Support Vector Machines Problem Solving

Four problems, small enough to trace every number by hand

CSCI 567 — Machine Learning • Discussion 05 • Sep 25, 2026

Outline: four problems that put Lectures 04 and 05 in tension

1. **Hard-margin SVM by hand**: max-margin separator and support vectors from four points.
2. **Kernel SVM on XOR**: linear SVM has no solution; a degree-2 kernel makes the dual solvable by hand, and the answer is $f(\mathbf{x}) = x_1 x_2$.
3. **Class imbalance across losses**: the same lopsided dataset through square, logistic, hinge, and hard margin — which boundaries move, and in which direction.
4. **Gradient descent on separable data**: logistic loss has no minimizer, yet gradient descent's direction converges to the max-margin separator — Lecture 04's implicit regularization, for classification.

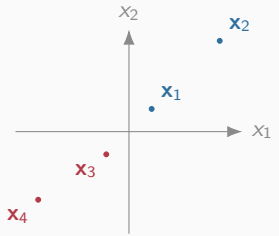
⚠ HW2 (Lectures 03–04) is due Mon 9/28, 11:59pm.

1. Hard-margin SVM by hand

Finding the max-margin separator, from four points

Problem 1: setup, four points in the plane

Data. $\mathbf{x}_1 = (1, 1)$, $\mathbf{y}_1 = +1$; $\mathbf{x}_2 = (4, 4)$, $\mathbf{y}_2 = +1$; $\mathbf{x}_3 = (-1, -1)$, $\mathbf{y}_3 = -1$; $\mathbf{x}_4 = (-4, -3)$, $\mathbf{y}_4 = -1$.



blue: $\mathbf{y} = +1$ red: $\mathbf{y} = -1$

Goal: find the max-margin separating hyperplane $\mathbf{w}, b$ by hand — no solver, just symmetry and a small linear system

Guessing the support vectors, by symmetry

$\mathbf{x}_1 = (1, 1)$ and $\mathbf{x}_3 = (-1, -1)$ are symmetric about the origin, and the closest pair across the two classes.

Guess: these two are exactly on the margin — the support vectors — and $\mathbf{x}_2, \mathbf{x}_4$ (farther out) are not

If the guess is right, both sit at margin value exactly 1:

$$\mathbf{y}_1(\mathbf{w}^\top \mathbf{x}_1 + b) = 1, \qquad \mathbf{y}_3(\mathbf{w}^\top \mathbf{x}_3 + b) = 1.$$

Two equations, but $\mathbf{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$ is three unknowns — symmetry supplies the missing constraint (next slide).

From the guess to two equations

Expand the two tightness conditions ($\mathbf{y}_1 = +1$, $\mathbf{y}_3 = -1$):

$$w_1 + w_2 + b = 1, \qquad -(-w_1 - w_2 + b) = 1 \implies w_1 + w_2 - b = 1.$$

Add the two equations: $2(w_1 + w_2) = 2 \Rightarrow w_1 + w_2 = 1$

Subtract them: $2b = 0 \Rightarrow b = 0$

One equation, $w_1 + w_2 = 1$, and one unknown pinned ($b = 0$) — still need a second equation for w_1, w_2 individually.

Pinning down w: the missing constraint

Symmetry argument. The max-margin hyperplane separating two points is their *perpendicular bisector* — its normal vector $\mathbf{w}$ must point along the line connecting them, $\mathbf{x}_1 - \mathbf{x}_3 = (2, 2)$.

$$\text{so } \mathbf{w} \quad = \quad (t, t) \text{ for some } t$$

Combined with $w_1 + w_2 = 1$ from the previous slide: $2t = 1 \Rightarrow t = 0.5$.

$$\boxed{\mathbf{w} = (0.5, 0.5), \quad b = 0}$$

Verifying the guess: $\mathbf{x}_2, \mathbf{x}_4$ outside the margin

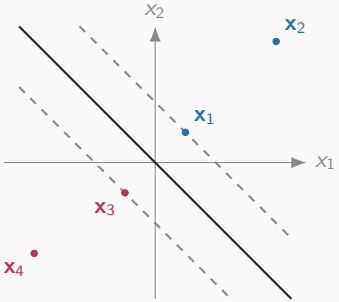
The guess in Part 1 is only valid if the other two points are *not* tighter than margin 1 — check directly:

$$\mathbf{y}_2(\mathbf{w}^\top \mathbf{x}_2 + b) \;=\; 1 \cdot (0.5(4) + 0.5(4) + 0) \;=\; 4 \;>\; 1 \quad \checkmark$$

$$\mathbf{y}_4(\mathbf{w}^\top \mathbf{x}_4 + b) = -1 \cdot (0.5(-4) + 0.5(-3) + 0) = -1 \cdot (-3.5) = 3.5 > 1 \quad \checkmark$$

both strictly outside the margin — the guess is confirmed; $\mathbf{x}_2, \mathbf{x}_4$ are **not** support vectors

The margin, geometrically



solid: separator $x_1 + x_2 = 0$ dashed: margin boundaries $x_1 + x_2 = \pm 2$

$$\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{0.5^2 + 0.5^2}} = \frac{1}{\sqrt{0.5}} = \sqrt{2} \approx 1.414$$

Verified numerically

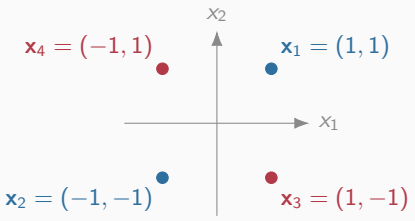
✅ subgradient descent on this exact dataset (soft-margin, $\lambda = 10^{-4}$, essentially hard margin) converges to $\mathbf{w} \approx (0.505, 0.496)$, $b \approx 0$

Matches the hand solution $\mathbf{w} = (0.5, 0.5)$, $b = 0$ to two decimal places — Lecture 05's subgradient-descent recipe, run as written.

2. Kernel SVM on XOR

Where duality earns its keep

Problem 2: setup, and why no linear SVM exists



Claim. No $(\mathbf{w}, b)$ satisfies all four hard-margin constraints $\mathbf{y}_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$.

Proof. Add the two $\mathbf{y} = +1$ constraints: $\mathbf{w}^\top(\mathbf{x}_1 + \mathbf{x}_2) + 2b \geq 2$, and $\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{0}$, so $b \geq 1$. Add the two $\mathbf{y} = -1$ constraints: $-\mathbf{w}^\top(\mathbf{x}_3 + \mathbf{x}_4) - 2b \geq 2$, and $\mathbf{x}_3 + \mathbf{x}_4 = \mathbf{0}$, so $b \leq -1$. Contradiction. ■

Goal: fix this with Lecture 04's kernel trick, and solve the resulting dual entirely by hand

The degree-2 kernel, and where it sends the four points

Use $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^2$, Lecture 04's degree-2 polynomial kernel; its explicit feature map in $d = 2$:

$$(\mathbf{x}^\top \mathbf{z})^2 \;=\; (x_1 z_1 \,+\, x_2 z_2)^2 \;=\; x_1^2 z_1^2 \,+\, 2 x_1 x_2 z_1 z_2 \,+\, x_2^2 z_2^2 \;=\; \phi(\mathbf{x})^\top \phi(\mathbf{z}), \quad \phi(\mathbf{x}) \;=\; \left(x_1^2,\ \sqrt{2}\, x_1 x_2,\ x_2^2\right)$$

For every XOR point $x_1^2 = x_2^2 = 1$, so only the middle coordinate varies:

point	$\mathbf{y}$	$\phi(\mathbf{x})$
$\mathbf{x}_1 = (1,1)$, $\mathbf{x}_2 = (-1,-1)$	$+1$	$(1,\ +\sqrt{2},\ 1)$
$\mathbf{x}_3 = (1,-1)$, $\mathbf{x}_4 = (-1,1)$	-1	$(1,\ -\sqrt{2},\ 1)$

in feature space the classes sit at $\pm\sqrt{2}$ along one axis: separable, and the max-margin hyperplane is visibly $\sqrt{2}\, x_1 x_2 \;=\; 0$

The kernel matrix has a block structure

Inner products between XOR points take three values: $\mathbf{x}_i^\top \mathbf{x}_i = 2$, $\mathbf{x}_i^\top (-\mathbf{x}_i) = -2$, and 0 between the classes' points. Squaring:

$$K \;=\; \left[(\mathbf{x}_i^\top \mathbf{x}_j)^2\right] \;=\; \begin{pmatrix} 4 & 4 & 0 & 0 \\ 4 & 4 & 0 & 0 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \end{pmatrix}$$

ordering $(\mathbf{x}_1, \mathbf{x}_2 \;\mid\; \mathbf{x}_3, \mathbf{x}_4)$: the same-class blocks are all 4s, the cross-class blocks are all 0s — every cross term in the dual objective vanishes before we start

Sanity check against the feature map: $\phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_3) = 1 - 2 + 1 = 0$ and $\phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_2) = 1 + 2 + 1 = 4$.

The hard-margin dual, reduced by symmetry to one variable

Lecture 05's dual with $\lambda \to 0$ (the classic Lagrangian form), kernelized:

$$\max_{\alpha \geq 0,\ \sum_i \alpha_i \mathbf{y}_i = 0} \sum_{i=1}^{4} \alpha_i \ - \ \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j K_{ij}$$

With the block K, the double sum has only same-class terms ($\mathbf{y}_i \mathbf{y}_j = 1$, $K_{ij} = 4$): $\sum_{i,j} \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j K_{ij} = 4(\alpha_1 + \alpha_2)^2 + 4(\alpha_3 + \alpha_4)^2$. The equality constraint says $\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4 =: s$, so the whole dual collapses to

$$\max_{s \geq 0} \ 2s - \tfrac{1}{2} \cdot 8s^2 \ = \ 2s - 4s^2 \ \Rightarrow \ s^\star \ = \ \tfrac{1}{4}$$

Only the sums are pinned down; by the problem's symmetry (every point is a reflection of every other) take $\alpha_i^\star = \frac{1}{8}$ for all four — all four points are support vectors.

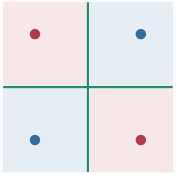
Recovering the classifier: $f(\mathbf{x}) = x_1 x_2$, margin $\sqrt{2}$

Prediction rule from the dual, $f(\mathbf{x}) = \sum_i \alpha_i^\star \mathbf{y}_i \, \kappa(\mathbf{x}_i, \mathbf{x}) + b^\star$, with $\kappa(\mathbf{x}_i, \mathbf{x}) = (\mathbf{x}_i^\top \mathbf{x})^2$:

$$f(\mathbf{x}) \;=\; \tfrac{1}{8}\big[(x_1 + x_2)^2 + (-x_1 - x_2)^2 - (x_1 - x_2)^2 - (-x_1 + x_2)^2\big] + b^\star \;=\; \tfrac{1}{8} \cdot 2 \cdot 4 x_1 x_2 + b^\star \;=\; x_1 x_2 + b^\star$$

$b^\star$ from any support vector: $\mathbf{y}_1 = f(\mathbf{x}_1)$ gives $1 = 1 + b^\star$, so $b^\star = 0$.

$f(\mathbf{x}) = x_1 x_2$: boundary $x_1 x_2 = 0$, the two axes — exactly XOR. In feature space $\mathbf{w}_\phi = \sum_i \alpha_i^\star \mathbf{y}_i \phi(\mathbf{x}_i) = (0, \frac{1}{\sqrt{2}}, 0)$, margin $1/\|\mathbf{w}_\phi\| = \sqrt{2}$; every $\mathbf{y}_i f(\mathbf{x}_i) = 1$ (all on the margin, matching $\alpha_i^\star > 0$) and $\|\mathbf{w}_\phi\|^2 = \frac{1}{2} = \sum_i \alpha_i^\star$



green: $f(\mathbf{x}) = 0$; shading: sign of f

3. Class imbalance across losses

Same data, four losses, three different answers

Problem 3: setup, Lecture 05's cluster dataset

Lecture 05's example, one dimension: class -1 at $x = -1$ (five points) and the edge point $x = -0.2$; class $+1$ at $x = +1$ (five points) and the edge point $x = +0.2$; plus N class-$+1$ points at $x = 3$, all safely correct from the start.

Task. Fit four classifiers to this data for $N = 0,\ 50,\ 200$ and record the boundary $-b/w$: square loss (least squares on $\mathbf{y} = \pm 1$), ℓ_2-regularized logistic and hinge (both $\lambda = 0.03$, in Lecture 05's $\hat{\mathcal{R}} + \frac{\lambda}{2} w^2$ form), and hard-margin SVM.

before looking: with no cluster every method puts the boundary at 0 by symmetry — predict which way each moves as N grows, and which edge point pays

The boundaries: two move in opposite directions, two do not move

loss	$N = 0$	$N = 50$	$N = 200$	which edge point is sacrificed
square (least squares)	0.000	0.432	0.449	the $+0.2$ point (its own class) at $N \geq 50$
logistic, $\lambda = 0.03$	0.000	-0.118	-0.660	the -0.2 point (other class) at $N = 200$
hinge, $\lambda = 0.03$	0	0.000	0.000	neither
hard-margin SVM	0	0	0	neither

Direct minimization, several starts. Hinge at $N = 0$: the minimizer is not unique in b (both edge points sit inside the margin, where the objective is flat in b), but the minimizer set is symmetric and contains 0; for $N \geq 50$ the solution is exactly $w = 1$, $b = 0$, cluster at margin 3.

the same imbalance moves the square-loss boundary *toward* the cluster and the logistic boundary *away* from it, and leaves hinge and hard margin where they were — three behaviors from one dataset

Why each loss does what it does

Square wants every score *equal* to ± 1; the cluster's scores overshoot $+1$, so the fit shrinks w and shifts toward the cluster — the $+0.2$ point, on the cluster's own side, is lost.

Logistic is monotone and never zero: each cluster point pays $\log(1 + e^{-m}) > 0$ and wants m larger; raising b does that for all N at once, so the boundary moves away from the cluster — the -0.2 point, on the *other* side, is lost.

Hinge is exactly zero once $m \geq 1$; the cluster sits at margin 3 and contributes nothing to the objective or its subgradient — the twelve base points decide the fit, whatever N is.

Hard margin sees only the closest points, ± 0.2: their midpoint, 0, for every N.

The usual fix, class reweighting, fixes the count and breaks the geometry

Reweight logistic regression so each class has total weight 1 (each class-$+1$ point weighted $1/(6+N)$, each class--1 point $1/6$), same $\lambda = 0.03$:

	$N=0$	$N=50$	$N=200$
logistic, unweighted	0.000	-0.118	-0.660
logistic, class-reweighted	0.000	0.716	0.977

reweighting overshoots the other way: the six base class-$+1$ points now weigh $1/56$ or $1/206$ each, so the $+0.2$ edge point is sacrificed instead — the boundary passes 0.7

Lesson. Class counts are the wrong knob: the damage comes from the loss's tail. A loss that is zero past the margin (hinge) is immune without reweighting; hard margin is immune by construction.

4. Gradient descent on separable data

Implicit regularization picks the max-margin separator

Problem 4: setup, Problem 1's data under logistic loss with no regularizer

Problem 1's four points, $\mathbf{x}_1 = (1, 1)$, $\mathbf{x}_2 = (4, 4)$ ($\mathbf{y} = +1$), $\mathbf{x}_3 = (-1, -1)$, $\mathbf{x}_4 = (-4, -3)$ ($\mathbf{y} = -1$); no bias ($b = 0$, which Problem 1's solution had anyway). Run plain gradient descent on the *unregularized* logistic loss, from $\mathbf{w}_0 = \mathbf{0}$, step size $\eta = 0.1$:

$$\hat{\mathcal{R}}_{\log}(\mathbf{w}) = \tfrac{1}{4}\sum_{i=1}^{4} \log\left(1 + e^{-\mathbf{y}_i \mathbf{w}^\top \mathbf{x}_i}\right), \qquad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \hat{\mathcal{R}}_{\log}(\mathbf{w}_t)$$

Lecture 05: the data is separable, so the loss has no minimizer and $\|\mathbf{w}_t\| \to \infty$. Lecture 04's sharper question: among all the solutions, *which one* does gradient descent pick?

Task. Track the *direction* $\mathbf{w}_t / \|\mathbf{w}_t\|$ and its geometric margin $\min_i \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w}_t / \|\mathbf{w}_t\|$. Problem 1's hard-margin answer: $\mathbf{w}^\star = (0.5, 0.5)$, direction $(1, 1)/\sqrt{2}$, margin $1/\|\mathbf{w}^\star\| = \sqrt{2} \approx 1.4142$.

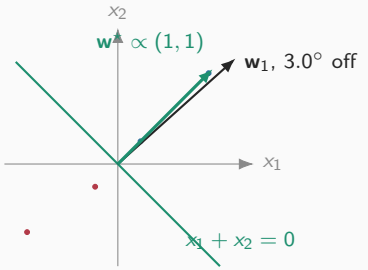
Gradient descent: the norm diverges, the direction converges to max margin

step t	$\|\mathbf{w}_t\|$	angle to $(1,1)/\sqrt{2}$	geometric margin of $\mathbf{w}_t/\|\mathbf{w}_t\|$	$\hat{\mathcal{R}}_{\log}(\mathbf{w}_t)$
1	0.17	3.01°	1.4123	4.6×10^{-1}
10	0.65	2.46°	1.4129	1.8×10^{-1}
10^2	1.68	1.21°	1.4139	4.5×10^{-2}
10^3	3.25	0.63°	1.4141	5.0×10^{-3}
10^4	4.88	0.42°	1.4142	5.0×10^{-4}
10^6	8.14	0.25°	1.4142	5.0×10^{-6}

Three things at once, at three speeds:

- the loss goes to 0 like $1/t$ — Lecture 05's "never done";
- the norm grows like $\log t$ — 1.6 more per factor of 10 in t, never settling;
- the direction converges to Problem 1's max-margin direction, and the induced margin reaches $\sqrt{2}$ to four decimals by $t = 10^4$.

Reading the result (optional): the SVM answer, with no margin written down



the iterates' direction starts 3° off $\mathbf{w}^\star$ and closes in; the norm keeps growing

Theorem (Soudry et al., 2018). For almost every separable dataset, gradient descent on logistic loss (small enough step) has $\mathbf{w}_t / \|\mathbf{w}_t\| \to \mathbf{w}^\star / \|\mathbf{w}^\star\|$, the hard-margin SVM direction; $\|\mathbf{w}_t\|$ grows like $\log t$.

Lecture 04's pattern: for regression, gradient descent from 0 picks the minimum-norm interpolant; for separable classification, the max-margin separator. The optimizer's path supplied the margin.

Caveats. The rate is logarithmic (0.25° off after a million steps), so the SVM's explicit margin is far cheaper; an ℓ_2 regularizer changes the answer to the regularized minimizer.

Optional: the course uses the previous frame's experiment, not this theorem.

Wrap-up

Four problems, one habit: verify, don't just guess

Wrap-up: what each problem showed

Hard-margin SVM by hand: guess support vectors from symmetry, solve a small linear system, *verify* the rest are outside the margin — the verification makes the guess a proof.

Kernel SVM on XOR: no linear separator (two lines of algebra); a degree-2 kernel gives a block K, a one-variable dual, $\alpha_i = \frac{1}{8}$, $f(\mathbf{x}) = x_1 x_2$.

Imbalance across losses: the same cluster moves square toward it, logistic away from it, and leaves hinge and hard margin alone; reweighting overshoots. The loss's tail, not the class count, is the knob.

Gradient descent on separable data: the loss never bottoms out, but the direction converges to the max-margin separator — implicit regularization for classification (the general theorem, Soudry et al. 2018, is optional).

Coming up, and reference (optional)

Lecture 06: Unsupervised Learning — PCA, clustering, GMM/EM

Discussion 06: HW2 Review and Exam 1 Problem Discussion

Reference (optional). D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, N. Srebro. *The Implicit Bias of Gradient Descent on Separable Data.* Journal of Machine Learning Research, 19(70):1–57, 2018. arXiv:1710.10345.