\|w\| \approx (\eta G^2/2\lambda)^{1/4}$  and effective learning rate  $\eta/\|w\|^2 \approx \sqrt{2\eta\lambda}/G$
7.  $\|w_T\| \approx (n_0^2 + 2\eta^2 G^2 T)^{1/4}$ : the norm grows like  $T^{1/4}$  and the effective learning rate on the direction falls like  $1/\sqrt{T}$ , independent of  $\eta$
8.  $\theta_T = (1 - \eta\lambda)^T \theta_0 - \eta \sum_t (1 - \eta\lambda)^{T-t} u_t$ ; the update weights sum to  $1/\lambda$ , so  $\theta$  is  $1/\lambda$  times an EMA of the Adam directions with time constant  $1/(\eta\lambda)$  steps;  $\theta_0$  is multiplied by about  $e^{-\lambda \sum_t \eta_t}$ , which is  $e^{-10}$  in the example
9.  $\theta_t - \theta^* = (I - \eta(A + \lambda I))(\theta_{t-1} - \theta^*)$ ,  $\theta^* = (A + \lambda I)^{-1}b$ ; converges iff  $0 < \eta < 2/(\lambda_{\max} + \lambda)$ ; for this  $A$  and  $\lambda = 1$ :  $\eta < \frac{2}{5}$ , best  $\eta = \frac{1}{4}$  with factor  $\frac{1}{4}$  (was  $\frac{1}{3}$ ), condition number  $\frac{5}{3}$  (was 2)
10. SGD + L2:  $(cL, \eta, \lambda) \equiv (L, c\eta, \lambda/c)$ ; AdamW is invariant to  $c$  (up to  $\epsilon$ ); Adam + L2:  $(cL, \lambda) \equiv (L, \lambda/c)$  at the same  $\eta$

## Worked solutions

### Problem 1

Show that SGD on  $L_\lambda$  is the decoupled update  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta g_t$ . With no data gradient, write  $\theta_T$  in terms of  $\theta_0$ , and find the number of steps that halves the weights when  $\eta = 0.1$  and  $\lambda = 10^{-4}$ .

|                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                             |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\nabla\left(\frac{\lambda}{2}\ \theta\ ^2\right) = \lambda\theta.$                                                                                                                                                                                                                                   | $\nabla_\theta(\theta^\top\theta) = 2\theta$ (the matrix-calculus page); the $\frac{1}{2}$ cancels the 2.                                                                                                                                                                                                                                                                                                   |
| $\theta_t = \theta_{t-1} - \eta(g_t + \lambda\theta_{t-1}) = (1 - \eta\lambda)\theta_{t-1} - \eta g_t.$                                                                                                                                                                                               | SGD on $L_\lambda$ uses the gradient of step 1 added to $g_t$ ; collect the $\theta_{t-1}$ terms.                                                                                                                                                                                                                                                                                                           |
| With $g_t = 0$ : $\theta_T = (1 - \eta\lambda)^T\theta_0.$                                                                                                                                                                                                                                            | Step 2 applied $T$ times is multiplication by the same scalar $T$ times.                                                                                                                                                                                                                                                                                                                                    |
| $(1 - \eta\lambda)^T = \frac{1}{2}$ gives $T = \frac{\ln 2}{-\ln(1 - \eta\lambda)} \approx \frac{\ln 2}{\eta\lambda}.$                                                                                                                                                                                | Take logs; $-\ln(1 - x) = x + \frac{x^2}{2} + \dots \approx x$ for small $x$ .                                                                                                                                                                                                                                                                                                                              |
| $\eta\lambda = 10^{-5}$ , so $T \approx 0.693 \times 10^5 \approx 69,300$ steps.                                                                                                                                                                                                                      | $\ln 2 \approx 0.693$ ; the exact value from step 4 is 69,315.                                                                                                                                                                                                                                                                                                                                              |
| <b>SGD with an L2 penalty is <math>\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta g_t</math>: the decay and the penalty are the same update; with no gradient, <math>\theta_T = (1 - \eta\lambda)^T\theta_0</math>, and <math>\eta\lambda = 10^{-5}</math> halves a weight in about 69,000 steps</b> | Under plain SGD the two conventions agree exactly, step for step, with the same $\lambda$ ; everything after this problem is about optimisers where they do not. The shrink per step is $\eta\lambda$ , a product: halving the learning rate halves the decay, and the characteristic time $1/(\eta\lambda)$ steps (here 100,000) is the horizon over which a weight forgets its initial value (Problem 8). |

## Problem 2

Show that the L2 and decoupled SGD updates on the ridge loss share the fixed point  $\theta^*$  of Before you start, and that the fixed point does not depend on  $\eta$ .

A fixed point of  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta g(\theta_{t-1})$  satisfies  $\theta = (1 - \eta\lambda)\theta - \eta g(\theta)$ .

Fixed point: the update returns the same  $\theta$ .

$\eta\lambda\theta + \eta g(\theta) = 0$ , so  $g(\theta) + \lambda\theta = 0$ .

Rearrange and divide by  $\eta > 0$ ;  $\eta$  has dropped out.

$\frac{1}{N}X^\top(X\theta - y) + \lambda\theta = 0$ , so  $(\frac{1}{N}X^\top X + \lambda I)\theta = \frac{1}{N}X^\top y$ .

The ridge gradient; collect the  $\theta$  terms.

$\theta = \theta^* = (\frac{1}{N}X^\top X + \lambda I)^{-1} \frac{1}{N}X^\top y$ .

$\frac{1}{N}X^\top X + \lambda I$  is positive definite for  $\lambda > 0$  (the regression page), so it is invertible.

The L2 update is the same map (Problem 1), so it has the same fixed point.

Problem 1, step 2.

**Both forms have the fixed point**

$g(\theta) + \lambda\theta = 0$ , **which is the ridge**

**solution**  $\theta^* = (\frac{1}{N}X^\top X + \lambda I)^{-1} \frac{1}{N}X^\top y$ ,

**for every  $\eta$**

A fixed point of SGD on  $L_\lambda$  is a stationary point of  $L_\lambda$ , whatever the learning rate;  $\eta$  decides how fast and whether the iteration gets there (Problem 9), not where it goes. This is the property AdamW gives up (Problem 5).

## Problem 3

Momentum with an L2 penalty is  $v_t = \beta v_{t-1} + g_t + \lambda\theta_{t-1}$ ,  $\theta_t = \theta_{t-1} - \eta v_t$ ; decoupled is  $v_t = \beta v_{t-1} + g_t$ ,  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta v_t$ . With no data gradient, show that the L2 version satisfies  $\theta_t = (1 + \beta - \eta\lambda)\theta_{t-1} - \beta\theta_{t-2}$ , find its asymptotic shrink factor per step to first order in  $\eta\lambda$ , and compare with the decoupled version.

$v_t = (\theta_{t-1} - \theta_t)/\eta$  for every  $t$ .

Rearrange  $\theta_t = \theta_{t-1} - \eta v_t$ ; this holds for both versions.

|                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| L2 with $g_t = 0$ : $\theta_t = \theta_{t-1} - \eta(\beta v_{t-1} + \lambda \theta_{t-1}) = \theta_{t-1} - \beta(\theta_{t-2} - \theta_{t-1}) - \eta \lambda \theta_{t-1}$ .                                                                                                          | Substitute the recurrence for $v_t$ , then step 1 for $v_{t-1}$ .                                                                                                                                                                                                                                                                                                                                                |
| $\theta_t = (1 + \beta - \eta \lambda) \theta_{t-1} - \beta \theta_{t-2}$ .                                                                                                                                                                                                           | Collect terms. A second-order linear recurrence with constant coefficients.                                                                                                                                                                                                                                                                                                                                      |
| Solutions are $r^t$ with $r^2 - (1 + \beta - \eta \lambda)r + \beta = 0$ .                                                                                                                                                                                                            | Substitute $\theta_t = r^t$ and divide by $r^{t-2}$ . The general solution is a combination of the two roots' powers, and the larger root dominates.                                                                                                                                                                                                                                                             |
| Write $r = 1 - \varepsilon$ : $(1 - \varepsilon)^2 - (1 + \beta - \eta \lambda)(1 - \varepsilon) + \beta = \eta \lambda - (1 - \beta)\varepsilon + \varepsilon^2 - \eta \lambda \varepsilon = 0$ .                                                                                    | Expand and cancel: the constant terms give $1 - (1 + \beta) + \beta = 0$ , the terms in $\varepsilon$ give $-2 + (1 + \beta) = -(1 - \beta)$ , and $\eta \lambda$ survives from the middle product.                                                                                                                                                                                                              |
| To first order, $\varepsilon = \frac{\eta \lambda}{1 - \beta}$ , so $r \approx 1 - \frac{\eta \lambda}{1 - \beta}$ .                                                                                                                                                                  | Drop the quadratic terms $\varepsilon^2$ and $\eta \lambda \varepsilon$ ; the next term is of order $\varepsilon^2/(1 - \beta)$ .                                                                                                                                                                                                                                                                                |
| Decoupled with $g_t = 0$ : $v_t = 0$ throughout, so $\theta_t = (1 - \eta \lambda) \theta_{t-1}$ exactly.                                                                                                                                                                             | $v_0 = 0$ and nothing is added to it.                                                                                                                                                                                                                                                                                                                                                                            |
| <b>L2 + momentum shrinks by about <math>1 - \eta \lambda / (1 - \beta)</math> per step, ten times the decoupled <math>1 - \eta \lambda</math> at <math>\beta = 0.9</math>; to match a decoupled <math>\lambda</math>, the L2 coefficient must be <math>(1 - \beta) \lambda</math></b> | The penalty gradient $\lambda \theta$ enters the momentum buffer and is applied $1/(1 - \beta)$ times over, exactly as the optimiser page's Problem 4 found for any steady gradient. PyTorch's SGD with momentum=0.9, weight_decay= $\lambda$ therefore decays weights ten times faster than the number suggests; a $\lambda$ tuned there and moved to an optimiser with decoupled decay is ten times too small. |

### Problem 4

Adam with an L2 penalty feeds  $g_t + \lambda \theta_{t-1}$  to the moment estimates. For a constant gradient vector  $g$ , and once the moment estimates have caught up with their input (so that  $\hat{m}_t \approx g +$

$\lambda\theta_{t-1}$  and  $\hat{v}_t \approx (g + \lambda\theta_{t-1})^2$ , write the step. Show that an entry with  $g_i = 0$  shrinks by about  $\eta$  per step whatever  $\lambda$  is, and that an entry with  $|g_i| \gg \lambda|\theta_i|$  receives a decay of about  $\eta\lambda\theta_i/|g_i|$ .

Step =  $-\eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \approx -\eta \frac{g + \lambda\theta_{t-1}}{|g + \lambda\theta_{t-1}| + \epsilon}$ ,  
entry by entry.

Adam's update with the converged estimates;  $\sqrt{x^2} = |x|$ .

Entry with  $g_i = 0$ :  $\text{step}_i \approx -\eta \frac{\lambda\theta_i}{|\lambda\theta_i| + \epsilon} \approx -\eta \text{sign}(\theta_i)$ .

$\lambda$  cancels between numerator and denominator when  $\lambda|\theta_i| \gg \epsilon$ .

Entry with  $|g_i| \gg \lambda|\theta_i|$ :  $\frac{g_i + \lambda\theta_i}{|g_i + \lambda\theta_i|} \approx \frac{g_i + \lambda\theta_i}{|g_i|}$ , so  $\text{step}_i \approx -\eta \frac{g_i}{|g_i|} - \eta \frac{\lambda\theta_i}{|g_i|}$ .

The denominator is dominated by  $|g_i|$  (and  $\epsilon$  is negligible next to it); split the numerator. The first term is the plain Adam step.

**Adam + L2 step  $\approx -\eta (g + \lambda\theta)/(|g + \lambda\theta| + \epsilon)$ : a weight with no data gradient moves by  $\eta$  per step towards 0 regardless of  $\lambda$ , while a weight with a large gradient is decayed by only  $\eta\lambda\theta_i/|g_i|$  per step**

The decay is normalised by the same  $\sqrt{\hat{v}_t}$  as the data gradient, so it is strongest exactly where the gradient history is smallest and weakest where it is largest: the opposite of a uniform shrink, and a  $\lambda$  that is invisible on busy weights and irrelevant on idle ones. The optimiser page's Problem 10 is the first step of this; here it is the steady state, approximate because the estimates trail the slowly moving  $\lambda\theta$ .

## Problem 5

AdamW with  $\epsilon > 0$ :  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$ . Show that a stationary point with gradient  $g = g(\theta)$  satisfies  $\lambda\theta_i(|g_i| + \epsilon) = -g_i$  for every entry. Deduce what happens for a constant gradient  $g$  with  $|g_i| \gg \epsilon$ , and, for a loss whose gradient vanishes at its minimiser, that the stationary point is the ridge solution with penalty coefficient  $\lambda\epsilon$ .

At a stationary point the gradient is a constant  $g$ , so  $\hat{m}_t = g$  and  $\hat{v}_t = g^2$  once the averages settle.

The optimiser page, Problem 8: constant input gives exact bias-corrected moments.

|                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\theta = (1 - \eta\lambda)\theta - \eta \frac{g}{ g  + \epsilon}, \text{ so } \lambda\theta_i = -\frac{g_i}{ g_i  + \epsilon}.$                                                                                                                                                                                                                                                                                        | Fixed-point condition; cancel $\eta$ .                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| $\lambda\theta_i( g_i  + \epsilon) = -g_i.$                                                                                                                                                                                                                                                                                                                                                                             | Multiply out.                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Constant $g$ with $ g_i  \gg \epsilon$ : $\lambda\theta_i \approx -\text{sign}(g_i)$ , so $ \theta_i  \approx 1/\lambda$ .                                                                                                                                                                                                                                                                                              | $g_i/( g_i  + \epsilon) \approx \text{sign}(g_i)$ . The size of the gradient has no effect on where the weight settles.                                                                                                                                                                                                                                                                                                                                                                   |
| Gradient that vanishes at the minimiser: near the stationary point $ g_i $ is small, and if $ g_i  \ll \epsilon$ then $\lambda\theta_i\epsilon \approx -g_i$ , that is $g(\theta) + \lambda\epsilon\theta \approx 0$ .                                                                                                                                                                                                  | Step 3 with $ g_i  + \epsilon \approx \epsilon$ . Consistency: it requires $ g_i  = \lambda\epsilon \theta_i  \ll \epsilon$ , that is $\lambda \theta_i  \ll 1$ , which holds for the usual $\lambda = 0.01$ and weights of order 1.                                                                                                                                                                                                                                                      |
| On the ridge loss, $g(\theta) + \lambda\epsilon\theta = 0$ is Problem 2's fixed-point equation with $\lambda\epsilon$ in place of $\lambda$ .                                                                                                                                                                                                                                                                           | Compare with Problem 2, step 2.                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| <b>AdamW's stationary points satisfy <math>\lambda\theta_i( g_i  + \epsilon) = -g_i</math>; under a persistent gradient the weight settles at <math> \theta_i  \approx 1/\lambda</math> whatever <math> g_i </math>; where the gradient vanishes at the optimum, the stationary point is the ridge solution with penalty <math>\lambda\epsilon</math>, essentially unregularised at <math>\epsilon = 10^{-8}</math></b> | AdamW is not gradient descent on any penalised loss: its decay regularises the <i>trajectory</i> (a weight needs a persistent gradient to stay large, and cannot exceed $1/\lambda$ , 100 at the default $\lambda$ ) but not the <i>solution</i> of a problem it can solve exactly. Adam with the L2 penalty, by contrast, is a method for $L_\lambda$ and does converge to the ridge solution with coefficient $\lambda$ , at the cost of Problem 4's normalised decay on the way there. |

## Problem 6

A scale-invariant weight:  $L(w) = f(w/\|w\|)$ . Show that  $w^\top \nabla L(w) = 0$  and  $\|\nabla L(w)\| = G/\|w\|$ . For decoupled SGD,  $w_t = (1 - \eta\lambda)w_{t-1} - \eta g_t$ , show that  $\|w_t\|^2 = (1 - \eta\lambda)^2 \|w_{t-1}\|^2 + \eta^2 \|g_t\|^2$ ,

find the equilibrium norm when  $\|g_t\| = G/\|w_{t-1}\|$  with  $G$  constant, and the resulting effective learning rate  $\eta/\|w\|^2$  on the direction of  $w$ .

$$L(\alpha w) = L(w) \text{ for } \alpha > 0, \text{ so}$$

$$\left. \frac{d}{d\alpha} L(\alpha w) \right|_{\alpha=1} = w^\top \nabla L(w) = 0.$$

$\alpha w/\|\alpha w\| = w/\|w\|$ ; chain rule along the ray, as on the contrastive page.

$$\nabla L(w) = \frac{1}{\|w\|} (I - \hat{w}\hat{w}^\top) \nabla f(\hat{w}).$$

Chain rule through  $\hat{w} = w/\|w\|$  with the Jacobian  $\frac{1}{\|w\|} (I - \hat{w}\hat{w}^\top)$  (the Jacobians page).

At  $\|w\| = 1$  the same formula gives  
 $\nabla L(\hat{w}) = (I - \hat{w}\hat{w}^\top) \nabla f(\hat{w})$ , so  $\nabla L(w) = \nabla L(\hat{w})/\|w\|$  and  $\|\nabla L(w)\| = G/\|w\|$ .

Step 2 at  $w = \hat{w}$  and again at  $w$ : the projection is the same, only the  $1/\|w\|$  differs.  $G = \|\nabla L(\hat{w})\|$  by definition.

$$\|w_t\|^2 = (1 - \eta\lambda)^2 \|w_{t-1}\|^2 - 2\eta(1 - \eta\lambda) w_{t-1}^\top g_t + \eta^2 \|g_t\|^2 = (1 - \eta\lambda)^2 \|w_{t-1}\|^2 + \eta^2 \|g_t\|^2.$$

Expand the squared norm; the cross term vanishes because  $g_t = \nabla L(w_{t-1})$  is orthogonal to  $w_{t-1}$  (step 1). Pythagoras: the decay shrinks  $w$  along itself and the gradient step moves it sideways.

With  $n = \|w\|^2$  at equilibrium and  
 $\|g_t\|^2 = G^2/n$ :  $n = (1 - \eta\lambda)^2 n + \eta^2 G^2/n$ ,  
 so  $n^2(1 - (1 - \eta\lambda)^2) = \eta^2 G^2$ .

Set  $\|w_t\|^2 = \|w_{t-1}\|^2 = n$  and multiply through by  $n$ .

$$1 - (1 - \eta\lambda)^2 = 2\eta\lambda - \eta^2\lambda^2 \approx 2\eta\lambda, \text{ so}$$

$$n^2 \approx \frac{\eta G^2}{2\lambda} \text{ and } \|w\| \approx \left( \frac{\eta G^2}{2\lambda} \right)^{1/4}.$$

$\eta\lambda \ll 1$ ; divide and take the fourth root.

$$\frac{\eta}{\|w\|^2} = \frac{\eta}{n} \approx \eta \sqrt{\frac{2\lambda}{\eta G^2}} = \frac{\sqrt{2\eta\lambda}}{G}.$$

Substitute step 6.

$w^\top \nabla L = 0$  **and**  $\|\nabla L(w)\| = G/\|w\|$ ;  
 $\|w_t\|^2 = (1 - \eta\lambda)^2 \|w_{t-1}\|^2 + \eta^2 \|g_t\|^2$ ;  
**equilibrium**  $\|w\| \approx (\eta G^2/2\lambda)^{1/4}$  **and**  
**effective learning rate**  $\eta/\|w\|^2 \approx \sqrt{2\eta\lambda}/G$

For a weight the loss cannot see the size of, decay does not regularise the function at all: it sets the norm, and through the norm the angular step  $\eta\|g\|/\|w\| = \eta G/\|w\|^2$  per iteration. That step is  $\sqrt{2\eta\lambda}/G$ : it depends on  $\eta$  and  $\lambda$  only through their product, and only as a square root, so dividing  $\eta$  by 10 at the end of a schedule slows the rotation of a normalised layer by only  $\sqrt{10}$ , and the same decay with no normalisation layer would be doing something else entirely.

### Problem 7

Same layer with no decay ( $\lambda = 0$ ). Show that  $n_t = \|w_t\|^2$  satisfies  $n_t^2 \approx n_{t-1}^2 + 2\eta^2 G^2$ , hence  $\|w_T\| \approx (n_0^2 + 2\eta^2 G^2 T)^{1/4}$ , and describe how the effective learning rate behaves over training.

$$n_t = n_{t-1} + \eta^2 G^2 / n_{t-1}.$$

Problem 6, step 4 with  $\lambda = 0$  and  $\|g_t\|^2 = G^2/n_{t-1}$ .

$$n_t^2 = n_{t-1}^2 + 2\eta^2 G^2 + \eta^4 G^4 / n_{t-1}^2 \approx n_{t-1}^2 + 2\eta^2 G^2.$$

Square step 1; the last term is smaller than the middle one by the factor  $\eta^2 G^2 / (2n_{t-1}^2)$ , which is tiny once  $n$  is of order 1.

$$n_T^2 \approx n_0^2 + 2\eta^2 G^2 T, \text{ so } \|w_T\| = n_T^{1/2} \approx (n_0^2 + 2\eta^2 G^2 T)^{1/4}.$$

Sum step 2 over  $T$  steps.

The effective learning rate  $\eta/n_T \approx \eta/\sqrt{2\eta^2 G^2 T} = 1/(G\sqrt{2T})$  for large  $T$ .

Substitute step 3 and drop  $n_0^2$ .

$\|w_T\| \approx (n_0^2 + 2\eta^2 G^2 T)^{1/4}$ : **the norm grows like  $T^{1/4}$  and the effective learning rate on the direction falls like  $1/\sqrt{T}$ , independent of  $\eta$**

With no decay, a normalised layer applies its own  $1/\sqrt{T}$  learning-rate schedule, because every gradient step lengthens  $w$  (step 1: the orthogonal step can only add to the norm) and a longer  $w$  turns more slowly. Weight decay (Problem 6) is what stops this: it pulls the norm back to an equilibrium and holds the angular step at  $\sqrt{2\eta\lambda}/G$  instead of letting it decay to zero. This is also why removing decay from such a layer makes training appear to stall late on rather than diverge.

### Problem 8

Write AdamW as  $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta u_t$  with  $u_t$  the normalised Adam direction. Unroll it to a closed form for  $\theta_T$ , show that the coefficients of the  $u_t$  sum to  $1/\lambda$  as  $T \rightarrow \infty$ , and, for a learning-rate schedule  $\eta_t$ , show that the factor multiplying  $\theta_0$  after  $T$  steps is about  $e^{-\lambda \sum_t \eta_t}$ . Evaluate it for  $\eta = 10^{-3}$ ,  $\lambda = 0.1$  and  $T = 100,000$ .

$$\theta_1 = (1 - \eta\lambda)\theta_0 - \eta u_1, \theta_2 = (1 - \eta\lambda)^2\theta_0 - \eta(1 - \eta\lambda)u_1 - \eta u_2.$$

Apply the update twice; each earlier term picks up one more factor of  $1 - \eta\lambda$ .

$$\theta_T = (1 - \eta\lambda)^T \theta_0 - \eta \sum_{t=1}^T (1 - \eta\lambda)^{T-t} u_t.$$

Induction on  $T$ : multiplying by  $1 - \eta\lambda$  raises every exponent by one and the new update enters with exponent 0.

$$\eta \sum_{t=1}^T (1 - \eta\lambda)^{T-t} = \eta \frac{1 - (1 - \eta\lambda)^T}{\eta\lambda} \rightarrow \frac{1}{\lambda}.$$

Geometric series with ratio  $1 - \eta\lambda < 1$ ; the  $\eta$  cancels.

So  $\theta_T \approx -\frac{1}{\lambda} \cdot$  (weighted average of the  $u_t$ )  
with weights  $\eta\lambda(1 - \eta\lambda)^{T-t}$ , an  
exponential moving average with time  
constant  $1/(\eta\lambda)$  steps.

The weights of step 3 normalised to sum to 1; they halve every  $\ln 2/(\eta\lambda)$  steps (Problem 1).

With a schedule,  $\theta_0$ 's factor is  $\prod_t (1 - \eta_t \lambda)$ , and  $\ln \prod_t (1 - \eta_t \lambda) = \sum_t \ln(1 - \eta_t \lambda) \approx -\lambda \sum_t \eta_t$ .

$\ln(1 - x) \approx -x$  for small  $x$ ; the error is of order  $\lambda^2 \sum_t \eta_t^2$ .

$\lambda \sum_t \eta_t = 0.1 \times 10^{-3} \times 10^5 = 10$ , so the factor is  $e^{-10} \approx 4.5 \times 10^{-5}$ .

Constant  $\eta$ :  $\sum_t \eta_t = \eta T$ .

$\theta_T = (1 - \eta \lambda)^T \theta_0 - \eta \sum_t (1 - \eta \lambda)^{T-t} u_t$ ;  
**the update weights sum to  $1/\lambda$ , so  $\theta$  is  $1/\lambda$  times an EMA of the Adam directions with time constant  $1/(\eta \lambda)$  steps;  $\theta_0$  is multiplied by about  $e^{-\lambda \sum_t \eta_t}$ , which is  $e^{-10}$  in the example**

Each Adam direction has entries of size about 1, so  $|\theta_i|$  cannot exceed  $1/\lambda$  (Problem 5 again) and a weight reflects only the last  $1/(\eta \lambda)$  steps of updates: 10,000 steps in the example, against a run of 100,000. The decay in PyTorch's AdamW is multiplied by the scheduled  $\eta_t$ , so the EMA horizon stretches as the learning rate decays and the total forgetting is set by the area under the schedule, not by  $\lambda$  alone.

## Problem 9

Decoupled SGD on the quadratic  $L(\theta) = \frac{1}{2} \theta^\top A \theta - b^\top \theta$  with  $A$  symmetric positive definite. Show that  $\theta_t - \theta^* = (I - \eta(A + \lambda I))(\theta_{t-1} - \theta^*)$  with  $\theta^* = (A + \lambda I)^{-1} b$ , give the condition on  $\eta$  for convergence, and for the optimiser page's  $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$  with  $\lambda = 1$  find the range of  $\eta$ , the best  $\eta$ , its worst-case factor, and the condition number with and without the decay.

$$\theta_t = (1 - \eta \lambda) \theta_{t-1} - \eta (A \theta_{t-1} - b) = (I - \eta(A + \lambda I)) \theta_{t-1} + \eta b.$$

The gradient of the quadratic is  $A\theta - b$ ; collect the  $\theta_{t-1}$  terms.

$$\theta^* = (A + \lambda I)^{-1} b \text{ satisfies } \theta^* = (I - \eta(A + \lambda I)) \theta^* + \eta b.$$

$$(A + \lambda I) \theta^* = b, \text{ so } \eta(A + \lambda I) \theta^* = \eta b.$$

$$\text{Subtracting: } \theta_t - \theta^* = (I - \eta(A + \lambda I))(\theta_{t-1} - \theta^*).$$

Step 1 minus step 2; the  $\eta b$  terms cancel.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                                                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $A + \lambda I$ has eigenvalues $\lambda_i + \lambda$ with the eigenvectors of $A$ , so convergence from every start needs $ 1 - \eta(\lambda_i + \lambda)  < 1$ for all $i$ : $0 < \eta < \frac{2}{\lambda_{\max} + \lambda}$ .                                                                                                                                                                                                                | The optimiser page's Problem 2 with $A + \lambda I$ in place of $A$ . Adding $\lambda I$ shifts every eigenvalue by $\lambda$ and leaves the eigenvectors alone.                                                                                                                                                                                                                                     |
| For the given $A$ , eigenvalues 2 and 4 become 3 and 5: $0 < \eta < \frac{2}{5}$ , best $\eta = \frac{2}{3+5} = \frac{1}{4}$ , worst-case factor $\frac{5-3}{5+3} = \frac{1}{4}$ .                                                                                                                                                                                                                                                              | The optimiser page's formulas $2/(\lambda_{\min} + \lambda_{\max})$ and $(\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} + \lambda_{\min})$ applied to 3 and 5.                                                                                                                                                                                                                                    |
| Condition number $\frac{\lambda_{\max} + \lambda}{\lambda_{\min} + \lambda} = \frac{5}{3}$ , against $\frac{4}{2} = 2$ without decay.                                                                                                                                                                                                                                                                                                           | The ratio of the largest to the smallest eigenvalue of the matrix the iteration sees.                                                                                                                                                                                                                                                                                                                |
| $\theta_t - \theta^* = (I - \eta(A + \lambda I))(\theta_{t-1} - \theta^*)$ ,<br>$\theta^* = (A + \lambda I)^{-1}b$ ; <b>converges iff</b> $0 < \eta < 2/(\lambda_{\max} + \lambda)$ ; <b>for this <math>A</math> and <math>\lambda = 1</math>:</b><br>$\eta < \frac{2}{5}$ , <b>best <math>\eta = \frac{1}{4}</math> with factor <math>\frac{1}{4}</math> (was <math>\frac{1}{3}</math>), condition number <math>\frac{5}{3}</math> (was 2)</b> | Decay moves the minimiser from $A^{-1}b$ to $(A + \lambda I)^{-1}b$ , which is the bias it pays for, and makes the problem better conditioned, which is why it speeds up convergence along the flat directions most: an eigenvalue $\lambda_i \ll \lambda$ is effectively replaced by $\lambda$ . The stability limit tightens slightly, from $2/\lambda_{\max}$ to $2/(\lambda_{\max} + \lambda)$ . |

### Problem 10

The loss is multiplied by a constant  $c > 0$  (a change of units, or a different reduction over the batch), so the gradient becomes  $cg_t$ . Show that SGD with an L2 penalty on  $cL$  with  $(\eta, \lambda)$  is SGD on  $L$  with  $(c\eta, \lambda/c)$ ; that AdamW with  $\epsilon \rightarrow 0$  produces the same iterates for  $cL$  as for  $L$ ; and that Adam with an L2 penalty on  $cL$  with  $\lambda$  equals Adam with an L2 penalty on  $L$  with  $\lambda/c$ .

|                                                                                                                                               |                                                                                               |
|-----------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| SGD + L2 on $cL$ : $\theta_t = \theta_{t-1} - \eta(cg_t + \lambda\theta_{t-1}) = \theta_{t-1} - c\eta(g_t + \frac{\lambda}{c}\theta_{t-1})$ . | Factor $c$ out of the bracket.                                                                |
| That is SGD + L2 on $L$ with learning rate $c\eta$ and coefficient $\lambda/c$ .                                                              | Compare with Problem 1, step 2. The regularisation relative to the data has weakened by $c$ . |

|                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AdamW on $cL$ : $\hat{m}_t \rightarrow c\hat{m}_t$ and $\hat{v}_t \rightarrow c^2\hat{v}_t$ , so $\hat{m}_t/\sqrt{\hat{v}_t}$ is unchanged.                                                                                                         | Both moment estimates are built from $cg_t$ ; the first is linear in it and the second quadratic, and $\sqrt{c^2\hat{v}_t} = c\sqrt{\hat{v}_t}$ . With $\epsilon \rightarrow 0$ the ratio cancels $c$ exactly.                                                                                                                                                                                                                       |
| The decay $(1 - \eta\lambda)\theta_{t-1}$ does not involve the gradient, so AdamW's iterates are identical.                                                                                                                                         | The decoupled form touches only $\theta$ and $\eta\lambda$ .                                                                                                                                                                                                                                                                                                                                                                         |
| Adam + L2 on $cL$ feeds $cg_t + \lambda\theta_{t-1} = c(g_t + \frac{\lambda}{c}\theta_{t-1})$ to the moments; by step 3 the factor $c$ cancels in the normalised step.                                                                              | Factor out $c$ inside the input to the moment estimates, then apply the scale invariance.                                                                                                                                                                                                                                                                                                                                            |
| <b>SGD + L2:</b> $(cL, \eta, \lambda) \equiv (L, c\eta, \lambda/c)$ ;<br><b>AdamW is invariant to <math>c</math> (up to <math>\epsilon</math>);</b><br><b>Adam + L2:</b> $(cL, \lambda) \equiv (L, \lambda/c)$ <b>at the same <math>\eta</math></b> | Two of the three change meaning when the loss is rescaled, and only AdamW's $\lambda$ is a property of the optimiser alone. Switching a loss from a sum over the batch to a mean is $c = 1/B$ ; under Adam with <code>weight_decay</code> it multiplies the effective regularisation by $B$ , while the Adam step itself does not change, which is a quiet way to make a $\lambda$ that worked at one batch size useless at another. |

## Where this goes wrong

### 1. Carrying $\lambda$ from SGD to SGD with momentum

Adding `momentum=0.9` to an SGD run looks like changing the direction, not the regularisation.

|                                                                                          |                          |
|------------------------------------------------------------------------------------------|--------------------------|
| SGD with <code>weight_decay</code> = $\lambda$ shrinks weights by $\eta\lambda$ per step | Right so far: Problem 1. |
|------------------------------------------------------------------------------------------|--------------------------|

“Momentum averages the gradients; the decay is a separate term and stays  $\eta\lambda$ .”

The shortcut that causes the mistake: in PyTorch's SGD the penalty gradient  $\lambda\theta$  is added to the gradient *before* the momentum buffer, so it is accumulated like everything else.

With momentum  $\beta$  the weights still shrink by  $\eta\lambda$  per step ✗

They shrink by about  $\eta\lambda/(1 - \beta)$  (Problem 3), ten times more at  $\beta = 0.9$ . The run trains, with weights ten times more strongly decayed than intended; the  $\lambda$  that reproduces the old behaviour is  $(1 - \beta)\lambda$ , and a decoupled implementation would have needed no change.

## 2. Decay on a normalised layer taken to shrink the layer's output

Weight decay is introduced as shrinking the weights, and a smaller weight matrix sounds like a smaller, more regular function.

$L(w) = f(w/\|w\|)$  for a weight matrix followed by a normalisation layer

Right so far: the layer's output is unchanged by the scale of  $w$ .

“Decay pulls  $\|w\|$  down, so the layer's output gets smaller and the model is regularised.”

The assumption that causes the mistake: the output is invariant to  $\|w\|$  (step 1), so it cannot get smaller.

Smaller  $\|w\|$  means smaller activations and a simpler function ✗

Nothing downstream changes with  $\|w\|$ . What decay sets on such a layer is the equilibrium norm  $(\eta G^2/2\lambda)^{1/4}$  and through it the effective learning rate  $\sqrt{2\eta\lambda}/G$  on the direction of  $w$  (Problem 6); raising  $\lambda$  there makes the layer *learn faster*, and removing decay makes its learning rate decay like  $1/\sqrt{T}$  (Problem 7). Decay on the parameters that the normalisation does not absorb, the gains and biases and the final layer, is where a function-space effect lives.

### 3. Expecting AdamW to converge to the ridge solution

AdamW is described as "Adam with weight decay", and weight decay is described as L2 regularisation.

|                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| The ridge loss $L_\lambda$ has minimiser $\theta^* = (\frac{1}{N}X^\top X + \lambda I)^{-1} \frac{1}{N}X^\top y$ | Right so far: Problem 2.                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| "AdamW minimises $L + \frac{\lambda}{2}\ \theta\ ^2$ , so it converges to $\theta^*$ ."                          | The assumption that causes the mistake: the decoupled decay is not the gradient of a term in the loss once the data gradient is normalised by $\sqrt{\hat{v}_t}$ , so there is no penalised loss whose stationary points AdamW finds.                                                                                                                                                                                                                                                 |
| AdamW with <code>weight_decay=<math>\lambda</math></code> converges to $\theta^*$ ✗                              | Its stationary points satisfy $\lambda\theta_i( g_i  + \epsilon) = -g_i$ (Problem 5): on the ridge loss, where the gradient vanishes at the optimum, that is the ridge solution with coefficient $\lambda\epsilon$ , about $10^{-10}$ at the defaults, so AdamW converges to the unregularised least-squares solution. Adam with the L2 penalty does converge to $\theta^*$ . The decay in AdamW bounds and forgets the trajectory (Problems 5 and 8); it does not pick the solution. |

### 4. Halving the learning rate with the decay held fixed

The decay coefficient is called  $\lambda$  and the learning rate  $\eta$ , and tuning one is not supposed to touch the other.

|                                                              |                                                                                                                                                                                |
|--------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AdamW: $\theta_t = (1 - \eta\lambda)\theta_{t-1} - \eta u_t$ | Right so far.                                                                                                                                                                  |
| "The decay is $\lambda$ , independent of the learning rate." | The slip that causes the mistake: the shrink per step is the product $\eta\lambda$ (Problem 1), and in PyTorch's AdamW the decay is multiplied by the scheduled learning rate. |

Halving  $\eta$  leaves the weight decay unchanged ✗

It halves the decay per step and doubles the EMA time constant  $1/(\eta\lambda)$  (Problem 8): the weights remember twice as many updates and  $\theta_0$  is forgotten half as fast, and on a normalised layer the effective learning rate  $\sqrt{2\eta\lambda}/G$  drops by  $\sqrt{2}$  rather than 2 (Problem 6). A learning-rate sweep with  $\lambda$  fixed is also a sweep over the decay; to vary one alone, hold  $\eta\lambda$  fixed.

## 5. Rescaling the loss under Adam with an L2 penalty

Adam does not care about the scale of the loss, and the penalty is just another term in it.

Adam's step  $\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$  is unchanged when the gradient is multiplied by  $c$

Right so far: Problem 10, step 3.

“So multiplying the loss by  $c$  (switching from a mean to a sum over the batch) changes nothing under Adam.”

The assumption that causes the mistake: the penalty gradient  $\lambda\theta$  is added to the *data* gradient before normalisation, and it was not multiplied by  $c$ .

Adam with `weight_decay`= $\lambda$  gives the same iterates for  $cL$  as for  $L$  ✗

It gives the iterates for  $L$  with  $\lambda/c$  (Problem 10, step 5): a sum over a batch of 256 instead of a mean divides the effective regularisation by 256, while the loss curve, in its new units, looks the same. AdamW is the version for which the claim holds, which is one reason its  $\lambda$  transfers between setups and Adam's does not.