HW3

Kevin Chang

October 15, 2025

1

Recall that for a vector $w \in \mathbb{R}^d$, $\mathcal{H}_w := \{z :< w, z >= 0\}$. Let $S = \{(x_i, y_i)\}$ be a set of linearly separable data in $\mathbb{R}^d$ (i.e., $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$). Define the set $\mathcal{M}_S$ to be the set of all vectors which separate the data with large dot product:

$$\mathcal{M}_S = \{w : y_i \langle w, x_i \rangle \geq 1 \text{ for } i = 1, ..., n\}.$$

- Let w^* denote the element of $\mathcal{M}_S$ with smallest norm.

  Show that for any other w that separates the data

  $$\min dist_{1 \leq i \leq n}(x_i, \mathcal{H}_w) \leq \min_{1 \leq i \leq n} dist(x_i, \mathcal{H}_{w^*}).$$

  Proof. For any nonzero w, $\text{dist}(x, \mathcal{H}_w) = \frac{|\langle w, x \rangle|}{\|w\|}$. Let $\gamma_w := \min_i y_i \langle w, x_i \rangle > 0$ be the margin of w. Then

  $$\min_i \text{dist}(x_i, \mathcal{H}_w) = \frac{\min_i |\langle w, x_i \rangle|}{\|w\|} = \frac{\gamma_w}{\|w\|}.$$

  Scale w to $\tilde{w} := \frac{w}{\gamma_w}$; then $\tilde{w} \in \mathcal{M}_S$ and $\|\tilde{w}\| = \frac{\|w\|}{\gamma_w}$. By optimality of w^* on $\mathcal{M}_S$, $\|w^*\| \leq \|\tilde{w}\| = \frac{\|w\|}{\gamma_w}$, hence

  $$\frac{\gamma_w}{\|w\|} \leq \frac{1}{\|w^*\|} = \min_i \text{dist}(x_i, \mathcal{H}_{w^*}),$$

  where the last equality uses $\min_i y_i \langle w^*, x_i \rangle = 1$ (if it were > 1, rescaling down would contradict minimality of $\|w^*\|$). This proves the claim. □

- Show that there are real numbers α_i such that $w^* = \sum_{i=1}^n \alpha_i x_i$.

  Proof. Suppose, for the sake of contradiction, that w^* cannot be written as a linear combination of the training examples $\{x_i\}$. Then there exists a decomposition

  $$w^* = \sum_{i=1}^n \alpha_i x_i + v,$$

  where v is orthogonal to all x_i, i.e., $\langle v, x_i \rangle = 0$ for all i.

  Define $w' := \sum_{i=1}^n \alpha_i x_i$. Since $\langle v, x_i \rangle = 0$, we have

  $$y_i \langle w', x_i \rangle = y_i \langle w^*, x_i \rangle \geq 1,$$

  which implies $w' \in \mathcal{M}_S$.

  Moreover, $\|w'\| < \|w^*\|$ because $w^* = w' + v$ and $v \neq 0$ adds an orthogonal component, increasing the norm:

  $$\|w^*\|^2 = \|w'\|^2 + \|v\|^2 > \|w'\|^2.$$

  This contradicts the minimality of w^* as the smallest–norm element in $\mathcal{M}_S$. Hence, such a v cannot exist, and therefore

  $$w^* = \sum_{i=1}^n \alpha_i x_i.$$

  □

- Show that the α_i can be chosen so that $y_i\alpha_i$ are all nonnegative.

  Define $\tilde{\alpha}_i := \alpha_i y_i$. Then

$$w^* = \sum_{i=1}^{n} \tilde{\alpha}_i x_i, \quad \text{and} \quad y_i\tilde{\alpha}_i = y_i^2\alpha_i = \alpha_i \geq 0.$$

  Hence, the coefficients can be chosen so that $y_i\alpha_i \geq 0$ for all i.

2

Let u and v be D-dimensional unit vectors. Let M be a random matrix of dimension $d \times D$. Each entry of M is generated iid from a normal distribution with mean 0 and variance $1/d$.

1. Show that $\mathbb{E}[\langle Mu, Mv\rangle] = \langle u, v\rangle$.

   Answer:

$$\mathbb{E}\big[\langle Mu, Mv\rangle\big] = \mathbb{E}\big[u^\top M^\top Mv\big] = u^\top\, \mathbb{E}[M^\top M]\, v = u^\top I_D v = \langle u, v\rangle.$$

   (We used $\mathbb{E}[M^\top M] = I_D$ since each row of M has covariance $\frac{1}{d}I_D$ and there are d rows.)

2. Suppose $d \geq \frac{8}{\epsilon^2}$. Show that with probability at least $1 - e^{-1} - e^{-2}$,

$$\langle Mu, Mv\rangle \geq \langle u, v\rangle - \epsilon.$$

   Answer: Using the polarization identity,

$$\langle Mu, Mv\rangle = \tfrac{1}{4}\big(\|M(u+v)\|^2 - \|M(u-v)\|^2\big)\,.$$

   Let $z_\pm := u \pm v$. Then $\|Mz_\pm\|^2 = \frac{1}{d}\sum_{j=1}^{d} g_{j,\pm}^2$ with $g_{j,\pm} \sim \mathcal{N}(0, \|z_\pm\|^2)$ i.i.d., so $\frac{\|Mz_\pm\|^2}{\|z_\pm\|^2}$ is the average of d i.i.d. $\mathcal{N}(0,1)^2$ variables. By the bounds, for any $\varepsilon' > 0$,

$$\Pr\big[\big|\|Mz_+\|^2 - \|z_+\|^2\big| > \varepsilon'\|z_+\|^2\big] \leq e^{-d\varepsilon'^2/8},$$

$$\Pr\big[\big|\|Mz_-\|^2 - \|z_-\|^2\big| > \varepsilon'\|z_-\|^2\big] \leq e^{-d\varepsilon'^2/4}.$$

   On the intersection of these two events,

$$\big|\langle Mu, Mv\rangle - \langle u, v\rangle\big| = \tfrac{1}{4}\big|\big(\|Mz_+\|^2 - \|z_+\|^2\big) - \big(\|Mz_-\|^2 - \|z_-\|^2\big)\big| \leq \varepsilon'.$$

   Taking $\varepsilon' = \varepsilon$ and using the union bound gives

$$\Pr[\langle Mu, Mv\rangle \geq \langle u, v\rangle - \varepsilon] \geq 1 - e^{-d\varepsilon^2/8} - e^{-d\varepsilon^2/4}.$$

   If $d \geq 8/\varepsilon^2$, then $e^{-d\varepsilon^2/8} \leq e^{-1}$ and $e^{-d\varepsilon^2/4} \leq e^{-2}$, hence

$$\Pr[\langle Mu, Mv\rangle \geq \langle u, v\rangle - \varepsilon] \;\geq\; 1 - e^{-1} - e^{-2}.$$

3. Now let's apply this to machine learning. Consider a set of n examples in D dimensional space that is linearly separable with margin y. That is, there are n examples, (x_i, y_i) with $y_i \in \{-1, 1\}$ and $||x_i|| \leq R$, and there is a unit vetor w so that $y_i\langle w, x_i\rangle \geq y$ for all i.

   Suppose that

$$d \geq 32\frac{R^2}{\gamma^2}\log(4n).$$

   Show that with probability at least $1/2$, $y_i\langle Mw, Mx_i\rangle \geq \frac{\gamma}{2}$ for all i. We can think of the vectors Mx_i as embeddings of the original data set in a lower dimensional space. This problem shows a random embedding already preserves much of the linear separability of data. An optimized embedding can do only better.

Answer: Assume a linearly separable dataset $\{(x_i, y_i)\}_{i=1}^n$ with $\|x_i\| \le R$, labels $y_i \in \{\pm 1\}$, and a unit vector w such that $y_i\langle w, x_i\rangle \ge \gamma$ for all i. Fix i and apply part (2) with the unit pair $u = w$ and $v = \frac{x_i}{\|x_i\|}$. With probability at least $1 - e^{-d\varepsilon^2/8} - e^{-d\varepsilon^2/4}$,

$$\langle Mw, M(x_i/\|x_i\|)\rangle \;\ge\; \langle w, x_i/\|x_i\|\rangle - \varepsilon.$$

Multiplying by $\|x_i\|$ and then by y_i yields

$$y_i\langle Mw, Mx_i\rangle \;\ge\; y_i\langle w, x_i\rangle - \|x_i\|\,\varepsilon \;\ge\; \gamma - R\varepsilon.$$

Choose $\varepsilon = \gamma/(2R)$. Then for this i,

$$\Pr[y_i\langle Mw, Mx_i\rangle \ge \gamma/2] \;\ge\; 1 - e^{-d\gamma^2/(32R^2)} - e^{-d\gamma^2/(16R^2)}.$$

If

$$d \;\ge\; 32\,\frac{R^2}{\gamma^2}\,\log(4n),$$

then $e^{-d\gamma^2/(32R^2)} \le \frac{1}{4n}$ and $e^{-d\gamma^2/(16R^2)} \le \frac{1}{4n}$, so for this i,

$$\Pr[y_i\langle Mw, Mx_i\rangle \ge \gamma/2] \;\ge\; 1 - \frac{1}{2n}.$$

Applying the union bound over all $i = 1, \ldots, n$,

$$\Pr[y_i\langle Mw, Mx_i\rangle \ge \gamma/2 \text{ for all } i] \;\ge\; 1 - n\cdot\frac{1}{2n} \;=\; \frac{1}{2}.$$

Thus, with probability at least $1/2$, every embedded example maintains margin at least $\gamma/2$ against Mw.

For parts 2 and 3, you can use the following fact about Gaussian random variables. If $g_1, \ldots, g_k$ are independent Gaussian random variables with mean zero and variance 1, then

$$Pr\left[\frac{1}{m}\sum_{i=1}^{m} g_i^2 \ge 1 + \epsilon\right] \le exp\left(-\frac{m\epsilon^2}{8}\right)$$

$$Pr\left[\frac{1}{m}\sum_{i=1}^{m} g_i^2 \ge 1 - \epsilon\right] \le exp\left(-\frac{m\epsilon^2}{4}\right)$$

3

Consider the function $k : (0,1) \times (0,1) \to \mathbb{R}$ defined by $k(x_1, x_2) = min\{x_1, x_2\}$.

1. Prove that k is a valid kernel (Hint: write k as the integral of a product of two simple functions and then prove that its Gram matrices are positive semi-definite).

   Answer: For $x \in (0,1)$ define the feature map $\phi_x : [0,1] \to \mathbb{R}$ by

   $$\phi_x(t) := \mathbf{1}\{t \le x\}.$$

   Then

   $$k(x_1, x_2) = \int_0^1 \phi_{x_1}(t)\,\phi_{x_2}(t)\,dt = \langle \phi_{x_1}, \phi_{x_2}\rangle_{L^2[0,1]}.$$

   Hence k is an inner-product kernel.

2. Now, consider a training set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ with $y_i \in \mathbb{R}$ and distinct points x_i in $(0,1)$. Show that if we ran kernel regression without regularization on this data set, we would obtain zero training error. More precisely, find explicit coefficients α_j, in terms of the training data, such that for all points (x_i, y_i) in the training set we have

   $$\sum_{j=1}^{n} \alpha_j \min\{x_j, x_i\} = y_i.$$

Answer: Without loss of generality, assume the inputs are sorted:

$$0 < x_1 < x_2 < \cdots < x_n < 1.$$

Define

$$f(x) = \sum_{j=1}^{n} \alpha_j \min\{x, x_j\}.$$

Then f is continuous and piecewise linear, with slope on each interval $(x_{i-1}, x_i]$ given by

$$f'(x) = \sum_{j:\, x_j \geq x} \alpha_j.$$

We want $f(x_i) = y_i$ for all i. To achieve this, define the slopes between adjacent points:

$$\beta_1 = \frac{y_1}{x_1}, \qquad \beta_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \quad \text{for } i = 2, \ldots, n.$$

These β_i describe the desired piecewise-linear interpolant through $(0, 0)$ and (x_i, y_i).

Since $f'(x) = \sum_{j=i}^{n} \alpha_j = \beta_i$ for $x \in (x_{i-1}, x_i]$, we can recover α from the backward differences:

$$\alpha_i = \beta_i - \beta_{i+1} \quad (i = 1, \ldots, n-1), \qquad \alpha_n = \beta_n,$$

where we take $\beta_{n+1} := 0$.

Then, for each i,

$$f(x_i) = \sum_{j=1}^{n} \alpha_j \min\{x_i, x_j\} = y_i,$$

so the regression interpolates the data exactly—yielding zero training error.

4

In the high-dimensional problems, there are usually an infinite number of possible models that perfectly fit the observed data. When a problem has multiple solutions, different optimization algorithms can find entirely different solutions to the same problem. Even though all of the solutions perfectly fit the training data, their generalization performance can be vastly different. In this problem, we explore this phenomenon for two widely used optimization algorithms: gradient descent and Adam. Consider a linear, binary classification problem under the squared loss. Let $X \in \mathbb{R}^{n \times d}$ be an $n \times d$ matrix of features, $y \in \{-1, 1\}^n$ be the corresponding vector of labels, and $\theta \in \mathbb{R}^d$ be the parameter vector. We wish to minimize the empirical risk

$$R_S[\theta] = \frac{1}{2}||X\theta - y||_2^2 \tag{1}$$

Assume that the rows of X are linearly independent and that $d > n$.

- Prove that there are infinite many $\theta \in \mathbb{R}^d$ such that $R_S[\theta] = 0$.

  Since the rows of $X \in \mathbb{R}^{n \times d}$ are linearly independent and $d > n$, we have $\text{rank}(X) = n$. Hence, the $n \times n$ matrix $XX^\top$ is invertible.

  - **Existence of one interpolating solution.**
    Define

    $$\theta_0 = X^\top (XX^\top)^{-1} y.$$

    Then

    $$X\theta_0 = XX^\top (XX^\top)^{-1} y = y,$$

    which implies

    $$R_S[\theta_0] = \frac{1}{2}\|X\theta_0 - y\|_2^2 = 0.$$

    Thus, θ_0 achieves zero training error.

– **Characterization of all interpolating solutions.**
Suppose θ satisfies $R_S[\theta] = 0$. Then $X\theta = y$. Subtracting the equation $X\theta_0 = y$, we obtain

$$X(\theta - \theta_0) = 0.$$

Hence, $\theta - \theta_0 \in \ker(X)$, and every such θ can be written as

$$\theta = \theta_0 + v, \qquad v \in \ker(X).$$

Because $\text{rank}(X) = n < d$, the nullspace $\ker(X)$ has dimension $d - n \geq 1$, implying there are infinitely many distinct θ satisfying $R_S[\theta] = 0$.

- Gradient descent generates a sequence of points $\{\theta_k^g d\}$ according to:

$$\theta_{k+1}^{gd} = \theta_k^{gd} - \alpha_k \nabla R_S[\theta_k^{gd}],$$

where α_k is a fixed sequence of learning rates. Assume the sequence α_k is chosen so that gradient descent converges to a minimizer of the objective (1). (You don't need to show how to select α_k). Suppose we initialize $\theta_0^{gd} = 0$. Show that gradient descent converges to the minimum Euclidean norm solution to $X\theta = y$.

Answer:

Consider gradient descent on $R_S(\theta) = \frac{1}{2}\|X\theta - y\|_2^2$:

$$\theta_{k+1}^{gd} \;=\; \theta_k^{gd} - \alpha_k \nabla R_S(\theta_k^{gd}) \;=\; \theta_k^{gd} - \alpha_k X^\top (X\theta_k^{gd} - y) \;=\; (I - \alpha_k X^\top X)\theta_k^{gd} + \alpha_k X^\top y.$$

With $\theta_0^{gd} = 0$, an induction shows $\theta_k^{gd} \in \text{range}(X^\top)$ for all k: the right-hand side is a sum of $(I - \alpha_k X^\top X)\theta_k^{gd} \in \text{range}(X^\top)$ and $X^\top y \in \text{range}(X^\top)$.

By the assumption on the step sizes (α_k), gradient descent converges to a minimizer θ_∞ of R_S. Since the problem is interpolable ($\text{rank}(X) = n$ and $d > n$), every minimizer satisfies $X\theta_\infty = y$, and thus θ_∞ is a solution to $X\theta = y$. Moreover, because each iterate lies in $\text{range}(X^\top)$ and $\text{range}(X^\top)$ is closed, we have $\theta_\infty \in \text{range}(X^\top)$.

Let $\theta^\dagger := X^\top (XX^\top)^{-1} y$ denote the minimum-norm solution (the pseudoinverse solution). Every solution of $X\theta = y$ can be written uniquely as

$$\theta \;=\; \theta^\dagger + v, \qquad v \in \ker(X),$$

and $\text{range}(X^\top)$ is orthogonal to $\ker(X)$ because $\langle X^\top u, v\rangle \;=\; \langle u, Xv\rangle \;=\; 0$ for all u and $v \in \ker(X)$. Hence, for any solution $\theta = \theta^\dagger + v$,

$$\|\theta\|_2^2 \;=\; \|\theta^\dagger\|_2^2 + \|v\|_2^2 \;\geq\; \|\theta^\dagger\|_2^2,$$

with equality iff $v = 0$ (i.e., $\theta \in \text{range}(X^\top)$). Since θ_∞ is both a solution and lies in $\text{range}(X^\top)$, it must equal $\theta^\dagger$. Therefore, gradient descent initialized at 0 converges to the minimum Euclidean norm solution:

$$\theta_\infty \;=\; X^\top (XX^\top)^{-1} y.$$

- Rather than use a fixed learning rate, Adam attempts to adapt the learning rate for each parameter using past gradient information. In particular, Adam generates a sequence of points $\{\theta_k^{ad}\}$ according to:

$$\theta_{k+1}^{ad} = \theta_k^{ad} - \alpha_k H_k^{-1} \nabla R_S[\theta_k^{ad}] + \beta_k H_k^{-1} H_{k-1}(\theta_k^{ad} - \theta_{k-1}^{ad}),$$

where α_k and β_k are fixed sequences, and H_k is a positive definite, diagonal matrix

$$H_k = diag\left(\left\{\sum_{i=1}^{k} \eta_i g_i \circ g_i\right\}^{1/2}\right),$$

where η_k is another fixed set of coefficients, $g_k = \nabla R_S[\theta_k^{ad}]$, and $\circ$ denotes an entry-wise product. Assume the sequences α_k, β_k , η_k are chosen so that Adam converges to a minimizer

of the objective (1). (You don't need to show how to choose these sequences). Suppose there exists some scalar c such that $X sign(X^\top y) = cy$, and we initialize $\theta_0^{ad}, \theta_{-1}^{ad} = 0$ Prove that Adam converges to the unique solution $\theta \propto sign(Xy)$. Hint: Use induction to show every iterate satisfies $\theta_k^{ad} = \lambda_k sign(X^\top y)$ for some scalar λ_k.

Answer:

Assume there exists $c \in \mathbb{R}$ such that

$$X \operatorname{sign}(X^\top y) = c\, y,$$

and initialize $\theta_0^{ad} = \theta_{-1}^{ad} = 0$. Let $g_k = \nabla R_S(\theta_k^{ad}) = X^\top(X\theta_k^{ad} - y)$ and

$$H_k = \operatorname{diag}\left(\Big\{ \sum_{i=1}^{k} \eta_i\, g_i \circ g_i \Big\}^{1/2} \right), \quad \alpha_k, \beta_k, \eta_k > 0,$$

with the Adam update

$$\theta_{k+1}^{ad} = \theta_k^{ad} - \alpha_k H_k^{-1} g_k + \beta_k H_k^{-1} H_{k-1}(\theta_k^{ad} - \theta_{k-1}^{ad}).$$

Step 1: Invariance of the span of $s := \operatorname{sign}(X^\top y)$. We prove by induction that for all $k \geq 0$ there exists a scalar λ_k with

$$\theta_k^{ad} = \lambda_k\, s.$$

The base case holds with $\lambda_0 = \lambda_{-1} = 0$. Assume $\theta_k^{ad} = \lambda_k s$ and $\theta_{k-1}^{ad} = \lambda_{k-1} s$. Then

$$g_k \;=\; X^\top(X\theta_k^{ad} - y) \;=\; X^\top(\lambda_k X s - y) \;=\; (\lambda_k c - 1)\, X^\top y \;=\; a_k\, X^\top y,$$

where $a_k := \lambda_k c - 1$.

Since each g_i is a scalar multiple of $X^\top y$, we have

$$g_i \circ g_i \;=\; a_i^2\, (X^\top y) \circ (X^\top y), \quad \sum_{i=1}^{k} \eta_i\, g_i \circ g_i \;=\; S_k\, (X^\top y) \circ (X^\top y),$$

with the scalar $S_k := \sum_{i=1}^{k} \eta_i a_i^2 > 0$. Hence

$$H_k \;=\; \sqrt{S_k}\, \operatorname{diag}\big(|X^\top y|\big), \quad H_k^{-1} \;=\; \frac{1}{\sqrt{S_k}} \operatorname{diag}\Big(\tfrac{1}{|X^\top y|}\Big).$$

Therefore,

$$H_k^{-1} g_k = \frac{a_k}{\sqrt{S_k}} \operatorname{diag}\Big(\tfrac{1}{|X^\top y|}\Big)\, X^\top y = \frac{a_k}{\sqrt{S_k}}\, s.$$

Similarly,

$$H_k^{-1} H_{k-1} = \frac{\sqrt{S_{k-1}}}{\sqrt{S_k}}\, I,$$

since the diagonal factors $\operatorname{diag}(|X^\top y|)$ cancel.

Plugging into Adam's update and using the induction hypothesis,

$$\theta_{k+1}^{ad} = \Big[\lambda_k - \alpha_k \frac{a_k}{\sqrt{S_k}} + \beta_k \frac{\sqrt{S_{k-1}}}{\sqrt{S_k}} (\lambda_k - \lambda_{k-1})\Big] s =: \lambda_{k+1}\, s.$$

Thus the form is preserved, completing the induction.

Step 2: Limit point and interpolation. By assumption on $(\alpha_k, \beta_k, \eta_k)$, Adam converges to a minimizer θ_∞ of R_S. Since the iterates remain in $\operatorname{span}\{s\}$ and this subspace is closed, $\theta_\infty = \lambda_\infty s$ for some λ_∞. Every minimizer of R_S in the overparameterized, full-row-rank setting satisfies $X\theta_\infty = y$. Hence

$$X(\lambda_\infty s) = y \iff \lambda_\infty X s = y \iff \lambda_\infty c\, y = y \iff \lambda_\infty = \frac{1}{c}.$$

Step 3: Uniqueness within the span and conclusion. If $\lambda_1 s$ and $\lambda_2 s$ both solve $X\theta = y$, then

$$0 = X(\lambda_1 s - \lambda_2 s) = (\lambda_1 - \lambda_2)Xs = (\lambda_1 - \lambda_2)c\,y,$$

forcing $\lambda_1 = \lambda_2$. Thus the interpolating point in $\mathrm{span}\{s\}$ is unique:

$$\theta_\infty \;=\; \frac{1}{c}\,\mathrm{sign}(X^\top y) \;\propto\; \mathrm{sign}(X^\top y).$$

Therefore, under the stated condition and initialization, Adam converges to the unique solution proportional to $\mathrm{sign}(X^\top y)$.

- Fix the labels $y \in \{-1, 1\}$, and let $X = [y; I_{n\times n}]$. Hence, only the first feature is discriminative, and the others are unrelated to the true label. Compute the solutions found by running (a) gradient descent and (b) Adam on this problem instance.

(a) Gradient descent (from $\theta_0 = 0$). Gradient descent on $R_S(\theta) = \frac{1}{2}\|X\theta - y\|_2^2$ with a convergent stepsize schedule and $\theta_0 = 0$ converges to the minimum-norm interpolator

$$\theta^{\mathrm{GD}} \;=\; X^\top (XX^\top)^{-1} y.$$

Here

$$XX^\top \;=\; yy^\top + I_n, \qquad (yy^\top + I_n)^{-1} y \;=\; \frac{1}{1 + \|y\|_2^2}\, y \;=\; \frac{1}{n+1}\, y,$$

so

$$\theta^{\mathrm{GD}} \;=\; X^\top \frac{1}{n+1} y \;=\; \frac{1}{n+1} X^\top y \;=\; \frac{1}{n+1} \begin{bmatrix} y^\top y \\ y \end{bmatrix} \;=\; \frac{1}{n+1} \begin{bmatrix} n \\ y \end{bmatrix}.$$

One can verify $X\theta^{\mathrm{GD}} = \frac{n}{n+1} y + \frac{1}{n+1} y = y$.

(b) Adam (from $\theta_0^{\mathbf{ad}} = \theta_{-1}^{\mathbf{ad}} = 0$). From the previous part's result, if there exists c such that $X\,\mathrm{sign}(X^\top y) = c\,y$, Adam converges to

$$\theta^{\mathrm{ADAM}} \;=\; \frac{1}{c}\,\mathrm{sign}(X^\top y).$$

In this instance,

$$X^\top y = \begin{bmatrix} y^\top y \\ y \end{bmatrix} = \begin{bmatrix} n \\ y \end{bmatrix} \quad \Rightarrow \quad \mathrm{sign}(X^\top y) = \begin{bmatrix} 1 \\ y \end{bmatrix} =: s,$$

and

$$X\,s \;=\; y \cdot 1 + I_n y \;=\; 2y,$$

so $c = 2$. Therefore

$$\theta^{\mathrm{ADAM}} \;=\; \frac{1}{2} \begin{bmatrix} 1 \\ y \end{bmatrix}, \qquad X\theta^{\mathrm{ADAM}} = \tfrac{1}{2}(y + y) = y.$$

Summary.

$$\boxed{\theta^{\mathrm{GD}} = \frac{1}{n+1} \begin{bmatrix} n \\ y \end{bmatrix}, \qquad \theta^{\mathrm{ADAM}} = \frac{1}{2} \begin{bmatrix} 1 \\ y \end{bmatrix}}$$

Both interpolate the data ($X\theta = y$), but θ^{GD} is the minimum-norm interpolator, whereas θ^{ADAM} generally is not.

- Compare the relative weight the solutions found in the above part place on the discriminative feature relative to the remaining features, i.e. compute $\frac{|\theta[1]|}{|\theta[i]|}$ for both gradient descent and Adam, where $\theta[i]$ denotes the i-th coordinate of θ. Heuristically, which solution do you expect to generalize better to new data?

Relative weight on the discriminative feature. From the previous part,

$$\theta^{\mathrm{GD}} = \frac{1}{n+1}\begin{bmatrix} n \\ y \end{bmatrix}, \qquad \theta^{\mathrm{ADAM}} = \frac{1}{2}\begin{bmatrix} 1 \\ y \end{bmatrix}.$$

For any $i \in \{2, \ldots, n+1\}$, we have $|\theta^{\mathrm{GD}}[1]| = \frac{n}{n+1}$ and $|\theta^{\mathrm{GD}}[i]| = \frac{1}{n+1}$, while $|\theta^{\mathrm{ADAM}}[1]| = \frac{1}{2}$ and $|\theta^{\mathrm{ADAM}}[i]| = \frac{1}{2}$. Hence

$$\frac{|\theta^{\mathrm{GD}}[1]|}{|\theta^{\mathrm{GD}}[i]|} = n, \qquad \frac{|\theta^{\mathrm{ADAM}}[1]|}{|\theta^{\mathrm{ADAM}}[i]|} = 1.$$

Heuristic generalization. Only the first feature is truly predictive. Gradient descent (minimum-norm interpolator) assigns n times more weight to this discriminative feature than to any spurious feature, and this selectivity *increases* with n. Adam, by contrast, spreads equal magnitude across all coordinates. Heuristically, the GD solution should generalize better to new data (lower variance on irrelevant features), whereas Adam's uniform weighting is more prone to overfit the spurious coordinates.

5

Apply the concepts from the course lectures to your final project.

- Provide your project abstract. You can copy it from the previous homework if you'd like. This may have changed from last week, and that's ok.

  We aim to use the BlueSky air traffic simulator to study and predict air traffic management (ATM) dynamics under high-density and partially faulty sensing conditions. BlueSky is an open-source simulator capable of running scenarios with hundreds of aircraft simultaneously, making it a powerful platform for generating realistic performance and navigation datasets. In this project, we extend the baseline scenario by introducing stochastic sensor-loss events, where drones have a random probability of losing sensor functionality during flight. The ultimate objective is to develop and evaluate machine-learning–based predictive models that can detect or anticipate sensor failure and maintain safe and efficient operations in dense advanced air mobility (AAM) environments. Our broader motivation is to contribute toward compositional and structured ATM policies that ensure scalability and safety in multi-agent airspaces.

- Collect a representative data set of (X, Y) pairs as you described in last week's homework. Describe what the features and labels mean for this prediction problem. Describe the geometric invariances of the data (e.g., do you expect it to behave like a sequence or an image? Or something else?)

  Features (X): Each data sample encodes the *state of a drone or aircraft* at a given simulation time step. Representative features include:

  $$X_t = [p_x,\, p_y,\, p_z,\, v,\, \psi,\, d_{\mathrm{sep}},\, n_{\mathrm{local}},\, w,\, h_t],$$

  where:

  - p_x, p_y, p_z: position coordinates (longitude, latitude, altitude),
  - v: airspeed,
  - ψ: heading angle,
  - d_{sep}: separation distance to the nearest neighbor,
  - n_{local}: local traffic density,
  - f_{plan}: flight plan waypoints or intended route,
  - w: wind speed or environmental disturbance (if modeled),
  - h_t: recent history window of past states.

These features are represented numerically as time-series vectors or aggregated feature arrays.

Labels (Y): whether the sensor of the drone is lost.

The target variable indicates whether the drone's sensor has failed:

$$Y = \begin{cases} 1, & \text{if the drone loses its sensor at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the problem is a binary classification task.

- Find an appropriate vector encoding of X so that you can apply linear classification. Describe why this encoding is reasonable for this set of data.

  – **Baseline (hand-crafted) encoding.**
  We encode each drone at time t using a fixed-length window of recent states of length L (simulation steps). Let $s_\tau \in \mathbb{R}^{d_0}$ be the per-step state

  $$s_\tau = [p_x(\tau),\, p_y(\tau),\, p_z(\tau),\, v(\tau),\, \psi(\tau),\, d_{\text{sep}}(\tau),\, n_{\text{local}}(\tau),\, w(\tau)].$$

  We form a flattened feature vector with dynamics- and interaction-aware components:

  $$\tilde{X}_t \;=\; \Big[\underbrace{s_{t-L+1} \,\|\, \cdots \,\|\, s_t}_{\text{raw history}} \;\|\; \underbrace{\Delta s_{t-L+2} \,\|\, \cdots \,\|\, \Delta s_t}_{\text{first differences}}\Big] \in \mathbb{R}^D,$$

  where $\Delta s_\tau := s_\tau - s_{\tau-1}$, computed over the window $[t-L+1,\, t]$.

  – **LLM-based lift (learned embedding).** To compare against the hand-crafted baseline, we apply a sequence encoder ("LLM") to obtain a high-dimensional lift

  $$\phi_\theta(\tilde{X}_t) \in \mathbb{R}^{d'}, \qquad d' \gg D,$$

  then train a linear classifier on $\phi_\theta(\tilde{X}_t)$. *Rationale:* A learned lift can linearize complex dependencies (e.g., cumulative disturbance, interaction effects) while keeping the downstream model simple and data-efficient. We will vary d' and regularization to assess the efficiency of embedding extraction versus the baseline.

- Fit a linear model to this data. You can use whichever software you'd like, but specify the code you used to fit this model. Is the data linearly separable? Why or why not?

  A *linear* classifier is required to assess linear separability; a DNN/MLP is *not* linear. We therefore train (i) a logistic-regression model (single affine layer) to test linear separability on $\tilde{X}_t$, and (ii) a shallow 2-layer MLP as a nonlinear baseline. Both use `BCEWithLogitsLoss`.

- Find a simple nonlinear lift of the data that is linearly separable. Describe the lift in equations or code. What is the associated margin? Again, you can use whichever solver you'd like to make the data separable. Be creative in constructing your lift so that it is computationally efficient to compute.

  We design a lightweight sequence encoder ("LLM") that maps the L-step history into a high-dimensional embedding $\phi_\theta(\tilde{X}_t) \in \mathbb{R}^{d'}$ on which a *linear* classifier is trained.