DSC 190/291 · Assignment 7 Solutions

UCSD · Spring 2026

Part A: Kernel Methods (10 points)

This homework continues the scale-sensitive complexity theme from Homework 6, where generalization is controlled by norms and margins rather than by dimension.

Let $\mathcal{U}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$, and let $\varphi : \mathcal{X} \to \mathcal{U}$ be a feature map with associated kernel $K(x, x') = \langle \varphi(x), \varphi(x') \rangle$. For $B > 0$, let $\mathcal{H}_{K,B} = \{x \to \langle w, \varphi(x) \rangle : w \in \mathcal{U}, \|w\| \leq B\}$ be the predictors of RKHS norm at most B. For a sample $S = (x_1, ..., x_n)$ and a real-valued function class $\mathcal{F}$, write the empirical Rademacher complexity

$$\hat{\mathcal{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i),$$

where $\sigma_1, ..., \sigma_n$ are independent uniform signs, and write $\mathcal{R}_{\mathcal{D}^n}(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{D}^n} \hat{\mathcal{R}}_S(\mathcal{F})$.

1. **(5 points) Norm-based kernel bound.**

   Prove the empirical bound

   $$\hat{\mathcal{R}}_S(\mathcal{H}_{K,B}) \leq \frac{B}{n} \sqrt{\sum_{i=1}^{n} K(x_i, x_i)},$$

   and conclude the distribution-level bound

   $$\mathcal{R}_{\mathcal{D}^n}(\mathcal{H}_{K,B}) \leq B \sqrt{\frac{\mathbb{E}[K(X, X)]}{n}}.$$

   Solution.

   Let

   $$v_\sigma = \sum_{i=1}^{n} \sigma_i \varphi(x_i).$$

   By Hilbert-space duality,

   $$\hat{\mathcal{R}}_S(\mathcal{H}_{K,B}) = \mathbb{E}_\sigma \sup_{\|w\| \leq B} \frac{1}{n} \langle w, v_\sigma \rangle = \frac{B}{n} \mathbb{E}_\sigma \|v_\sigma\|.$$

   Jensen's inequality gives

   $$\mathbb{E}_\sigma \|v_\sigma\| \leq \sqrt{\mathbb{E}_\sigma \|v_\sigma\|^2}.$$

   The squared norm is

$$\|v_\sigma\|^2 = \sum_{i,j} \sigma_i \sigma_j \langle \varphi(x_i), \varphi(x_j) \rangle.$$

Taking expectation over the independent signs removes the cross terms, so

$$\mathbb{E}_\sigma \|v_\sigma\|^2 = \sum_{i=1}^n \|\varphi(x_i)\|^2 = \sum_{i=1}^n K(x_i, x_i).$$

Therefore

$$\hat{\mathcal{R}}_S(\mathcal{H}_{K,B}) \leq \frac{B}{n} \sqrt{\sum_{i=1}^n K(x_i, x_i)}.$$

Now take expectation over $S \sim \mathcal{D}^n$ and use Jensen again:

$$\mathcal{R}_{\mathcal{D}^n}(\mathcal{H}_{K,B}) \leq \frac{B}{n} \mathbb{E}_S \sqrt{\sum_{i=1}^n K(X_i, X_i)} \leq \frac{B}{n} \sqrt{n \mathbb{E}[K(X,X)]} = B \sqrt{\frac{\mathbb{E}[K(X,X)]}{n}}.$$

2. **(5 points) Representer theorem.**

The space $\mathcal{U}$ may be infinite-dimensional, yet for any $\lambda > 0$ the regularized objective

$$J(w) = \frac{1}{n} \sum_{i=1}^n \text{loss}(\langle w, \varphi(x_i) \rangle; y_i) + \lambda \|w\|^2$$

can be minimized over finitely many coefficients. This is the representer theorem: prove that whenever J attains a minimum over $w \in \mathcal{U}$, some minimizer has the form

$$w^* = \sum_{j=1}^n \alpha_j \varphi(x_j), \quad \alpha \in \mathbb{R}^n.$$

Then rewrite J as a minimization over $\alpha \in \mathbb{R}^n$ in which the data enter only through the Gram matrix K_S, with entries $(K_S)_{ij} = K(x_i, x_j)$.

Solution.

Let $V = \text{span}\{\varphi(x_1), ..., \varphi(x_n)\}$. Every $w \in \mathcal{U}$ decomposes uniquely as $w = u + v$ with $u \in V$ and $v \perp V$. For each training point,

$$\langle w, \varphi(x_i) \rangle = \langle u, \varphi(x_i) \rangle,$$

because v is orthogonal to V. Thus the empirical loss term depends only on u, while

$$\|w\|^2 = \|u\|^2 + \|v\|^2.$$

Since $\lambda > 0$, $J(u + v) \geq J(u)$, with equality only if $v = 0$. Hence, whenever a minimizer exists, a minimizer lies in V and has the form

$$w^* = \sum_{j=1}^n \alpha_j \varphi(x_j).$$

For such a vector, the prediction on x_i is

$$\langle w, \varphi(x_i)\rangle = \sum_{j=1}^{n} \alpha_j K\big(x_j, x_i\big) = (K_S\alpha)_i,$$

where $(K_S)_{ij} = K\big(x_i, x_j\big)$. Also

$$\|w\|^2 = \sum_{i,j} \alpha_i\alpha_j K\big(x_i, x_j\big) = \alpha^\top K_S\alpha.$$

Therefore the finite-dimensional objective is

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \mathrm{loss}\Big((K_S\alpha)_i; y_i\Big) + \lambda\alpha^\top K_S\alpha.$$

Only the Gram matrix K_S appears.

Part B: ℓ_1 Rademacher Complexity (25 points)

Throughout this part, $\hat{\mathcal{R}}_S$ denotes the empirical Rademacher complexity defined in Part A.

1. **(8 points) Convex hulls.**

   Let $\mathcal{F}$ be a finite class of real-valued functions. Its convex hull is

   $$\mathrm{conv}(\mathcal{F}) = \left\{\sum_{f\in\mathcal{F}} a_f f : a_f \geq 0, \sum_{f\in\mathcal{F}} a_f = 1\right\}.$$

   Prove that empirical Rademacher complexity is unchanged by taking the convex hull:

   $$\hat{\mathcal{R}}_S(\mathrm{conv}(\mathcal{F})) = \hat{\mathcal{R}}_S(\mathcal{F}).$$

   Solution.

   Fix the sample S and signs σ. For any $F \in \mathrm{conv}(\mathcal{F})$, write $F = \sum_{f\in\mathcal{F}} a_f f$ with $a_f \geq 0$ and $\sum_f a_f = 1$. Then

   $$\frac{1}{n}\sum_{i=1}^{n} \sigma_i F(x_i) = \sum_{f\in\mathcal{F}} a_f \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right) \leq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i).$$

   This proves

   $$\hat{\mathcal{R}}_S(\mathrm{conv}(\mathcal{F})) \leq \hat{\mathcal{R}}_S(\mathcal{F}).$$

   The reverse inequality holds because $\mathcal{F} \subseteq \mathrm{conv}(\mathcal{F})$. Therefore the two empirical Rademacher complexities are equal.

2. **(10 points) ℓ_1 linear predictors.**

   Let $\varphi : \mathcal{X} \to \mathbb{R}^d$ satisfy $\|\varphi(x)\|_\infty \leq R$ for all x, and define

   $$\mathcal{H}_B^1 = \{x \to \langle w, \varphi(x)\rangle : \|w\|_1 \leq B\}.$$

Derive a logarithmic-in-dimension Rademacher bound for $\mathcal{H}_B^1$. Massart's finite-class lemma bounds the Rademacher complexity of a finite class; the convex-hull identity from the previous problem is what connects $\mathcal{H}_B^1$ to such a class.

Solution.

For each coordinate define $\psi_j(x) = \varphi_j(x)$ and include both signed coordinate functions $\mathcal{F}_{\text{pm}} = \{\psi_1, -\psi_1, ..., \psi_d, -\psi_d\}$. Since $\|\varphi(x)\|_\infty \le R$, every function in $\mathcal{F}_{\text{pm}}$ is bounded in absolute value by R. Massart's finite-class lemma gives

$$\hat{\mathcal{R}}_S(\mathcal{F}_{\text{pm}}) \le R\sqrt{\frac{2\log(2d)}{n}}.$$

Every w with $\|w\|_1 \le B$ can be written as B times an element of $\text{conv}(\mathcal{F}_{\text{pm}})$, allowing unused mass to be placed on canceling signed pairs. Equivalently, split $w_j = w_j^+ - w_j^-$ and normalize the nonnegative coefficients. By scaling and the convex-hull identity,

$$\hat{\mathcal{R}}_S(\mathcal{H}_B^1) \le B\hat{\mathcal{R}}_S(\mathcal{F}_{\text{pm}}) \le BR\sqrt{\frac{2\log(2d)}{n}}.$$

The same bound holds after expectation over S:

$$\mathcal{R}_{\mathcal{D}^n}(\mathcal{H}_B^1) \le BR\sqrt{\frac{2\log(2d)}{n}}.$$

3. **(7 points) Comparison with sparsity bounds.**

In Homework 4 you showed that the class of k-sparse linear classifiers in $\mathbb{R}^d$ has VC dimension $O(k\log(ed/k))$. With the VC generalization bound, with probability at least $1-\delta$ over the draw of $S \sim \mathcal{D}^n$, every k-sparse linear classifier h satisfies

$$L_{\mathcal{D}}^{0/1}(h) - L_S^{0/1}(h) \le C\sqrt{\frac{k\log(ed/k) + \log(1/\delta)}{n}}$$

for a universal constant C; structural risk minimization makes the bound adapt to the unknown sparsity level k.

Using your Rademacher bound from the previous problem and the week 7 uniform convergence theorem, derive the analogous bound for $\mathcal{H}_B^1$: a bound on $L_{\mathcal{D}}^{\ell}(w) - L_S^{\ell}(w)$, holding with probability at least $1-\delta$ over the draw of $S \sim \mathcal{D}^n$, uniformly over $\|w\|_1 \le B$, for a fixed loss ℓ that is G-Lipschitz in the prediction and bounded in $[0,1]$.

These two bounds control different loss functionals, so they are not interchangeable guarantees. Compare them: which quantity is the complexity in each and which one is scale-sensitive; what extra step the ℓ_1 bound needs before it can say anything about 0/1 error; and exhibit a predictor for which the sparsity bound is vacuous while the ℓ_1 bound is not.

Solution.

Let h_w denote the predictor $x \to \langle w, \varphi(x)\rangle$. Write $\mathcal{L}$ for the loss class induced by applying ℓ to these predictors. Contraction and the previous bound give

$$\mathcal{R}_{\mathcal{D}^n}(\mathcal{L}) \le GBR\sqrt{\frac{2\log(2d)}{n}}.$$

The one-sided Rademacher uniform convergence theorem gives, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, uniformly over all $\|w\|_1 \le B$,

$$L^{\ell}_{\mathcal{D}}(w) - L^{\ell}_{S}(w) \le 2GBR\sqrt{\frac{2\log(2d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Equivalently, this is $O\Big(GBR\sqrt{\log d/n} + \sqrt{\log(1/\delta)/n}\Big)$.

The sparse-VC bound and the ℓ_1 Rademacher bound control different objects. The sparse bound is a $\frac{0}{1}$ classifier guarantee whose complexity term is $k\log(ed/k)$. It is scale-insensitive: multiplying a nonzero sparse predictor does not change its classifier. The ℓ_1 bound is a Lipschitz-loss guarantee whose complexity term is controlled by $BR\sqrt{\log d}$ in the deviation bound. It is scale-sensitive through B and through the choice of loss. To obtain a $\frac{0}{1}$ statement from the ℓ_1 route, one needs a surrogate or margin loss that upper-bounds $\frac{0}{1}$ error, such as the ramp or hinge loss after a margin normalization.

For example, take a dense predictor $w = \big(\frac{1}{d}, ..., \frac{1}{d}\big)$. It has support d but $\|w\|_1 = 1$. If $d \gg n$, the sparse guarantee applied with $k = d$ has a term of order $\sqrt{\frac{d}{n}}$ and can be vacuous. For a bounded Lipschitz surrogate with $R = 1$, the ℓ_1 bound has only the logarithmic dimension dependence $\sqrt{\log d/n}$.

Part C: Normalized ℓ_1 Margins and FixedBoost (30 points)

In this part the sample is $S = ((x_1, y_1), ..., (x_n, y_n))$ with labels $y_i \in \{-1, +1\}$, and $\mathcal{G}$ is a finite base class of binary predictors $g : \mathcal{X} \to \{-1, +1\}$, symmetric in the sense that $-g \in \mathcal{G}$ whenever $g \in \mathcal{G}$. A coefficient vector $w = \big(w_g\big)_{g\in\mathcal{G}}$ parameterizes the real-valued predictor $f_w(x) = \sum_{g\in\mathcal{G}} w_g g(x)$, with empirical normalized ℓ_1 margin

$$\text{margin}_S(w) = \min_i \frac{y_i f_w(x_i)}{\|w\|_1},$$

defined whenever $\|w\|_1 > 0$. For a distribution D over the n training examples, the weighted error of a predictor g is $\sum_i D_i \mathbb{1}[g(x_i) \ne y_i]$.

Homework 6′s AdaBoost chooses its step size adaptively to drive training error down quickly. FixedBoost instead uses a fixed step size, so the ℓ_1 scale of the ensemble grows in a controlled way with the number of rounds. This makes the round count an explicit complexity knob and lets boosting be read as a search for a large normalized ℓ_1 margin, the theme of this part.

FixedBoost. Given a step size $\eta > 0$ and a number of rounds T, FixedBoost starts from $w_0 = 0$ and the uniform distribution D^1 over the n training examples. For each round $t = 1, ..., T$:

- pick a base predictor g_t maximizing the weighted correlation $\sum_i D^t_i y_i g_t(x_i)$;
- set $w_t = w_{t-1} + \eta e_{g_t}$, adding η to the weight of g_t;
- reweight the examples by $D^{t+1}_i \propto \exp\Big(-y_i f_{w_t}(x_i)\Big)$.

FixedBoost outputs the coefficient vector w_T.

1. **(10 points) Margin generalization.**

   Assume S is drawn i.i.d. from a population distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$. Prove a population 0/1 error bound for predictors with empirical normalized margin $\text{margin}_S(w) \geq \gamma$ for some $\gamma > 0$. Build on the generalization-gap bound from B3, specialized to the ramp surrogate from the week 7 lecture, rescaled to γ. State the final bound explicitly in terms of $|\mathcal{G}|$, n, γ, and δ.

   Solution.

   Normalize the coefficients by $v = \frac{w}{\|w\|_1}$. Then $\|v\|_1 = 1$ and the empirical margin condition says

   $$y_i f_v(x_i) \geq \gamma$$

   for every training example. Let

   $$r_\gamma(z) = \min\left(1, \left(1 - \frac{z}{\gamma}\right)_+\right)$$

   be the ramp loss at margin scale γ. It is $\frac{1}{\gamma}$-Lipschitz, is bounded in $[0, 1]$, upper-bounds $\mathbf{1}[z \leq 0]$, and equals 0 whenever $z \geq \gamma$.

   The feature map is $\varphi(x) = (g(x))_{g \in \mathcal{G}}$, so $\|\varphi(x)\|_\infty = 1$. Because $\mathcal{G}$ is symmetric, the signed coordinate functions are already contained in $\mathcal{G}$, and the ℓ_1 Rademacher bound with $B = 1$ gives

   $$\mathcal{R}_{\mathcal{D}^n}(\{x \to \langle v, \varphi(x)\rangle : \|v\|_1 \leq 1\}) \leq \sqrt{\frac{2\log|\mathcal{G}|}{n}}.$$

   After contraction by the ramp loss, with probability at least $1 - \delta$, uniformly over $\|v\|_1 \leq 1$,

   $$L_{\mathcal{D}}^{r_\gamma}(v) - L_S^{r_\gamma}(v) \leq \frac{2}{\gamma}\sqrt{\frac{2\log|\mathcal{G}|}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

   The ramp loss upper-bounds the 0/1 loss, so $L_{\mathcal{D}}^{0/1}(v) \leq L_{\mathcal{D}}^{r_\gamma}(v)$. The empirical margin assumption gives $L_S^{r_\gamma}(v) = 0$. Also, v is a positive rescaling of w, so $\operatorname{sign} f_w = \operatorname{sign} f_v$. Therefore

   $$L_{\mathcal{D}}^{0/1}(\operatorname{sign} f_w) \leq \frac{2}{\gamma}\sqrt{\frac{2\log|\mathcal{G}|}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

2. **(8 points) FixedBoost implication.**

   Use the following theorem about FixedBoost as a black box: if every distribution D over the training sample admits a base predictor of weighted error at most $\frac{1}{2} - \frac{\gamma}{2}$, then after $T = O\left(\frac{\log(n)}{\gamma^4}\right)$ rounds FixedBoost returns coefficients w_T satisfying

   $$\min_i y_i f_{w_T}(x_i) \geq \frac{\gamma}{2}\|w_T\|_1.$$

Derive a generalization guarantee for the classifier returned by FixedBoost.

Solution.

The black-box FixedBoost theorem says that under the stated weak-learning condition, after $T = O(\log(n)/\gamma^4)$ rounds,

$$\min_i y_i f_{w_T}(x_i) \geq \frac{\gamma}{2}\|w_T\|_1.$$

Thus the returned ensemble has empirical normalized margin at least $\frac{\gamma}{2}$. Applying the previous result with margin parameter $\frac{\gamma}{2}$ gives, with probability at least $1-\delta$ over the training sample,

$$L_{\mathcal{D}}^{0/1}\left(\operatorname{sign} f_{w_T}\right) \leq \frac{4}{\gamma}\sqrt{\frac{2\log|\mathcal{G}|}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

The bound is uniform over the ℓ_1-normalized ensemble class, so no additional union bound over the boosting trajectory is needed.

3. **(8 points) Weak learning and normalized margin.**

Encode the sample and base class as the matrix $A \in \{-1,+1\}^{n\times|\mathcal{G}|}$ with $A_{ig} = y_i g(x_i)$. The weak-learning edge of a base predictor g under a distribution D over the sample is $\frac{1}{2}$ minus its weighted error, its advantage over random guessing under D; the sample's best weak-learning edge is the largest γ such that every distribution D over the sample admits some $g \in \mathcal{G}$ with edge at least γ.

Prove the exact relationship between the best weak-learning edge and the best normalized ℓ_1 margin $\max_w \operatorname{margin}_S(w)$, using von Neumann's minimax theorem applied to A.

Solution.

Because $\mathcal{G}$ is symmetric, every coefficient vector can be replaced by a nonnegative coefficient vector over the signed base class without changing the predictor or increasing the ℓ_1 norm. After normalization, the best empirical normalized margin is therefore

$$\rho^* = \max_{q\geq 0,\, \sum_g q_g = 1} \min_i \sum_{g\in\mathcal{G}} q_g A_{ig}.$$

Writing the minimum over examples as a minimum over distributions p on the rows,

$$\rho^* = \max_{q\in\Delta_{\mathcal{G}}} \min_{p\in\Delta_n} p^\top A q.$$

By von Neumann's minimax theorem,

$$\rho^* = \min_{p\in\Delta_n} \max_{q\in\Delta_{\mathcal{G}}} p^\top A q = \min_{p\in\Delta_n} \max_{g\in\mathcal{G}} \sum_i p_i A_{ig}.$$

For a fixed row distribution p, the weighted error of g is

$$\sum_i p_i \mathbf{1}[g(x_i) \neq y_i].$$

Its weighted correlation is

$$\sum_i p_i y_i g(x_i) = 1 - 2\sum_i p_i \mathbf{1}[g(x_i) \neq y_i].$$

Thus the edge of g is one half of this correlation. If γ^*_{weak} is the best weak-learning edge of the sample, then

$$\gamma^*_{\text{weak}} = \frac{1}{2} \min_{p \in \Delta_n} \max_{g \in \mathcal{G}} \sum_i p_i A_{ig} = \frac{\rho^*}{2}.$$

Equivalently, the best normalized ℓ_1 margin equals twice the best weak-learning edge.

4. **(4 points) Connection to Homework 6.**

   Explain how this margin-based view refines the sparse-boosting analysis from Homework 6. Your answer should identify what changes in the comparator and what changes in the generalization argument.

   Solution.

   Homework 6 started from a sparse comparator with bounded coefficients. That comparator implied a weak coordinate edge of order $\frac{1}{\|w^*\|_1}$, and AdaBoost used the weak edge to drive empirical error to zero. Generalization was then argued through the VC dimension of the final sparse or boosted vote, with the number of selected weak rules entering the complexity term.

   The normalized-margin view changes the comparator from a sparse vector to an ℓ_1-normalized ensemble with large empirical margin. The generalization argument changes from a sparse-support VC bound to an ℓ_1 Rademacher margin bound, whose logarithmic dependence on $|\mathcal{G}|$ comes from the finite signed base class. FixedBoost is useful here because its fixed step size makes $\|w_T\|_1$ explicit, so the empirical quantity being certified is the normalized margin rather than only zero training error.

Part D: FixedBoost Margin Experiment (25 points)

The base class is the $2d$ signed coordinate predictors on Boolean inputs $x \in \{-1, +1\}^d$: the maps $x \to x_j$ and $x \to -x_j$ for $j \in [d]$. This is a finite symmetric base class in the sense of Part C. For a labeled sample $(x_1, y_1), ..., (x_n, y_n)$, form the sign matrix A with $A_{ig} = y_i g(x_i)$, equal to $+1$ when g is correct on example i and -1 otherwise. FixedBoost interacts with the data only through A: its weighted correlations, ensemble margins, and reweighting are all functions of A alone.

This experiment tests whether FixedBoost drives the normalized ℓ_1 margin toward the largest value attainable on a given A. Designing the data is part of the task: choose how the labeled examples (x_i, y_i) are generated, and describe it precisely enough that someone could reproduce your runs. You may use NumPy, plotting libraries, and a linear-programming solver for the optimal-margin computation, but the FixedBoost loop must be your own.

1. **(8 points) Implementation.**

   Implement FixedBoost from scratch, running directly on the matrix A, with a fixed step size. Track $\|w_t\|_1$ and the normalized empirical margin over rounds. You may reuse a boosting harness from Homework 6; state what differs from it.

   Solution.

   One possible implementation works entirely with the matrix A. Initialize $D_i^1 = \frac{1}{n}$, $w_0 = 0$, and choose a fixed step size $\eta > 0$. At round t, compute the correlations

   $$c_g^t = \sum_i D_i^t A_{ig},$$

   choose a column $g_t \in \operatorname{argmax}_g c_g^t$, set $w_t = w_{t-1} + \eta e_{g_t}$, and update

   $$D_i^{t+1} = \frac{\exp\left(-(Aw_t)_i\right)}{\sum_j \exp\left(-(Aw_t)_j\right)}.$$

   Track

   $$\|w_t\|_1 = \eta t$$

   and

   $$m_t = \min_i \frac{(Aw_t)_i}{\|w_t\|_1}.$$

   This differs from AdaBoost in Homework 6 because the step size is fixed rather than chosen from the current edge.

2. **(10 points) Optimal-margin comparison.**

   Formulate the optimal normalized ℓ_1 margin attainable on a matrix A as a linear program, and solve it. On the same matrix A, plot FixedBoost's normalized-margin trajectory against this optimal value. Do this for at least two instances: one with a large optimal margin, and one harder instance where the optimal margin is small. What structure in A makes the optimal margin large, and what makes it small? The relationship you proved in C3 is one way to reason about this. Construct the two instances accordingly, and report each construction.

   Solution.

   For a finite signed base class, the optimal normalized ℓ_1 margin is the value of the linear program

   $$\max_{\rho, q} \rho$$

   subject to

   $$q_g \geq 0, \quad \sum_g q_g = 1, \quad \sum_g A_{ig} q_g \geq \rho \ \text{ for every } \ i.$$

The LP variables are the distribution q over signed base predictors and the margin value ρ.

One large-margin instance is obtained by drawing arbitrary Boolean inputs and setting $y_i = x_{i,1}$. Then the signed coordinate predictor $g(x) = x_1$ is correct on every example, so the corresponding column of A is all $+1$. The optimal normalized margin is 1, and FixedBoost should choose that column immediately and keep normalized margin 1.

One small-margin construction is cyclic. Take even d, set $n = d$, let $y_i = +1$, and let the rows x_i be all cyclic shifts of a vector in $\{-1, +1\}^d$ with $\frac{d}{2} + 1$ entries equal to $+1$ and $\frac{d}{2} - 1$ entries equal to -1. Using only the positive coordinate columns uniformly gives every row average $\frac{2}{d}$, so the optimal margin is at least $\frac{2}{d}$. For any distribution over the signed coordinate columns, write its induced signed coordinate weights as $u \in \mathbb{R}^d$ with $\|u\|_1 \leq 1$. The average over the cyclic rows is

$$\left(\frac{2}{d}\right) \sum_j u_j \leq \frac{2}{d},$$

so the minimum row margin cannot exceed $\frac{2}{d}$. Thus the LP optimum is exactly $\frac{2}{d}$. This construction has small optimal margin because no single coordinate covers all rows; the mass must be spread across many weak rules.

The plot can show the FixedBoost trajectory m_t and the horizontal LP optimum on the same axes for both instances. On the perfect-coordinate instance, the trajectory reaches the optimum immediately. On the cyclic instance, the trajectory should move toward the small value $\frac{2}{d}$ as the reweighting shifts attention to rows with low current margin.

3. **(7 points) Interpretation.**

   Does FixedBoost's normalized margin approach the optimal value, and how fast? Compare the observed behavior with the rate implied by the FixedBoost theorem in C2. Identify at least one place where that theory is loose, vacuous, or silent relative to what you observe.

Solution.

The FixedBoost theorem predicts that if the optimal normalized margin is ρ, then the weak-learning edge is $\frac{\rho}{2}$ and a margin of order $\frac{\rho}{2}$ is guaranteed after $O(\log(n)/\rho^4)$ rounds, up to constants. For the cyclic construction with $\rho = \frac{2}{d}$, this becomes a bound of order $O(d^4 \log n)$ rounds to certify only a constant fraction of the optimum. In typical numerical runs this theorem is conservative: the observed trajectory may improve much sooner, the constants are large, and the theorem does not describe the detailed shape of the curve or the dependence on the chosen fixed step size η beyond the conditions needed for the proof.