

---

EECS 182      Deep Neural Networks

Fall 2025      Anant Sahai and Gireeja Ranade

# Homework 1

---

**This homework is due on Friday, Sep 12, 2025, at 10:59PM.**

## 1. Why Learning Rates Cannot be Too Big

To understand the role of the learning rate, it is useful to understand it in the context of the simplest possible problem first.

Suppose that we want to solve the scalar equation

$$\sigma w = y \tag{1}$$

where we know that  $\sigma > 0$ . We proceed with an initial condition  $w_0 = 0$  by using gradient descent to minimize the squared loss

$$L(w) = (y - \sigma w)^2 \tag{2}$$

which has a derivative with respect to the parameter  $w$  of  $-2\sigma(y - \sigma w)$ .

Gradient descent with a learning rate of  $\eta$  follows the recurrence-relation or discrete-time state evolution of:

$$\begin{aligned} w_{t+1} &= w_t + 2\eta\sigma(y - \sigma w_t) \\ &= (1 - 2\eta\sigma^2)w_t + 2\eta\sigma y. \end{aligned} \tag{3}$$

(a) **For what values of learning rate  $\eta > 0$  is the recurrence (3) stable?**

*(HINT: Remember the role of the unit circle in determining the stability or instability of such recurrences. If you keep taking higher and higher positive integer powers of a number, what does that number has to be like for this to converge?)*

**Solution:** We can rewrite the update rule as

$$\begin{aligned} w_{t+1} - \frac{y}{\sigma} &= (1 - 2\eta\sigma^2)(w_t - \frac{y}{\sigma}) \\ w_{t+1} &= \frac{y}{\sigma} + (1 - 2\eta\sigma^2)^{t+1}(w_0 - \frac{y}{\sigma}) \end{aligned} \tag{4}$$

To make the recurrence (3) stable with  $\eta > 0$ , we need  $|1 - 2\eta\sigma^2| < 1$ . This gives  $\eta < \frac{1}{\sigma^2}$ .

(b) The previous part gives you an upper bound for the learning rate  $\eta$  that depends on  $\sigma$  beyond which we cannot safely go. **If  $\eta$  is below that upper bound, how fast does  $w_t$  converge to its final solution  $w^* = \frac{y}{\sigma}$ ? i.e. if we wanted to get within a factor  $(1 - \epsilon)$  of  $w^*$ , how many iterations  $t$  would we need?**

*(HINT: The absolute value of the error of current  $w$  to the optimality might help.)*

**Solution:**

$$|w_T - w^*| < \epsilon |w^*| \tag{5}$$

Use the derived update rule in (4). We have

$$\begin{aligned} |w_T - \frac{y}{\sigma}| &< \epsilon \left| \frac{y}{\sigma} \right| \\ |(1 - 2\eta\sigma^2)^T| &< \epsilon \\ T &> \frac{\log(\epsilon)}{\log(|1 - 2\eta\sigma^2|)}, \end{aligned} \quad (6)$$

- (c) Suppose that we now have a vector problem where we have two parameters  $w[1], w[2]$ . One with a large  $\sigma_\ell$  and the other with a tiny  $\sigma_s$ . i.e.  $\sigma_\ell \gg \sigma_s$  and we have the vector equation we want to solve:

$$\begin{bmatrix} \sigma_\ell & 0 \\ 0 & \sigma_s \end{bmatrix} \begin{bmatrix} w[1] \\ w[2] \end{bmatrix} = \begin{bmatrix} y[1] \\ y[2] \end{bmatrix}. \quad (7)$$

We use gradient descent with a single learning rate  $\eta$  to solve this problem starting from an initial condition of  $\mathbf{w} = \mathbf{0}$ .

**For what learning rates  $\eta > 0$  will we converge? Which of the two  $\sigma_i$  is limiting our learning rate?**

**Solution:** Similarly, we can rewrite the loss function and update rule w.r.t the vector form.

$$\begin{aligned} L(\mathbf{w}) &= \|\mathbf{y} - \Sigma\mathbf{w}\|^2, \Sigma = \begin{bmatrix} \sigma_\ell & 0 \\ 0 & \sigma_s \end{bmatrix} \\ \nabla_{\mathbf{w}} L(\mathbf{w}) &= 2(\Sigma^2\mathbf{w} - \Sigma\mathbf{y}) \\ \mathbf{w}_{t+1} &= (I - 2\eta\Sigma^2)\mathbf{w}_t + 2\eta\Sigma\mathbf{y} \end{aligned} \quad (8)$$

To ensure the convergence, we need

$$\begin{cases} |1 - 2\eta\sigma_\ell^2| < 1 \\ |1 - 2\eta\sigma_s^2| < 1 \end{cases} \quad (9)$$

$$\eta < \min\left(\frac{1}{\sigma_\ell^2}, \frac{1}{\sigma_s^2}\right) = \frac{1}{\sigma_\ell^2} \quad (10)$$

- (d) **For the previous problem, depending on  $\eta, \sigma_\ell, \sigma_s$ , which of the two dimensions is converging faster and which is converging slower?**

**Solution:** We can rewrite the update rule w.r.t each dimension, this gives

$$\begin{aligned} w[1]_t &= \frac{y[1]}{\sigma_\ell} + (1 - 2\eta\sigma_\ell^2)^t \left(-\frac{y[1]}{\sigma_\ell}\right) \\ w[2]_t &= \frac{y[2]}{\sigma_s} + (1 - 2\eta\sigma_s^2)^t \left(-\frac{y[2]}{\sigma_s}\right) \end{aligned}$$

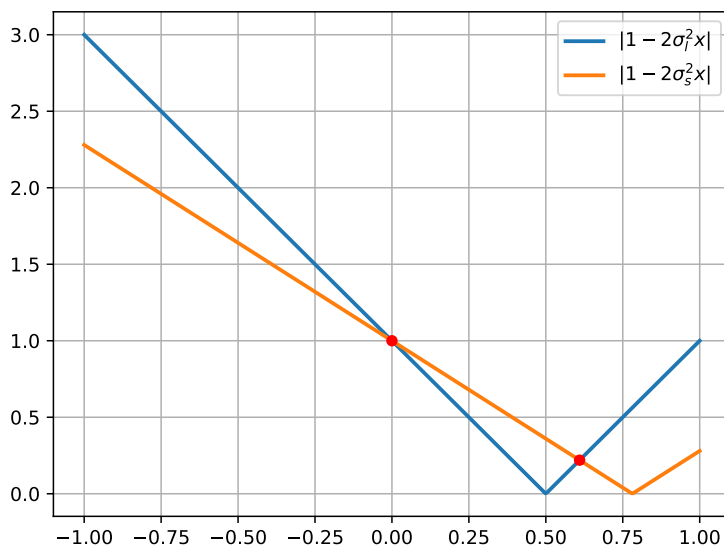
This faster convergence dimension is  $\min(|1 - 2\eta\sigma_\ell^2|, |1 - 2\eta\sigma_s^2|)$

- (e) The speed of convergence overall will be dominated by the slower of the two. **For what value of  $\eta$  will we get the fastest overall convergence to the solution?**

**Solution:** Recall that the minimum of the pointwise maximum of two functions occurs at a point where both functions are equal. Thus, the fastest convergence is achieved when  $|1 - 2\eta\sigma_\ell^2| = |1 - 2\eta\sigma_s^2|$ , and since  $\sigma_\ell \geq \sigma_s$ , there are two solutions, one is the trivial  $\eta = 0$ , and the other is

$$\eta = \frac{1}{\sigma_\ell^2 + \sigma_s^2}$$

See Figure 1 for a picture.



**Figure 1:** Plot of  $|1 - 2\eta\sigma_\ell^2|$ ,  $|1 - 2\eta\sigma_s^2|$ , showing their two intersections. One is the trivial  $\eta = 0$ , and the other is the nontrivial  $\eta = \frac{1}{\sigma_\ell^2 + \sigma_s^2}$ .

- (f) Comment on what would happen if we had more parallel problems with  $\sigma_i$  that all were in between  $\sigma_\ell$  and  $\sigma_s$ ? **Would they influence the choice of possible learning rates or the learning rate with the fastest convergence?**

**Solution:** The bounds on the learning rate should still be the same. For maximal convergence, the largest and smallest values of sigma are the two that cause gradient descent to take the longest if we poorly choose a learning rate.

- (g) Using what you know about the SVD, **how is the simple scalar and parallel scalar problem analysis above relevant to solving general least-squares problems of the form  $X\mathbf{w} \approx \mathbf{y}$  using gradient descent?**

**Solution:** We can think of SVD as a change of bases into and then back from a coordinate system where  $X$  instead is just diagonal, this will directly connect to the same problem with lots of parallel scalar problems.

## 2. Stochastic Gradient Descent (when it is possible to interpolate)

This is a problem about the convergence of SGD for least-squares problems when there is actually a solution that achieves zero loss.

For simplicity, suppose that the problem we are given is

$$X\mathbf{w} = \mathbf{y} \quad (11)$$

where  $X = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}$  is a wide matrix with  $\mathbf{x}_i$  being  $d$ -dimensional vectors and  $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$  is an  $n$ -dimensional vector. Here, we assume that  $X$  has full row-rank (i.e. the  $\mathbf{x}_i$  are linearly independent) and so (11) indeed has solutions. (Note that as  $d > n$ , there there would be infinitely many solutions.)

While we already know lots of ways of solving this problem, it is an illustrative toy example to make sure we understand why SGD works in such settings. (This material was covered in lecture but you really do need to understand it yourself so you can deal with variations you might encounter as well as internalize the kinds of manipulations required.) This problem has a Demo Notebook ([demo link](#)) associated with it to help you play around with things to get an even deeper set of intuitions.

In this problem, we will just initialize  $\mathbf{w}_0 = \mathbf{0}$  for simplicity.

- (a) Let's do some preliminaries. First, we want to change coordinates to notationally simplify our analysis of SGD.

Let  $\mathbf{w}^*$  be the min-norm solution to (11).

**Write out what  $\mathbf{w}^*$  is explicitly with respect to  $X$  and  $\mathbf{y}$**  and then, change coordinates to  $\mathbf{w}' = \mathbf{w} - \mathbf{w}^*$  to write the new equations as:

$$X\mathbf{w}' = \mathbf{0} \quad (12)$$

**What is the new initial condition for  $\mathbf{w}'_0$ ?**

**Solution:** The min-norm solution  $\mathbf{w}^*$  to (11) is

$$\mathbf{w}^* = X^T(XX^T)^{-1}\mathbf{y} \quad (13)$$

and so the new equation is

$$X\mathbf{w}' = X(\mathbf{w} - \mathbf{w}^*) = X\mathbf{w} - X\mathbf{w}^* = \mathbf{y} - \mathbf{y} = \mathbf{0} \quad (14)$$

and the new initial condition is  $\mathbf{w}'_0 = \mathbf{w}_0 - \mathbf{w}^* = -\mathbf{w}^*$ .

- (b) Next, let's leverage SVD coordinates to further simplify the problem. **Show that there exists an orthonormal transformation  $V$  of variables  $\mathbf{w}'' = V\mathbf{w}'$  so that (12) looks like**

$$[\tilde{X} \quad \mathbf{0}_{n \times (d-n)}]\mathbf{w}'' = \mathbf{0} \quad (15)$$

and furthermore, **show that the initial condition for  $\mathbf{w}'_0$  you computed in the previous part, when viewed as  $\mathbf{w}''_0$  has all zeros in the final  $(d - n)$  positions.**

**Solution:** SVD of  $X$  is  $X = U\Sigma V^T$  where  $U \in \mathbb{R}^{n \times n}$ ,  $\Sigma \in \mathbb{R}^{n \times d}$ ,  $V \in \mathbb{R}^{d \times d}$  and  $U$  and  $V$  are orthonormal. Non-zero singular values of  $X$  are  $\sigma_1, \sigma_2, \dots, \sigma_n$  and the rest are zero.  $\Sigma$  can be written

as

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \dots & 0 & 0 \\ & \sigma_2 & & \dots & 0 & 0 \\ & & \ddots & \dots & 0 & 0 \\ & & & \sigma_n & \dots & 0 & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & \mathbf{0}_{n \times (d-n)} \end{bmatrix} \quad (16)$$

$$\text{, where } \Sigma_1 = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \in \mathbb{R}^{n \times n}. \text{ Therefore, we can write } X\mathbf{w}' \text{ as}$$

$$X\mathbf{w}' = U\Sigma V^T \mathbf{w}' \quad (17)$$

$$= U \begin{bmatrix} \Sigma_1 & \mathbf{0}_{n \times (d-n)} \end{bmatrix} V^T \mathbf{w}' \quad (18)$$

$$= \begin{bmatrix} U\Sigma_1 & \mathbf{0}_{n \times (d-n)} \end{bmatrix} V^T \mathbf{w}' \quad (19)$$

$$= \begin{bmatrix} \tilde{X} & \mathbf{0}_{n \times (d-n)} \end{bmatrix} V^T \mathbf{w}' \quad (20)$$

$$= \begin{bmatrix} \tilde{X} & \mathbf{0}_{n \times (d-n)} \end{bmatrix} \mathbf{w}'' \quad (21)$$

where  $\tilde{X} = U\Sigma_1$  and  $\mathbf{w}'' = V^T \mathbf{w}'$ . Therefore, (12) becomes (15) and the initial condition for  $\mathbf{w}''_0$  is  $\mathbf{w}''_0 = V^T \mathbf{w}'_0 = V^T(-\mathbf{w}^*)$ . Let's derive initial condition for  $\mathbf{w}''_0$  with more details.

$$\mathbf{w}''_0 = -V^T X^T (X X^T)^{-1} \mathbf{y} \quad (22)$$

$$= -V^T (U\Sigma V^T)^T ((U\Sigma V^T)(U\Sigma V^T)^T)^{-1} \mathbf{y} \quad (23)$$

$$= -V^T V \Sigma^T U^T (U\Sigma V^T V \Sigma^T U^T)^{-1} \mathbf{y} \quad (24)$$

$$= -V^T V \begin{bmatrix} \Sigma_1 \\ \mathbf{0}_{(d-n) \times n} \end{bmatrix} U^T (U \begin{bmatrix} \Sigma_1 & \mathbf{0}_{n \times (d-n)} \end{bmatrix} \begin{bmatrix} \Sigma_1^T \\ \mathbf{0}_{(d-n) \times n} \end{bmatrix} U^T)^{-1} \mathbf{y} \quad (25)$$

$$= - \begin{bmatrix} \Sigma_1 \\ \mathbf{0}_{(d-n) \times n} \end{bmatrix} U^T (U \Sigma_1^2 U^T)^{-1} \mathbf{y} \quad (26)$$

$$= - \begin{bmatrix} \Sigma_1 \\ \mathbf{0}_{(d-n) \times n} \end{bmatrix} U^T U \Sigma_1^{-2} U^T \mathbf{y} \quad (27)$$

$$= - \begin{bmatrix} \Sigma_1^{-1} \\ \mathbf{0}_{(d-n) \times n} \end{bmatrix} U^T \mathbf{y} \quad (28)$$

$$= - \begin{bmatrix} \Sigma_1^{-1} U^T \mathbf{y} \\ \mathbf{0}_{d-n} \end{bmatrix} \quad (29)$$

Therefore, the initial condition for  $\mathbf{w}''_0$  has all zeros in the final  $(d - n)$  positions.

- (c) **Argue why what you have seen in the previous parts allows us to now focus on a square system of equations:**

$$\tilde{X} \tilde{\mathbf{w}} = \mathbf{0} \quad (30)$$

**and furthermore show that each of the  $n$  constituent equations (corresponding to rows) of (30) can be obtained by means of coordinate changes from the same indexed equation in (11).**

**Solution:** Let's first show that we can focus on a square system of equations. As we have shown in the previous parts, the initial condition for  $\mathbf{w}_0''$  has all zeros in the final  $(d-n)$  positions and the system of equations for  $\mathbf{w}''$  is all zeros in the final  $(d-n)$  columns. Therefore, the final  $(d-n)$  elements of  $\mathbf{w}_t''$  will always be zero. Therefore, we can focus on a square system of equations.

The  $i$ -th equation of (30) is  $\tilde{x}_i^T \tilde{\mathbf{w}} = 0$ . Let's show that this equation can be obtained by means of coordinate changes from the  $i$ -th equation of (11).

$$\tilde{x}_i^T \tilde{\mathbf{w}} = \begin{bmatrix} \tilde{x}_i^T & \mathbf{0}_{(d-n) \times 1}^T \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{w}} \\ \mathbf{0}_{(d-n) \times 1} \end{bmatrix} \quad (31)$$

$$= \begin{bmatrix} \tilde{x}_i^T & \mathbf{0}_{(d-n) \times 1}^T \end{bmatrix} \mathbf{w}'' \quad (32)$$

$$= \begin{bmatrix} \tilde{x}_i^T & \mathbf{0}_{(d-n) \times 1}^T \end{bmatrix} V^T \mathbf{w}' \quad (33)$$

$$= \begin{bmatrix} \text{row}_i(U) \Sigma_1 & \mathbf{0}_{n \times (d-n)} \end{bmatrix} V^T \mathbf{w}' \quad (34)$$

$$= \text{row}_i(U) \Sigma V^T \mathbf{w}' \quad (35)$$

$$= \mathbf{x}_i^T \mathbf{w}' \quad (36)$$

$$(37)$$

Therefore, the  $i$ -th equation of (30) can be obtained by means of coordinate changes from the  $i$ -th equation of (11).

- (d) Let's now engage with SGD itself. Here, we will just use a minibatch length of 1 and batch sampling with replacement. This means that at every iteration of SGD, we roll a fair  $n$ -sided die and choose the single equation in (11) that corresponds to the row that came up on the die. Let  $I_t$  be the iid uniform random variable on  $\{1, \dots, n\}$  that we roll after iteration  $t$ .

At the  $t + 1$ -th iteration, we compute

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}_{I_t}(\mathbf{w}_t) \quad (38)$$

where  $\mathcal{L}_i(\mathbf{w}) = (y[i] - \mathbf{x}_i^T \mathbf{w})^2$  is the squared loss on the  $i$ -th equation and  $\eta$  is the step-size (learning rate).

**Show that an SGD step taken in (38) for the original optimization problem matches exactly to an SGD step taken for  $\tilde{\mathbf{w}}$  for solving (30), and that in particular these steps look like:**

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t - 2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^T \tilde{\mathbf{w}}_t \quad (39)$$

**Solution:** Let's first derive the SGD update for the original optimization problem with respect to  $\mathbf{w}'$ . As  $\mathbf{w}'_t = \mathbf{w}_t - \mathbf{w}^*$ , we can write the SGD update for the original optimization problem as

$$\mathbf{w}'_{t+1} = \mathbf{w}'_t - \eta \nabla \mathcal{L}_{I_t}(\mathbf{w}'_t + \mathbf{w}^*) \quad (40)$$

$$= \mathbf{w}'_t - 2\eta (\mathbf{x}_{I_t}^T \mathbf{w}'_t + \mathbf{x}_{I_t}^T \mathbf{w}^* - y_{I_t}) \mathbf{x}_{I_t} \quad (41)$$

$$= \mathbf{w}'_t - 2\eta \mathbf{x}_{I_t} \mathbf{x}_{I_t}^T \mathbf{w}'_t \quad (42)$$

$$(43)$$

$\mathbf{w}' = V \mathbf{w}''$  and  $\mathbf{x}_i = V \tilde{x}_i$ . Therefore, we can write the SGD update for the original optimization

problem with respect to  $\mathbf{w}''$  as

$$V\mathbf{w}''_{t+1} = V\mathbf{w}''_t - 2\eta V\tilde{x}_{I_t}\tilde{x}_{I_t}^T V^T V\mathbf{w}''_t \quad (44)$$

$$= V\mathbf{w}''_t - 2\eta V\tilde{x}_{I_t}\tilde{x}_{I_t}^T \mathbf{w}''_t \quad (45)$$

$$\mathbf{w}''_{t+1} = \mathbf{w}''_t - 2\eta\tilde{x}_{I_t}\tilde{x}_{I_t}^T \mathbf{w}''_t \quad (46)$$

Therefore, the SGD update for the original optimization problem with respect to  $\mathbf{w}''$  is the same as the SGD update for (30).

- (e) At this point, we can focus entirely on the simplified square system (30) and the stochastic evolution of the iterations described by (39).

To show convergence to zero, we need to pick a suitable stochastic Lyapunov function  $\mathcal{L}(\tilde{w})$  that is bounded below by zero and will decrease in expectation at every time step. In particular, we want to establish

$$E[\mathcal{L}(\tilde{w}_{t+1})|\tilde{w}_t] < (1 - \rho)\mathcal{L}(\tilde{w}_t) \quad (47)$$

with a  $1 > \rho > 0$  so that this Lyapunov function tends to decrease exponentially to zero. We will have to have a suitably small step-size/learning-rate  $\eta$  for this to happen, of course.

**Show that ((e)) indeed implies that for every  $\epsilon > 0$  and  $\delta > 0$ , there exists a  $T > 0$  for which**

$$P(\mathcal{L}(\tilde{w}_T) < \epsilon) \geq 1 - \delta. \quad (48)$$

**Solution:** From , we can derive the relationship between  $\mathcal{L}(\tilde{w}_{t+1})$  and  $\mathcal{L}(\tilde{w}_t)$ .

$$E[\mathcal{L}(\tilde{w}_{t+1})] = E[E[\mathcal{L}(\tilde{w}_{t+1})|\tilde{w}_t]] < E[(1 - \rho)\mathcal{L}(\tilde{w}_t)] = (1 - \rho)E[\mathcal{L}(\tilde{w}_t)] \quad (49)$$

Therefore, we have

$$E[\mathcal{L}(\tilde{w}_T)] < (1 - \rho)^T E[\mathcal{L}(\tilde{w}_0)] \quad (50)$$

As Lyapunov function is bounded below by zero, we can apply Markov's inequality to get

$$P(\mathcal{L}(\tilde{w}_T) < \epsilon) \geq 1 - \frac{E[\mathcal{L}(\tilde{w}_T)]}{\epsilon} \geq 1 - \frac{(1 - \rho)^T E[\mathcal{L}(\tilde{w}_0)]}{\epsilon} \geq 1 - \delta \quad (51)$$

Let's choose  $T$  such that  $(1 - \rho)^T E[\mathcal{L}(\tilde{w}_0)] \leq \epsilon\delta$ . This proof implies that as  $t$  goes to infinity,  $\mathcal{L}(\tilde{w}_t) \rightarrow 0$  with probability 1. It converges exponentially to zero. The learning rate  $\eta$  determines  $\rho$  and therefore, convergence rate.

- (f) One natural guess for a stochastic Lyapunov function is

$$\mathcal{L}(\tilde{w}) = \tilde{w}^\top \tilde{X}^\top \tilde{X} \tilde{w}. \quad (52)$$

**Argue why the candidate Lyapunov function  $\mathcal{L}(\tilde{w})$  in (52) is non-negative and is only equal to zero at  $\tilde{w} = \mathbf{0}$ .**

**Solution:** Note that  $\tilde{X} \in \mathbb{R}^{n \times n}$  is full rank. Therefore, the matrix  $\tilde{X}^\top \tilde{X}$  is strictly positive definite. Therefore,  $\tilde{w}^\top \tilde{X}^\top \tilde{X} \tilde{w} \geq 0$  and  $\tilde{w}^\top \tilde{X}^\top \tilde{X} \tilde{w} = 0$  if and only if  $\tilde{w} = \mathbf{0}$ .

- (g) Now, with a guessed stochastic Lyapunov function in hand, we can try to show ((e)). The first step will

be to decompose the evolution of  $\mathcal{L}(\tilde{w})$  into three parts:

$$\mathcal{L}(\tilde{w}_{t+1}) = \mathcal{L}(\tilde{w}_t) + A + B \quad (53)$$

where the term  $A$  is linear in the actual stochastic update  $(\tilde{w}_{t+1} - \tilde{w}_t)$  and the term  $B$  is quadratic in that update.

**Expand out  $\mathcal{L}(\tilde{w}_t + (\tilde{w}_{t+1} - \tilde{w}_t))$  to give explicit forms for  $A$  and  $B$ .**

**Solution:**

$$\mathcal{L}(\tilde{w}_t + (\tilde{w}_{t+1} - \tilde{w}_t)) = (\tilde{w}_t + (\tilde{w}_{t+1} - \tilde{w}_t))^\top \tilde{X}^\top \tilde{X} (\tilde{w}_t + (\tilde{w}_{t+1} - \tilde{w}_t)) \quad (54)$$

$$= \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \tilde{w}_t + 2\tilde{w}_t^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) + (\tilde{w}_{t+1} - \tilde{w}_t)^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) \quad (55)$$

$$= \mathcal{L}(\tilde{w}_t) + 2\tilde{w}_t^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) + (\tilde{w}_{t+1} - \tilde{w}_t)^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) \quad (56)$$

Therefore, we have

$$A = 2\tilde{w}_t^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) \quad (57)$$

$$B = (\tilde{w}_{t+1} - \tilde{w}_t)^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) \quad (58)$$

- (h) We are counting on the term  $A$  in (53) to give us actual contraction in expectation since this looks like a gradient-descent step. **Show:**

$$E[A|\tilde{w}_t] \leq -c_1 \eta \mathcal{L}(\tilde{w}_t) \quad (59)$$

where the  $c_1 > 0$  is a positive constant that depends on the problem.

(Hint: you are going to want to leverage the actual updates in (39) as well as the singular value structure for  $\tilde{X}$ . Can the smallest singular value be zero?)

**Solution:** Plugging in the SGD update for (39) into  $\tilde{w}_{t+1}$ , we have

$$E[A|\tilde{w}_t] = E[2\tilde{w}_t^\top \tilde{X}^\top \tilde{X} (\tilde{w}_{t+1} - \tilde{w}_t) | \tilde{w}_t] \quad (60)$$

$$= 2\tilde{w}_t^\top \tilde{X}^\top \tilde{X} E[(\tilde{w}_{t+1} - \tilde{w}_t) | \tilde{w}_t] \quad (61)$$

$$= 2\tilde{w}_t^\top \tilde{X}^\top \tilde{X} E[-2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t | \tilde{w}_t] \quad (62)$$

$$= -4\eta \tilde{w}_t^\top \tilde{X}^\top \tilde{X} E[\tilde{x}_{I_t} \tilde{x}_{I_t}^\top] \tilde{w}_t \quad (63)$$

$$= -\frac{4}{n} \eta \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \tilde{X}^\top \tilde{X} \tilde{w}_t \quad (64)$$

$$\leq -\frac{4}{n} \eta \sigma_{\min}^2 \tilde{w}_t^\top \tilde{X}^\top \tilde{X} \tilde{w}_t \quad (65)$$

$$= -\frac{4}{n} \eta \sigma_{\min}^2 \mathcal{L}(\tilde{w}_t) \quad (66)$$

Therefore, we have  $c_1 = \frac{4}{n} \eta \sigma_{\min}^2$ . Note that  $\sigma_{\min} > 0$  as  $\tilde{X}$  is full rank.

- (i) We need to make sure that the “quadratic” term  $B$  in (53) cannot undo the progress made by  $A$  in expectation. **Show:**

$$E[B|\tilde{w}_t] \leq c_2 \eta^2 \mathcal{L}(\tilde{w}_t) \quad (67)$$

where  $c_2 > 0$  is another positive constant that depends on the problem.

(Hint: you are going to want to leverage the actual updates in (39), the singular value structure for  $\tilde{X}$ , and the fact that the rows of  $\tilde{X}$  can only be so big. You will want to leverage the largest singular value of  $\tilde{X}$  for one bounding step and then let  $\beta$  be the largest norm of the rows of  $\tilde{X}$  to do another bounding step.)

**Solution:** Plugging in the SGD update for (39) into  $\tilde{w}_{t+1}$ , we have

$$E[B|\tilde{w}_t] = E[(\tilde{w}_{t+1} - \tilde{w}_t)^\top \tilde{X}^\top \tilde{X}(\tilde{w}_{t+1} - \tilde{w}_t)|\tilde{w}_t] \quad (68)$$

$$= E[-2\eta \tilde{w}_t^\top \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{X}^\top \tilde{X}(-2\eta \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t)|\tilde{w}_t] \quad (69)$$

$$= 4\eta^2 E[\tilde{w}_t^\top \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{X}^\top \tilde{X} \tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{w}_t|\tilde{w}_t] \quad (70)$$

$$\leq 4\eta^2 \sigma_{\max}^2 \tilde{w}_t^\top E[\tilde{x}_{I_t} \tilde{x}_{I_t}^\top \tilde{x}_{I_t} \tilde{x}_{I_t}^\top] \tilde{w}_t \quad (71)$$

$$\leq 4\eta^2 \sigma_{\max}^2 \rho^2 \tilde{w}_t^\top E[\tilde{x}_{I_t} \tilde{x}_{I_t}^\top] \tilde{w}_t \quad (72)$$

$$= \frac{4}{n} \eta^2 \sigma_{\max}^2 \rho^2 \mathcal{L}(\tilde{w}_t) \quad (73)$$

where,  $\sigma_{\max}$  is the largest singular value of  $\tilde{X}$  and  $\rho$  is the largest norm of the rows of  $\tilde{X}$ . Therefore, we have  $c_2 = \frac{4}{n} \eta^2 \sigma_{\max}^2 \rho^2$ .

(j) Finally, we can put the pieces together to see that

$$E[\mathcal{L}(\tilde{w}_{t+1})|\tilde{w}_t] \leq (1 - c_1\eta + c_2\eta^2)\mathcal{L}(\tilde{w}_t) \quad (74)$$

where  $c_1 > 0$  and  $c_2 > 0$  as well. **Show that this means that there exists a small enough  $\eta$  so that  $1 - c_1\eta + c_2\eta^2 < 1$ .**

**Solution:**

$$E[\mathcal{L}(\tilde{w}_{t+1})|\tilde{w}_t] = E[\mathcal{L}(\tilde{w}_t) + A + B|\tilde{w}_t] \quad (75)$$

$$= E[\mathcal{L}(\tilde{w}_t)] + E[A|\tilde{w}_t] + E[B|\tilde{w}_t] \quad (76)$$

$$\leq \mathcal{L}(\tilde{w}_t) - \frac{4}{n} \eta \sigma_{\min}^2 \mathcal{L}(\tilde{w}_t) + \frac{4}{n} \eta^2 \rho^2 \sigma_{\max}^2 \mathcal{L}(\tilde{w}_t) \quad (77)$$

$$= (1 - \frac{4}{n} \eta \sigma_{\min}^2 + \frac{4}{n} \eta^2 \rho^2 \sigma_{\max}^2) \mathcal{L}(\tilde{w}_t) \quad (78)$$

Therefore, we have  $1 - c_1\eta + c_2\eta^2 < 1$  if  $\eta < \frac{c_1}{c_2}$ . To minimize  $\frac{c_1}{c_2}$ , we can choose  $\eta = \frac{c_1}{2c_2}$ . Note that  $c_1$  and  $c_2$  are positive constants.

(k) In earlier problem set, you saw how you could reinterpret ridge-regression using feature-augmentation. The earlier parts of this problem have now established that leveraging that trick, you can get SGD to converge exponentially for ridge regression. Check out Jupyter notebook in this [demo link](#), and **report what you observed in terms of the convergence rate**.

One of the lessons that you will observe from the code is that the implementation details matter. If you do ridge regression and just treat it as an optimization problem, you won't just be able to use SGD and get exponential convergence with a constant step size. (You would have to adjust the step sizes to make them smaller, but this would slow down your convergence considerably.) But if you intelligently use the feature-augmentation perspective on ridge regression, you'll get exponential convergence.

This is why it is vital for people in EECS to really understand machine learning at the level of detail that we are teaching you. Because in the real world, even if you are a practicing machine learning engineer, if you are working on cutting-edge systems, you need to understand how to implement what

you want to do so that it works fast. Equivalent formulations mathematically need not be equivalent from the point of view of implementation – this is one dramatic example of a case when they are not. Take EE227C and beyond if you want to understand these things more deeply.

**Solution:** This SGD with constant step size (??) converges exponentially towards zero in these settings.

For people coming from an optimization point of view who have seen SGD before, this might seem perplexing. After all, it is generically the case that the convergence of SGD to a solution requires an appropriately diminishing step size, and that the convergence is far slower than exponential. This is where the machine-learning understanding is important as well as understanding the special case of interpolating solutions. The machine-learning understanding tells us that by doing SGD in the augmented-features point of view, we are keeping a little bit of memory for every training point. This memory is being used to help us better “agree to disagree” on each individual point instead of constantly having the same argument over and over again between the individual data points and the collective impact of the rest of the data points. Whereas traditional SGD convergence requires an increasing stubbornness of the collective vis-a-vis each data point (that is what a diminishing step size means — it is harder to change the collective’s point of view), the augmented features perspective gets around this by allowing learning to focus on the new disagreements as opposed to old ones.

The augmented-features perspective on ridge regression is meant to help you understand what is spiritually going on with deep learning in the context of a giant model that might very well have more parameters in it than you have data points. The sheer mass of those extra parameters can act like the ridge augmentation done explicitly here. This is what permits what is sometimes called “benign overfitting” or “harmless interpolation” in the context of deep learning models.

### 3. Accelerating Gradient Descent with Momentum

Consider the problem of finding the minimizer of the following objective:

$$\mathcal{L}(w) = \|y - Xw\|_2^2 \quad (79)$$

In an earlier problem, we proved that gradient descent (GD) algorithm can converge and derive the convergence rate. In this problem, we will add the momentum term and see how it affects to the convergence rate. The optimization procedure of gradient descent+momentum is given below:

$$\begin{aligned} w_{t+1} &= w_t - \eta z_{t+1} \\ z_{t+1} &= (1 - \beta)z_t + \beta g_t, \end{aligned} \quad (80)$$

where  $g_t = \nabla \mathcal{L}(w_t)$ ,  $\eta$  is learning rate and  $\beta$  defines how much averaging we want for the gradient. Note that when  $\beta = 1$ , the above procedure is just the original gradient descent.

Let’s investigate the effect of this change. We’ll see that this modification can actually ‘accelerate’ the convergence by allowing larger learning rates.

(a) Recall that the gradient descent update of (79) is

$$w_{t+1} = \left( I - 2\eta(X^T X) \right) w_t + 2\eta X^T y \quad (81)$$

and the minimizer is

$$w^* = (X^T X)^{-1} X^T y \quad (82)$$

The geometric convergence rate (in the sense of what base is there for convergence as  $\text{rate}^t$ ) of this procedure is

$$\text{rate} = \max_i |1 - 2\eta\sigma_i^2| \quad (83)$$

You already saw if we choose the learning rate that maximizes (83), the optimal learning rate,  $\eta^*$  is

$$\eta^* = \frac{1}{\sigma_{\min}^2 + \sigma_{\max}^2}, \quad (84)$$

where  $\sigma_{\max}$  and  $\sigma_{\min}$  are the maximum and minimum singular value of the matrix  $X$ . The corresponding optimal convergence rate is

$$\text{optimal rate} = \frac{(\sigma_{\max}/\sigma_{\min})^2 - 1}{(\sigma_{\max}/\sigma_{\min})^2 + 1} \quad (85)$$

Therefore, how fast ordinary gradient descent converges is determined by the ratio between the maximum singular value and the minimum singular value as above.

Now, let's consider using momentum to smooth the gradients before taking a step in (80).

$$\begin{aligned} w_{t+1} &= w_t - \eta z_{t+1} \\ z_{t+1} &= (1 - \beta)z_t + \beta(2X^T X w_t - 2X^T y) \end{aligned} \quad (86)$$

We can use the SVD of the matrix  $X = U\Sigma V^T$ , where  $\Sigma = \text{diag}(\sigma_{\max}, \sigma_2, \dots, \sigma_{\min})$  with the same (potentially rectangular) shape as  $X$ . This allows us to reparameterize the parameters  $w_t$  and averaged gradients  $z_t$  as below:

$$\begin{aligned} x_t &= V^T(w_t - w^*) \\ a_t &= V^T z_t. \end{aligned} \quad (87)$$

**Please rewrite (86) with the reparameterized variables,  $x_t[i]$  and  $a_t[i]$ . ( $x_t[i]$  and  $a_t[i]$  are  $i$ -th components of  $x_t$  and  $a_t$  respectively.)**

**Solution:** Let's multiply  $V^T$  both sides in Eq.(86). Then,

$$\begin{aligned} V^T w_{t+1}[i] &= V^T w_t[i] - \eta V^T z_{t+1}[i] \\ x_{t+1}[i] + V^T w^*[i] &= x_t[i] + V^T w^*[i] - \eta a_{t+1}[i] \\ x_{t+1}[i] &= x_t[i] - \eta a_{t+1}[i] \end{aligned}$$

$$\begin{aligned} V^T z_{t+1} &= (1 - \beta)V^T z_t + \beta V^T(2X^T X w_t - 2X^T y) \\ a_{t+1} &= (1 - \beta)a_t + \beta V^T(2V\Sigma^2 V^T(Vx_t + w^*) - 2X^T y) \\ a_{t+1} &= (1 - \beta)a_t + \beta(2\Sigma^2 V^T Vx_t + \Sigma^2 V^T w^* - 2V^T X^T y) \end{aligned} \quad (88)$$

Note that,

$$w^* = (X^T X)^{-1} X^T y \quad (89)$$

$$= V \Sigma^{-2} V^T X^T y \quad (90)$$

Let's plug the above result into Eq.(88).

$$\begin{aligned} a_{t+1} &= (1 - \beta)a_t + \beta(2\Sigma^2 V^T V x_t + 2\Sigma^2 V^T w^* - 2V^T X^T y) \\ &= (1 - \beta)a_t + \beta(2\Sigma^2 V^T V x_t + 2\Sigma^2 V^T V \Sigma^{-2} V^T X^T y - 2V^T X^T y) \\ &= (1 - \beta)a_t + \beta 2\Sigma^2 V^T V x_t \\ &= (1 - \beta)a_t + \beta 2\Sigma^2 x_t \end{aligned}$$

If we express the above result element-wise,

$$\begin{aligned} a_{t+1}[i] &= (1 - \beta)a_t[i] + 2\beta\sigma_i^2 x_t[i] \\ x_{t+1}[i] &= x_t[i] - \eta a_{t+1}[i] \end{aligned}$$

- (b) Notice that the above  $2 \times 2$  vector/matrix recurrence has no external input. We can derive the  $2 \times 2$  system matrix  $R_i$  from above such that

$$\begin{bmatrix} a_{t+1}[i] \\ x_{t+1}[i] \end{bmatrix} = R_i \begin{bmatrix} a_t[i] \\ x_t[i] \end{bmatrix} \quad (91)$$

**Derive  $R_i$ .**

**Solution:**

Since  $x_{t+1}$  is expressed with  $x_t$  and  $a_{t+1}$ , let's reformulate with  $x_t$  and  $a_t$

$$\begin{aligned} x_{t+1}[i] &= x_t[i] - \eta a_{t+1}[i] \\ &= x_t[i] - \eta(1 - \beta)a_t[i] - 2\eta\beta\sigma_i^2 x_t[i] \\ &= (1 - 2\eta\beta\sigma_i^2)x_t[i] - \eta(1 - \beta)a_t[i] \end{aligned}$$

Therefore, the matrix  $R_i$  can be represented as below:

$$R_i = \begin{bmatrix} (1 - \beta) & 2\beta\sigma_i^2 \\ -\eta(1 - \beta) & 1 - 2\eta\beta\sigma_i^2 \end{bmatrix}$$

- (c) Use the computer to symbolically find the eigenvalues of the matrix  $R_i$ .

**When are they purely real? When are they repeated and purely real? When are they complex?**

**Solution:** Let's derive the characteristic equation of  $R_i$  to derive eigenvalues:

$$\begin{aligned} f(\lambda) &= |R_i - \lambda I| \\ &= \left| \begin{bmatrix} (1 - \beta) - \lambda & 2\beta\sigma_i^2 \\ -\eta(1 - \beta) & 1 - 2\eta\beta\sigma_i^2 - \lambda \end{bmatrix} \right| \end{aligned} \quad (92)$$

$$= \lambda^2 - \left(2 - \beta - 2\eta\beta\sigma_i^2\right) \lambda + 1 - \beta \quad (93)$$

Setting this to zero and solving, we get that the eigenvalues are:

$$\lambda_{1,i} = 1 - \beta\eta\sigma_i^2 - \frac{\beta}{2} - \frac{\sqrt{\beta(4\beta\eta^2\sigma_i^4 + 4\beta\eta\sigma_i^2 + \beta - 8\eta\sigma_i^2)}}{2}$$

$$\lambda_{2,i} = 1 - \beta\eta\sigma_i^2 - \frac{\beta}{2} + \frac{\sqrt{\beta(4\beta\eta^2\sigma_i^4 + 4\beta\eta\sigma_i^2 + \beta - 8\eta\sigma_i^2)}}{2}$$

The discriminant of (92) is:

$$D = \left(2 - \beta - 2\eta\beta\sigma_i^2\right)^2 - 4(1 - \beta) \quad (94)$$

$$= \beta \left(4\beta\eta^2\sigma_i^4 + 4\beta\eta\sigma_i^2 + \beta - 8\eta\sigma_i^2\right). \quad (95)$$

To have distinctive and real eigenvalues,  $D > 0$

To have repeated real eigenvalues,  $D = 0$

To have complex eigenvalues,  $D < 0$

- (d) **For the case when they are repeated, what is the condition on  $\eta, \beta, \sigma_i$  that keeps them stable (strictly inside the unit circle)? What is the highest learning rate  $\eta$  as a function of  $\beta$  and  $\sigma_i$  that results in repeated eigenvalues?**

**Solution:** To find the maximum value, let's put the  $D = 0$ . Then, we can find two solutions:

$$\frac{2 - \beta - 2\sqrt{1 - \beta}}{2\beta\sigma_i^2} \text{ or } \frac{2 - \beta + 2\sqrt{1 - \beta}}{2\beta\sigma_i^2}$$

Therefore, the highest learning rate is  $\frac{2 - \beta + 2\sqrt{1 - \beta}}{2\beta\sigma_i^2}$

Plugging in learning rate to the eigenvalues, we have:

$$\lambda_{1,i} = \lambda_{2,i} = 1 - \beta\eta\sigma_i^2 - \frac{\beta}{2} = \pm\sqrt{1 - \beta} \quad (96)$$

If eigenvalues are repeated or complex, this system is always stable as  $0 < \beta < 1$ . This result is very surprising in that the convergence does not depend neither on the learning rate nor on the singular values of the covariate matrix.

- (e) **For the case when the eigenvalues are real, what is the condition on  $\eta, \beta, \sigma_i$  that keeps them stable (strictly inside the unit circle)? What is the range of the learning rate? Express with  $\beta, \sigma_i$**

**Solution:** At first, to have the different real roots, the discriminant Eq.(94),  $D$  should be strictly positive:

$$D = \beta \left(4\beta\eta^2\sigma_i^4 + 4\beta\eta\sigma_i^2 + \beta - 8\eta\sigma_i^2\right) > 0 \quad (97)$$

If we solve Eq.(97) with respect to  $\eta$ ,

$$\eta < \frac{2 - \beta - 2\sqrt{1 - \beta}}{2\beta\sigma_i^2} \text{ or } \eta > \frac{2 - \beta + 2\sqrt{1 - \beta}}{2\beta\sigma_i^2} \quad (98)$$

Also, to guarantee that those two solutions are in the open interval  $(0, 1)$ <sup>1</sup>, the characteristic equation in Eq.(92) should satisfy  $f(1) > 0$  and  $f(-1) > 0$ . We can check this by drawing the graph of  $f(\lambda)$ . Let's first compute  $f(1)$

$$\begin{aligned} f(1) &= 1 - \left(2 - \beta - 2\eta\beta\sigma_i^2\right) + 1 - \beta \\ &= 2\eta\beta\sigma_i^2 > 0 \end{aligned}$$

Therefore,  $f(1)$  is always positive.

Now, let's look at  $f(-1)$ .

$$\begin{aligned} f(-1) &= 1 + \left(2 - \beta - 2\eta\beta\sigma_i^2\right) + 1 - \beta \\ &= 4 - 2\beta - 2\eta\beta\sigma_i^2 > 0 \end{aligned}$$

Thus,

$$0 < \eta < \frac{4 - 2\beta}{2\beta\sigma_i^2} \quad (99)$$

Let's compare Eq.(98) and Eq.(99). Since

$$4 - 2\beta \geq 2 - \beta + 2\sqrt{1 - \beta}, \text{ for all } \beta \in [0, 1]$$

the condition that Eq.(92) has two different real roots is

$$0 < \eta < \frac{2 - \beta - 2\sqrt{1 - \beta}}{2\beta\sigma_i^2} \text{ or } \frac{2 - \beta + 2\sqrt{1 - \beta}}{2\beta\sigma_i^2} < \eta < \frac{4 - 2\beta}{2\beta\sigma_i^2}$$

- (f) **For the case when the eigenvalues are complex, what is the condition on  $\eta, \beta, \sigma_i$  that keeps them stable (strictly inside the unit circle)? What is the highest learning rate  $\eta$  as a function of  $\beta$  and  $\sigma_i$  that results in complex eigenvalues?**

**Solution:** If eigenvalues are repeated or complex, this system is always stable as  $0 < \beta < 1$ . This result is very surprising in that the convergence does not depend neither on the learning rate nor on the singular values of the covariate matrix

To find the maximum value, let's put  $D < 0$ . Then, we can find two solutions:

$$\frac{2 - \beta - 2\sqrt{1 - \beta}}{2\beta\sigma_i^2} < \eta < \frac{2 - \beta + 2\sqrt{1 - \beta}}{2\beta\sigma_i^2}$$

Therefore, the highest learning rate is  $\frac{2 - \beta + 2\sqrt{1 - \beta}}{2\beta\sigma_i^2}$ .

From the questions (d) (f), the takeaway is that we have to choose  $\eta$  such that every eigenvalues of  $R_i$  are either the same real roots or complex roots. Because the rate of convergence  $\max(|\lambda_1|, |\lambda_2|)$  is higher for those cases.

- (g) (This question might take more time than others) Now, apply what you have learned to the following problem. Assume that  $\beta = 0.1$  and we have a problem with two singular values  $\sigma_{\max}^2 = 5$  and  $\sigma_{\min}^2 = 0.05$ . **What learning rate  $\eta$  should we choose to get the fastest convergence for gradient**

---

<sup>1</sup>For stability, the norm should be less than 1

**descent with momentum? Compare how many iterations it will take to get within 99.9% of the optimal solution (starting at 0) using this learning rate and momentum with what it would take using ordinary gradient descent.**

**Solution:** For OGD, referring Eq.(85) the optimal convergence rate is  $R_1 = (5 - 0.05)/(5 + 0.05) = 0.98$ . Therefore, the minimum number of iterations ( $T$ ) to get within 99.9% of the optimal solution is:

$$\|w_{T_1} - w^*\|_2 = R_1^{T_1} \|w_0 - w^*\|_2 = \left(\frac{5 - 0.05}{5 + 0.05}\right)^{T_1} \|w^*\| \leq \epsilon \|w^*\|$$

$$T_1 = \lfloor \frac{\log 1000}{\log 505/495} \rfloor = 346$$

For GD+momentum, let's derive the optimal learning rate first. The optimal learning rate  $\eta^*$  and corresponding convergence rate  $R_2$  are:

$$\eta^* = \operatorname{argmin}_{\eta} \max \left\{ \left\| \begin{bmatrix} 1 - \beta & 2\beta\sigma_{\max}^2 \\ -\eta(1 - \beta) & 1 - 2\eta\beta\sigma_{\max}^2 \end{bmatrix} \right\|, \left\| \begin{bmatrix} 1 - \beta & 2\beta\sigma_{\min}^2 \\ -\eta(1 - \beta) & 1 - 2\eta\beta\sigma_{\min}^2 \end{bmatrix} \right\| \right\}$$

$$= \operatorname{argmin}_{\eta} \max \{ |\lambda_{1,\max}|, |\lambda_{2,\max}|, |\lambda_{1,\min}|, |\lambda_{2,\min}| \}$$

$$R_2 = \min_{\eta} \max \{ |\lambda_{1,\max}|, |\lambda_{2,\max}|, |\lambda_{1,\min}|, |\lambda_{2,\min}| \}$$

, where  $\|\cdot\|$  is the maximum of absolute value of eigenvalues and  $\lambda_{i,\max}, \lambda_{i,\min}$  are eigenvalues of matrices.

As you can see the above objective is seemingly very difficult to solve. But we can get some hints from the previous parts. First, Let's look at Eq.(92). The constant term is  $1 - \beta$ , which is the product of two solutions. If one of solutions is  $\lambda$ , the other one is  $\frac{1-\beta}{\lambda}$ . In other words, if  $\lambda < \sqrt{1-\beta}$ , the other solution is larger than  $\sqrt{1-\beta}$ . Then  $R_2$  cannot be smaller than  $\sqrt{1-\beta}$ . Therefore  $R_2$  has the lower bound,  $R_2 \geq \sqrt{1-\beta}$ .

Then, let's hypothesize that we can achieve that lower bound. Remind that if two solutions are repeated real or complex, their absolute values are always  $\sqrt{1-\beta}$ . So to prove the hypothesis, we need to show that eigenvalues of each matrix are either repeated real or complex. In that case, Eq.(94) should be non-positive for both matrices.

$$\frac{2 - \beta - 2\sqrt{1-\beta}}{2\beta\sigma_{\max}^2} \leq \eta \leq \frac{2 - \beta + 2\sqrt{1-\beta}}{2\beta\sigma_{\max}^2}$$

$$\frac{2 - \beta - 2\sqrt{1-\beta}}{2\beta\sigma_{\min}^2} \leq \eta \leq \frac{2 - \beta + 2\sqrt{1-\beta}}{2\beta\sigma_{\min}^2}$$

Inserting  $\beta, \sigma_i^2$ , the above inequalities, we can get the following result:

$$0.002633 \leq \eta \leq 3.797$$

$$0.2633 \leq \eta \leq 379.7$$

The common interval that  $\eta$  satisfies both inequalities is

$$0.2633 \leq \eta \leq 3.797$$

If  $\eta$  is in that interval, the convergence rate is  $\sqrt{1-\beta} = 0.949$

$$\|w_{T_2} - w^*\|_2 \leq R_2^{T_2} \|w_0 - w^*\|_2 = (0.949)^{T_2} \|w^*\| \leq \epsilon \|w^*\|$$

$$T_2 \geq \left\lfloor \frac{\log 1/\epsilon}{\log 1/R_2} \right\rfloor = 132$$

We need at least 132 iterations to guarantee 0.1% error.

Note: if you find the convergence rate correctly  $\sqrt{1-\beta}$  and derive any  $\eta$  in the interval  $[0.263, 3.80]$ , then you can get the full credit for this part.

- (h) The 2 questions below are based on the Jupyter Notebook given in [the notebook](#). Please open the corresponding notebook and follow the instructions to answer the following questions. You don't need to submit the `ipynb` file.

**How does  $\sigma_i$  (the eigenvalues) influence the gradients and parameters updates?**

**Solution:** Dimension 0 has larger  $\sigma$ /eigenvalue, so the gradient is larger. With a relatively large stepsize (what we choose here), the gradients and parameters are oscillating a bit at the beginning before converging.

- (i) Question: Comparing gradient descent and gradient descent with momentum, **which one converges faster for this task? Why?**

**Solution:** Gradient descent with momentum is faster in convergence compared to traditional gradient descent. This is because gradient descent with momentum adjusts the magnitude of the parameter update in each dimension, allowing the optimizer to make bigger and more confident updates in the dimensions where the gradients are pointing in the same direction, and smaller updates in the dimensions where the gradients are oscillating. This leads to a more efficient optimization process and faster convergence to the optimal solution.

**Solution:** Another solution, which tries to avoid as much annoying algebra as possible. This solution uses the companion notebook [the companion notebook](#) or [the companion video](#).

Start the same way, until we get to the quadratic equation for the eigenvalues:

$$\lambda^2 - (2 - \beta - 2\beta(\eta\sigma^2))\lambda + 1 - \beta = 0$$

By Vieta's formula,

$$\lambda_1 \lambda_2 = 1 - \beta, \quad \lambda_1 + \lambda_2 = 2 - \beta - 2\beta(\eta\sigma^2)$$

- If we draw  $\lambda_1, \lambda_2$  as two vectors on the complex plane, we can see that if  $\lambda_1 = \lambda_2$  or  $\lambda_1, \lambda_2$  are complex, then they are two vectors on the circle of radius  $\sqrt{1-\beta}$ . Otherwise, they are both parallel to the real axis, but one of them is outside the circle and one inside.

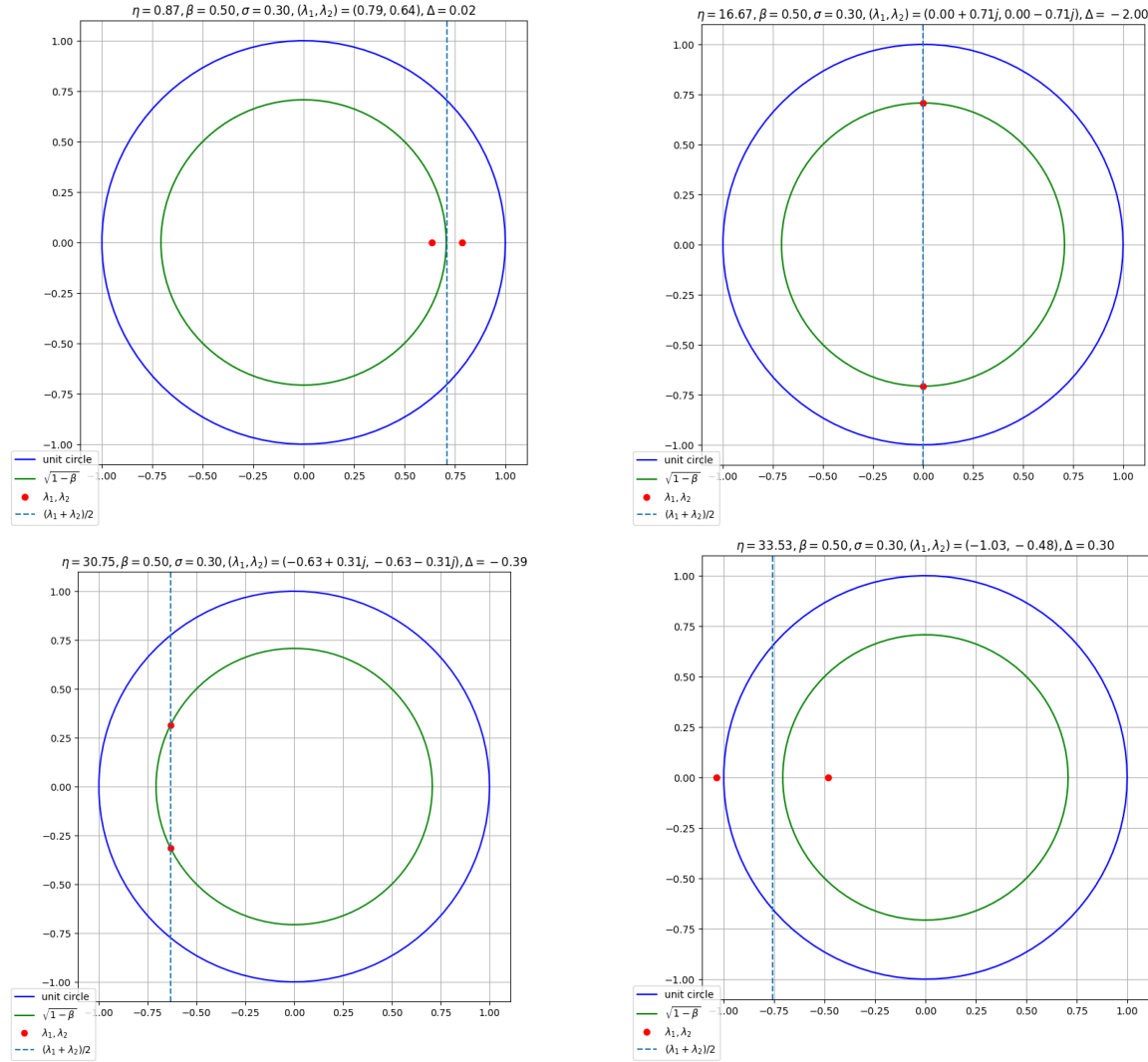
Let  $t := \eta\sigma^2$ . As  $t$  increases, the mid-point  $\frac{\lambda_1 + \lambda_2}{2}$  monotonically moves from  $+\infty$  to  $-\infty$ , and the two eigenvalues are on the  $\sqrt{1-\beta}$  circle exactly when  $\frac{\lambda_1 + \lambda_2}{2}$  falls on the diameter, that is,

$$\frac{\lambda_1 + \lambda_2}{2} \in [-\sqrt{1-\beta}, +\sqrt{1-\beta}]$$

that is,

$$t \in \left[ \frac{(1 \pm \sqrt{1-\beta})^2}{2\beta} \right]$$

**Figure 2:** Four examples, with different  $\eta$ , showing different behavior of the eigenvalues. Figures generated by the companion notebook.



- A simultaneous solution for all singular values exist, iff  $\sigma_{min}, \sigma_{max}$  are such that

$$\frac{\sigma_{max}}{\sigma_{min}} \leq \frac{1 + \sqrt{1-\beta}}{1 - \sqrt{1-\beta}}$$

As  $\beta$  increases from 0 to 1, the minimal achievable rate of convergence drops further and further, from 1 to 0, but there is a price to pay: the condition

$$\frac{\sigma_{max}}{\sigma_{min}} \leq \frac{1 + \sqrt{1-\beta}}{1 - \sqrt{1-\beta}}$$

becomes harder and harder to satisfy, dropping from  $+\infty$  down to 1.

So, we have a tradeoff: higher  $\beta$  makes the best rate of convergence smaller, but also makes it less likely to achieve the best rate of convergence.

The whole process can be easily seen in Figure 2. As  $\eta$  increases, first the two eigenvalues are real and on the positive axis, then they move to the circle of radius  $\sqrt{1-\beta}$  and finally they move to the

negative axis. Their motion is "puppeteered" by the dashed line showing  $\frac{\lambda_1 + \lambda_2}{2}$ , which starts to the right side of the circle, then it intersects the circle, then it moves off to the left.

Numerical example:

$\beta = 0.1$ , then the condition is

$$t \in [0.013, 18.98]$$

Simultaneous solution exists iff  $(\frac{\sigma_{max}}{\sigma_{min}})^2 \leq 1442$ , in which case we have convergence rate  $\sqrt{0.9} = 0.948$ .

Now compare this with the minimal rate of convergence for gradient descent,

$$\frac{(\sigma_{max}/\sigma_{min})^2 - 1}{(\sigma_{max}/\sigma_{min})^2 + 1} = 1 - \frac{2}{(\sigma_{max}/\sigma_{min})^2 + 1}$$

To make the comparison as dramatic as possible, we can use  $(\frac{\sigma_{max}}{\sigma_{min}})^2 = 1442$ , which makes the plain gradient descent have rate of convergence 0.9986.

Now, we have  $0.948 \approx 0.9986^{38}$ , meaning that each step of the optimal momentum gradient descent is equivalent to 38 steps of optimal gradient descent.

## 4. Optimizers

---

### Algorithm 1 SGD with Momentum

---

```

1: Given  $\eta = 0.001, \beta_1 = 0.9$ 
2: Initialize:
3:   time step  $t \leftarrow 0$ 
4:   parameter  $\theta_{t=0} \in \mathbb{R}^n$ 
5: Repeat
6:    $t \leftarrow t + 1$ 
7:    $g_t \leftarrow \nabla f_t(\theta_{t-1})$ 
8:    $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$ 
9:    $\theta_t \leftarrow \theta_{t-1} - \eta m_t$ 
10: Until the stopping condition is met
```

---



---

### Algorithm 2 Adam Optimizer (without bias correction)

---

```

1: Given  $\eta = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$ 
2: Initialize time step  $t \leftarrow 0$ , parameter  $\theta_{t=0} \in \mathbb{R}^n$ ,
    $m_{t=0} \leftarrow 0, v_{t=0} \leftarrow 0$ 
3: Repeat
4:    $t \leftarrow t + 1$ 
5:    $g_t \leftarrow \nabla f_t(\theta_{t-1})$ 
6:    $m_t \leftarrow \text{---(A)---}$ 
7:    $v_t \leftarrow \text{---(B)---}$ 
8:    $\theta_t \leftarrow \theta_{t-1} - \eta \cdot \frac{m_t}{\sqrt{v_t}}$ 
9: Until the stopping condition is met
```

---

(a) Complete part (A) and (B) in the pseudocode of Adam.

**Solution:**

$$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

(b) This question asks you to establish the relationship between

- **L2 regularization** for vector-valued weights  $\theta$  refers to adding a squared Euclidean norm of the weights to the loss function itself:

$$f_t^{reg} = f_t(\theta) + \frac{\lambda}{2} \|\theta\|_2^2$$

- **Weight decay** refers to explicitly introducing a scalar  $\gamma$  in the weight updates assuming loss  $f$ :

$$\theta_{t+1} = (1 - \gamma)\theta_t - \eta \nabla f(\theta_t)$$

where  $\gamma = 0$  would correspond to regular SGD since it has no weight-decay.

**Show that SGD with weight decay using the original loss  $f_t(\theta)$  is equivalent to regular SGD on the L2-regularized loss  $f_t^{reg}(\theta)$  when  $\gamma$  is chosen correctly, and find such a  $\gamma$  in terms of  $\lambda$  and  $\eta$ .**

**Solution:**

$$f_t^{reg}(\theta) = f_t(\theta) + \frac{\lambda}{2} \|\theta\|_2^2 \implies \nabla f_t^{reg}(\theta) = \nabla f_t(\theta) + \lambda \theta$$

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla f_t^{reg}(\theta_t) = \theta_t - \eta \nabla f_t - \eta \lambda \theta_t = (1 - \eta \lambda) \theta_t - \eta \nabla f_t(\theta_t)$$

The process above shows that L2 regularization and weight decay are equivalent when  $\gamma = \lambda \eta$ .

## 5. Regularization and Instance Noise

Say we have  $m$  labeled data points  $(\mathbf{x}_i, y_i)_{i=1}^m$ , where each  $\mathbf{x}_i \in \mathbb{R}^n$  and each  $y_i \in \mathbb{R}$ . We perform data augmentation by adding some noise to each vector every time we use it in SGD. This means for all points  $i$ , we have a true input  $\mathbf{x}_i$  and add noise  $\mathbf{N}_i$  to get the effective random input seen by SGD:

$$\tilde{\mathbf{X}}_i = \mathbf{x}_i + \mathbf{N}_i$$

The i.i.d. random noise vectors  $\mathbf{N}_i$  are distributed as  $\mathbf{N}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ .

We can conceptually arrange these noise-augmented data points into a random matrix  $\tilde{X} \in \mathbb{R}^{m \times n}$ , where row  $\tilde{\mathbf{X}}_i^\top$  represents one augmented datapoint. Similarly we arrange the labels  $y_i$  into a vector  $\mathbf{y}$ .

$$\tilde{X} = \begin{bmatrix} \tilde{\mathbf{X}}_1^\top \\ \tilde{\mathbf{X}}_2^\top \\ \dots \\ \tilde{\mathbf{X}}_m^\top \end{bmatrix}, \text{ where } \tilde{\mathbf{X}}_i \in \mathbb{R}^n, \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} \in \mathbb{R}^m$$

One way of thinking about what SGD might do is to consider learning weights that minimize the *expected* least squares objective for the **noisy** data matrix:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \mathbb{E}[\|\tilde{X}\mathbf{w} - \mathbf{y}\|^2] \quad (100)$$

- (a) **Show that this problem (100) is equivalent to a regularized least squares problem:**

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{m} \|\tilde{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2 \quad (101)$$

You will need to determine the value of  $\lambda$ .

*Hint: write the squared norm of a vector as an inner product, expand, and apply linearity of expectation.*

**Solution:** Following the hint, we write our objective as

$$\begin{aligned}
\mathbb{E}[\|\tilde{X}\mathbf{w} - \mathbf{y}\|^2] &= \mathbb{E}\left[\sum_{i=1}^m (\tilde{\mathbf{x}}_i^\top \mathbf{w} - y_i)^2\right] \\
&= \sum_{i=1}^m \mathbb{E}[(\mathbf{x}_i + \mathbf{N}_i)^\top \mathbf{w} - y_i]^2 \\
&= \sum_{i=1}^m \mathbb{E}[(\mathbf{x}_i^\top \mathbf{w} + \mathbf{N}_i^\top \mathbf{w} - y_i)^2] \\
&= \sum_{i=1}^m \mathbb{E}[(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 - 2(\mathbf{N}_i^\top \mathbf{w})(\mathbf{x}_i^\top \mathbf{w} - y_i) + (\mathbf{N}_i^\top \mathbf{w})^2] \\
&= \sum_{i=1}^m \mathbb{E}[(\mathbf{x}_i^\top \mathbf{w} - y_i)^2] - 2\mathbb{E}[(\mathbf{N}_i^\top \mathbf{w})(\mathbf{x}_i^\top \mathbf{w} - y_i)] + \mathbb{E}[\mathbf{w}^\top \mathbf{N}_i \mathbf{N}_i^\top \mathbf{w}] \\
&= \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 - 2\mathbb{E}[(\mathbf{N}_i^\top \mathbf{w})](\mathbf{x}_i^\top \mathbf{w} - y_i) + \mathbf{w}^\top \mathbb{E}[\mathbf{N}_i \mathbf{N}_i^\top] \mathbf{w} \\
&= \sum_{i=1}^m [(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 - 0 + \mathbf{w}^\top \sigma^2 I_n \mathbf{w}] \\
&= \|X\mathbf{w} - \mathbf{y}\|^2 + m\sigma^2 \|\mathbf{w}\|^2
\end{aligned}$$

Dividing through by  $m$  to match the desired form of (101) gives us  $\lambda = \sigma^2$ .

Now consider a simplified example where we only have a single scalar datapoint  $x \in \mathbb{R}$  and its corresponding label  $y \in \mathbb{R}$ . We are going to analyze this in the context of gradient descent. For the  $t$ -th step of gradient descent, we use a noisy datapoint  $\tilde{X}_t = x + N_t$  which is generated by adding different random noise values  $N_t \sim \mathcal{N}(0, \sigma^2)$  to our underlying data point  $x$ . The noise values for each iteration of gradient descent are i.i.d. We want to learn a weight  $w$  such that the squared-loss function  $\mathcal{L}(w) = \frac{1}{2}(\tilde{X}w - y)^2$  is minimized. We initialize our weight to be  $w_0 = 0$ .

- (b) Let  $w_t$  be the weight learned after the  $t$ -th iteration of gradient descent with data augmentation. **Write the gradient descent recurrence relation between  $\mathbb{E}[w_{t+1}]$  and  $\mathbb{E}[w_t]$  in terms of  $x$ ,  $\sigma^2$ ,  $y$ , and learning rate  $\eta$ .**

**Solution:** We can first compute the gradient in this case:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= (w\tilde{X} - y)\tilde{X} \\
&= (w(x + N_t) - y)(x + N_t) \\
&= w(x^2 + 2xN_t + N_t^2) - y(x + N_t).
\end{aligned}$$

And now, we can use it for gradient descent with stepsize  $\eta$ :

$$\begin{aligned}
w_{t+1} &= w_t - \eta \nabla_w \mathcal{L}(w_t) \\
&= w_t - \eta (w_t(x + N_t)^2 - y(x + N_t)) \\
&= w_t(1 - \eta(x^2 + 2xN_t + N_t^2)) + \eta y(x + N_t)
\end{aligned}$$

Taking expectations, and using the fact that  $N_t$  is independent of  $w_t$  since the noise values are i.i.d., we get:

$$\mathbb{E}[w_{t+1}] = \mathbb{E}[w_t](1 - \eta(x^2 + \sigma^2)) + \eta yx$$

- (c) **For what values of learning rate  $\eta$  do we expect the expectation of the learned weight to converge using gradient descent?**

**Solution:** For gradient descent to converge, we need the coefficient on the recurrent term to be between -1 and 1.

$$-1 < 1 - \eta(x^2 + \sigma^2) \text{ and } 1 - \eta(x^2 + \sigma^2) < 1$$

This implies

$$0 < \eta < \frac{2}{x^2 + \sigma^2}$$

- (d) Assuming that we are in the range of  $\eta$  for which gradient-descent converges, **what would we expect  $\mathbb{E}[w_t]$  to converge to as  $t \rightarrow \infty$ ? How does this differ from the optimal value of  $w$  if there were no noise being used to augment the data?**

(HINT: You can also use this to help check your work for part (a).)

**Solution:** One way of doing this is to take an expectation of the gradient and set it to 0. Using the fact that  $\mathbb{E}[N_t] = 0$  and  $\mathbb{E}[N_t^2] = \sigma^2$ ,

$$w^*(x^2 + \sigma^2) - yx = 0$$

$$\begin{aligned} w^* &= \frac{yx}{x^2 + \sigma^2} \\ &= \left(\frac{y}{x}\right) \frac{1}{1 + \frac{\sigma^2}{x^2}} \end{aligned}$$

The optimal value of  $w$  if there was no data-augmenting noise would simply be  $w = \frac{y}{x}$ . This means the optimal expected value with noise augmentation is scaled down by a factor of  $\frac{1}{1 + \frac{\sigma^2}{x^2}}$ . This looks just like the “shrinkage” term we saw on the eigenvalues with Ridge regression. When  $\sigma$  is small relative to  $x$ , the learned weight barely changes. When  $\sigma$  is large relative to  $x$ , the learned weight shrinks significantly towards zero.

Part (a) has solution  $w^* = (X^T X + m\sigma^2 I)^{-1} X^T y$ , which reduces to this when  $m = 1$ .

## 6. General Case Tikhonov Regularization

Consider the optimization problem:

$$\min_{\mathbf{x}} ||W_1(A\mathbf{x} - \mathbf{b})||_2^2 + ||W_2(\mathbf{x} - \mathbf{c})||_2^2$$

Where  $W_1$ ,  $A$ , and  $W_2$  are matrices and  $\mathbf{x}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  are vectors.  $W_1$  can be viewed as a generic weighting of the residuals and  $W_2$  along with  $c$  can be viewed as a generic weighting of the parameters.

- (a) **Solve this optimization problem manually** by expanding it out as matrix-vector products, setting the gradient to 0, and solving for  $\mathbf{x}$ .

**Solution:** Expand our objective function:

$$f(\mathbf{x}) = (\mathbf{Ax} - \mathbf{b})^T W_1^T W_1 (\mathbf{Ax} - \mathbf{b}) + (\mathbf{x} - \mathbf{c})^T W_2^T W_2 (\mathbf{x} - \mathbf{c})$$

$$f(\mathbf{x}) = \mathbf{x}^T A^T W_1^T W_1 A \mathbf{x} - 2\mathbf{b}^T W_1^T W_1 A \mathbf{x} + \mathbf{b}^T W_1^T W_1 \mathbf{b} + \mathbf{x}^T W_2^T W_2 \mathbf{x} - 2\mathbf{c}^T W_2^T W_2 \mathbf{x} + \mathbf{c}^T W_2^T W_2 \mathbf{c}$$

Now take gradients and set it to  $\mathbf{0}$ :

$$\nabla f = 2A^T W_1^T W_1 A \mathbf{x} - 2A^T W_1^T W_1 \mathbf{b} + 2W_2^T W_2 \mathbf{x} - 2W_2^T W_2 \mathbf{c} = \mathbf{0}$$

Isolating the  $\mathbf{x}$  terms on one side, we have:

$$(A^T W_1^T W_1 A + W_2^T W_2) \mathbf{x} = A^T W_1^T W_1 \mathbf{b} + W_2^T W_2 \mathbf{c}$$

So we can solve to get

$$\mathbf{x} = (A^T W_1^T W_1 A + W_2^T W_2)^{-1} (A^T W_1^T W_1 \mathbf{b} + W_2^T W_2 \mathbf{c})$$

(b) **Construct an appropriate matrix  $\mathbf{C}$  and vector  $\mathbf{d}$  that allows you to rewrite this problem as**

$$\min_x \|\mathbf{Cx} - \mathbf{d}\|^2$$

**and use the OLS solution  $(\mathbf{x}^* = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{d})$  to solve.** Confirm your answer is in agreement with the previous part.

**Solution:** We can rewrite our problem in least-squares form using

$$\mathbf{C} = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}, \text{ and } \mathbf{d} = \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix}$$

Now, using least squares solution and solving, we get

$$\begin{aligned} \mathbf{x}^* &= (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{d} = \left( \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}^T \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}^T \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix} \\ &= \left( \begin{bmatrix} A^T W_1^T & W_2^T \end{bmatrix} \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} A^T W_1^T & W_2^T \end{bmatrix} \begin{bmatrix} W_1 \mathbf{b} \\ W_2 \mathbf{c} \end{bmatrix} \\ \mathbf{x}^* &= (A^T W_1^T W_1 A + W_2^T W_2)^{-1} (A^T W_1^T W_1 \mathbf{b} + W_2^T W_2 \mathbf{c}) \end{aligned}$$

Which is the same as the previous part, as desired.

(c) **Choose a  $W_1$ ,  $W_2$ , and  $\mathbf{c}$  such that this reduces to the simple case of ridge regression that you've seen in the previous problem,  $\mathbf{x}^* = (A^T A + \lambda I)^{-1} A^T \mathbf{b}$ .**

**Solution:** This reduces to ridge regression when  $W_1 = I$ ,  $W_2 = \sqrt{\lambda} I$ , and  $\mathbf{c} = \mathbf{0}$ . You can see this in both the optimization problem and the result.

## 7. An Alternate MAP Interpretation of Ridge Regression

Consider the Ridge Regression estimator,

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|^2$$

We know this is solved by

$$\hat{\mathbf{w}} = (X^T X + \lambda I)^{-1} X^T \mathbf{y} \quad (102)$$

An alternate form of the Ridge Regression solution (often called the Kernel Ridge form) is given by

$$\hat{\mathbf{w}} = X^T (X X^T + \lambda I)^{-1} \mathbf{y}. \quad (103)$$

We know that Ridge Regression can be viewed as finding the MAP estimate when we apply a prior on the (now viewed as random parameters)  $\mathbf{W}$ . In particular, we can think of the prior for  $\mathbf{W}$  as being  $\mathcal{N}(\mathbf{0}, I)$  and view the random  $Y$  as being generated using  $Y = \mathbf{x}^T \mathbf{W} + \sqrt{\lambda} N$  where the noise  $N$  is distributed iid (across training samples) as  $\mathcal{N}(0, 1)$ . At the vector level, we have  $\mathbf{Y} = X\mathbf{W} + \sqrt{\lambda}\mathbf{N}$ , and then we know that when we try to maximize the log likelihood we end up minimizing

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{\lambda} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \|\mathbf{w}\|^2 = \underset{\mathbf{w}}{\operatorname{argmin}} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|^2.$$

The underlying probability space is that defined by the  $d$  iid standard normals that define the  $\mathbf{W}$  and the  $n$  iid standard normals that give the  $n$  different  $N_i$  on the training points. Note that the  $X$  matrix whose rows consist of the  $n$  different inputs for the  $n$  different training points are not random.

Based on what we know about joint normality, it is clear that the random Gaussian vectors  $\mathbf{W}$  and  $\mathbf{Y}$  are jointly normal. **Use the following facts to show that the two forms of solution are identical.**

- (102) is the MAP estimate for  $\mathbf{W}$  given an observation  $\mathbf{Y} = \mathbf{y}$  (We showed this in HW1 last week, and in discussion section)
- For jointly normal random variables, when you condition one set of variables on the values for the others, the resulting conditional distribution is still normal.
- A normal random variable has its density maximized at its mean.
- For jointly normal random vectors that are zero mean, the formula for conditional expectation is

$$E[\mathbf{W}|\mathbf{Y} = \mathbf{y}] = \Sigma_{WY} \Sigma_Y^{-1} \mathbf{y} \quad (104)$$

where the  $\Sigma_{YY}$  is the covariance  $E[\mathbf{Y}\mathbf{Y}^T]$  of  $\mathbf{Y}$  and  $\Sigma_{WY} = E[\mathbf{W}\mathbf{Y}^T]$  is the appropriate cross-covariance of  $\mathbf{W}$  and  $\mathbf{Y}$ .

### Solution:

We are given that (102) is  $\operatorname{MAP}(\mathbf{w} | \mathbf{Y} = \mathbf{y})$ . Let's try to find  $\operatorname{MAP}(\mathbf{w} | \mathbf{Y} = \mathbf{y})$  in a different way.

We can condition the random variables  $\mathbf{W}$  on the values for  $\mathbf{Y} = \mathbf{y}$ , and we know the resulting conditional distribution  $\mathbf{W}|\mathbf{Y} = \mathbf{y}$  is still normal.

Since normal random variables has density maximized at the mean, the MAP estimate is equivalent to the conditional expectation  $E[\mathbf{W}|\mathbf{Y} = \mathbf{y}]$ . We are given the formula to calculate the conditional expectation.

Before we plug into the formula, let's calculate  $\Sigma_{WY} = E[\mathbf{W}\mathbf{Y}^T]$ . Using  $\mathbf{Y} = X\mathbf{W} + \mathbf{N}$ , we have:

$$\Sigma_{WY} = E[\mathbf{W}(X\mathbf{W} + \sqrt{\lambda}\mathbf{N})^T]$$

Take transposes and distribute, apply linearity of expectation

$$\Sigma_{WY} = E[\mathbf{W}\mathbf{W}^T X^T] + \sqrt{\lambda}E[\mathbf{W}\mathbf{N}^T]$$

Since  $\mathbf{W}$  and  $\mathbf{N}$  are uncorrelated, and  $X$  does not involve any randomness, we can write

$$\Sigma_{WY} = E[\mathbf{W}\mathbf{W}^T]X^T + \sqrt{\lambda}E[\mathbf{W}]E[\mathbf{N}^T]$$

Use the fact that  $E[\mathbf{W}\mathbf{W}^T]$  is the covariance matrix of  $\mathbf{W}$ , which we are given is the identity. Also use the fact that  $\mathbf{W}$  and  $\mathbf{N}$  are zero-meaned.

$$\Sigma_{WY} = IX^T + 0 = X^T$$

Now let's find  $\Sigma_{YY} = E[\mathbf{Y}\mathbf{Y}^T]$ .

$$\Sigma_{YY} = E[(X\mathbf{W} + \mathbf{N})(X\mathbf{W} + \mathbf{N})^T]$$

Take transposes and distribute

$$\Sigma_{YY} = E[X\mathbf{W}\mathbf{W}^T X^T] + \sqrt{\lambda}^2 E[\mathbf{N}\mathbf{N}^T] + \sqrt{\lambda}E[X\mathbf{W}\mathbf{N}^T] + \sqrt{\lambda}E[\mathbf{N}\mathbf{W}^T X^T]$$

The cross-terms are 0 since the random vectors are zero-meaned. With some manipulation, we have

$$\Sigma_{YY} = XE[\mathbf{W}\mathbf{W}^T]X^T + \lambda I$$

$$\Sigma_{YY} = XX^T + \lambda I$$

Finally, plugging in  $\Sigma_{WY}$  and  $\Sigma_{YY}$  to the formula for  $E[\mathbf{W}|\mathbf{Y} = \mathbf{y}]$ , we get  $MAP(\mathbf{w}|\mathbf{Y} = \mathbf{y})$  as

$$\hat{\mathbf{w}} = X^T(XX^T + \lambda I)^{-1}\mathbf{y}$$

as desired.

## 8. Homework Process and Study Group

Citing sources and collaborators are an important part of life, including being a student!

We also want to understand what resources you find helpful and how much time homework is taking, so we can change things in the future if possible.

- (a) **What sources (if any) did you use as you worked through the homework?**
- (b) **If you worked with someone on this homework, who did you work with?**  
List names and student ID's. (In case of homework party, you can also just describe the group.)
- (c) **Roughly how many total hours did you work on this homework?**

### Contributors:

- Anant Sahai.
- Sheng Shen.
- Gireeja Ranade.

- Yaodong Yu.
- Suhong Moon.
- Gabriel Goh.
- Peter Wang.
- Yuxi Liu.
- Kevin Li.
- Saagar Sanghavi.