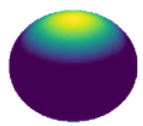
APPM 4490/5490
"Theory of Machine Learning"
Prof. Becker, spring 2026
Student projects

Forming and Analyzing Spherical Random Features

(Livia Betti)



Theorem 3 (Schoenberg's theorem on $\mathbb{S}^2$)**.** Let k be a continuous, zonal kernel on $\mathbb{S}^2$. Then k is positive definite on $\mathbb{S}^2$ if and only if

$$k(\langle \mathbf{x}, \mathbf{y} \rangle) = \sum_{l=0}^{\infty} a_l \, P_l(\langle \mathbf{x}, \mathbf{y} \rangle) = \sum_{l=0}^{\infty} a_l \, \frac{4\pi}{2\ell+1} \sum_{m=-l}^{l} Y_\ell^m(\mathbf{x}) \, Y_\ell^m(\mathbf{y}), \qquad a_l \geq 0, \quad \sum_{l=0}^{\infty} a_l < \infty, \tag{8}$$

where P_l are the Legendre polynomials and Y_ℓ^m the spherical harmonics.

Tight Bounds for Learning Polyhedra with a Margin

(Adithya Bhaskara)

In this report, we explore the recent result of Patel and Vempala [2026] and provide exposition building to their Theorem 1. Our goal is to prioritize intuition and high-level understanding over absolute rigor, so in our presentation, we may skip proofs as appropriate and skim through some details.

Patel and Vempala [2026] provide an algorithm to (ϵ, δ)-PAC learn intersections of k halfspaces, when given the promise that all negative points are at least ρR distance below some halfspace. Their algorithm runs in time

$$\text{poly}\left(k, \frac{1}{\epsilon}, \frac{1}{\rho}\right) \exp\left(O\left(\sqrt{n \log\left(\frac{1}{\rho}\right) \log k}\right)\right),$$

which is an improvement on existing work that has an exponential dependence on either k or $\frac{1}{\rho}$. Their algorithm also extends to continuous distibutions, where the promise is instead that *most* points are at least distance ρ away from the polyhedron (Theorem 3).

When Risk Bounds Become Risky: Rademacher Complexity in Non-IID Settings

(Chloe Chung)

5 Critical Analysis

Having examined the proof structure, we now examine the role of the IID assumption. Theorem 8 relies on IID in two distinct places, that being exchangeability, which allows for the symmetrization step, and independence, which is explicitly stated in the hypothesis of McDiarmid's inequality. In realistic machine learning settings where the IID assumption is violated, such as the Markov-reliant data which is foundational to reinforcement learning, both conditions simultaneously fail, meaning the resulting bound is no longer guaranteed to hold or is vacuous.

To demonstrate that the bounds derived in Theorem 8 can become vacuous in non-IID settings, consider a strongly correlated Markov chain with states 0 and 1 where the transition probabilities are $\mathbb{P}(X_i = 1 | X_{i-1} = 1) = 1 - \epsilon$ and $\mathbb{P}(X_i = 0 | X_{i-1} = 0) = 1 - \epsilon$ for small ϵ.

Generalization Bounds for Low-Rank Adaptation via Covering Numbers

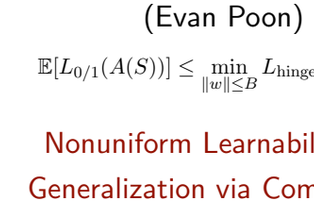
(Imad Dar)

Contributions. This paper closes that gap with a direct, covering-number argument. Specifically:

1. A covering number bound for the bilinear set $\{BA : \|B\|_F \leq s_B,\ \|A\|_F \leq s_A\}$ in Frobenius norm, with metric entropy scaling as $r(d+k)\log(s_B s_A/\epsilon)$ (Lemma 4.2).
2. A Rademacher complexity bound for the LoRA hypothesis class scaling as $\tilde{\mathcal{O}}(\sqrt{r(d+k)/n})$ (Theorem 4.5), compared to $\tilde{\mathcal{O}}(\sqrt{dk/n})$ for unconstrained fine-tuning.
3. A closed-form sample complexity corollary (Corollary 4.7) and the rank selection rule $r^\star \asymp n/(d+k)$.
4. Empirical validation on DistilBERT/SST-2 across a (r, n) grid, confirming the rank-dependence predicted by the bound and isolating a mismatch between the theorem's bounded-loss setting and the cross-entropy metric used in practice (Section 5).

Physics-Informed System Identification for Aerospace: Generalization in a Low-Data Regression Problem

(Brice Gillespie)



Learning Nonlinear Dynamical Systems from a Single Trajectory with Dependent Data

(Andy Gusty)

We now state our main result, which provides a finite-sample bound on the parameter estimation error for a nonlinear feature-based model learned from a single dependent trajectory. The proof combines ideas from [2] with martingale concentration tools from [4] that are accessible at the level of this course, which will give us a different (although likely less sharp) bounds than that found in [2].

Theorem 2 (Finite-Sample Identification Error)**.** *Under the stated assumptions, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2 \leq \frac{2L^2\sigma^2 p \log(2p/\delta)}{\kappa^2 T}.$$

Proof. We proceed in several steps.

Wasserstein K-means++ Clustering Algorithm

(James Hyun)

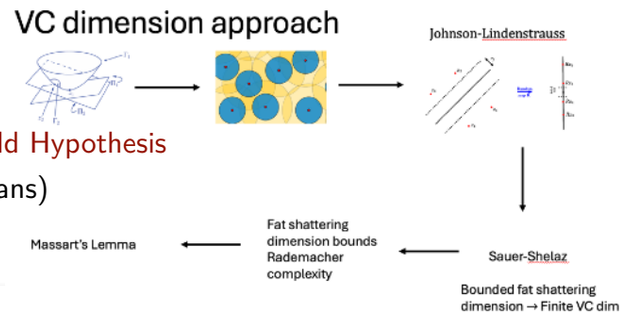
Proposition 1. Let $G = (V, E)$ and $G' = (V', E')$ with $|V| = n$ and $|V'| = m$ such that $n \neq m$. Via embedding Φ in (6), define

$$\mu_G = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \quad \text{and} \quad \mu_{G'} = \frac{1}{m}\sum_{j=1}^{m} \delta_{y_j},$$

where $x_i = \frac{\deg(i)}{\Delta_G} \in [0,1]$ and $y_j = \frac{\deg(j)}{\Delta_{G'}} \in [0,1]$ are the normalized degrees of nodes $i \in V$ and $j \in V'$. Here, a degree means the number of incident edges at a node and Δ_G is the maximum degree in a graph G. Then

$$W_1(\mu_G, \mu_{G'}) = \int_0^1 \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[x_i,\infty)}(t) - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{[y_j,\infty)}(t) \right| dt.$$

VC dimension approach



Testing the Manifold Hypothesis

(June Means)

A Brief Survey of Sequential Rademacher Complexity

(Chris Guthrie)

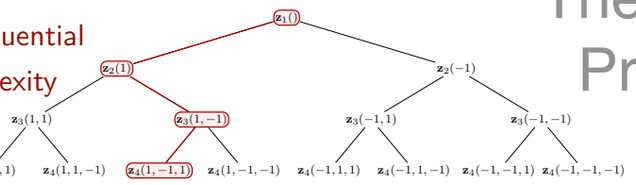


Figure 1: A $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth 4, with highlighted path $\sigma = (1, -1, 1, \ldots)$. At each depth, the left branch corresponds to $\sigma_t = 1$ and the right branch to $\sigma_t = -1$.

Student backgrounds:

- Applied Math (BS, MS, PhD)
- Math (BA)
- Computer Science (BS, MS, PhD)
- (and double-majors)

The Theory Behind Modern Image Denoising

(Robby Pawloski)

Score-Based Diffusion Models

Having outlined the general framework for denoising using deep learning, we now transition to focus on a particular type of model that has proven highly successful as a generative inverse problem solver: diffusion models. The main idea when using a diffusion model is to define a generative process that is the reverse of the noising process. Before going any further, though, we must briefly clarify some necessary terminology. A standard one-dimensional Wiener process is defined as a stochastic process $\{W_t\}_{t\geq 0}$ such that all of the following hold [6]:

1. $W_0 = 0$ almost surely.
2. The function $t \to W_t$ is almost surely continuous in t.
3. The process $\{W_t\}$ has stationary, independent increments.
4. The increment $W_{t+s} - W_s$ has normal distribution $\mathcal{N}(0, t)$.

Implicit Bias of Adam vs. SGD: Margin Geometry, Simplicity Bias, and Spurious Correlations

(Vishwas Kothari)

Proof mechanism: Adam as approximate sign descent. The Adam optimizer maintains two exponential moving averages for each coordinate j:

$$m_j^{(t)} = \beta_1 m_j^{(t-1)} + (1-\beta_1) g_j^{(t)}, \tag{15}$$

$$v_j^{(t)} = \beta_2 v_j^{(t-1)} + (1-\beta_2)(g_j^{(t)})^2, \tag{16}$$

where $g_j^{(t)} = \partial L / \partial \theta_j$ is the gradient at step t. The parameter update is:

Daily SPY Market Direction Prediction with Randomized Weighted Majority

(Bao Nguyen)

We approach financial market forecasting through the lens of online learning and frame market direction prediction as a binary classification problem. Using SPY to represent the overall U.S. stock market, we apply the randomized weighted majority algorithm to adaptively combine the daily predictions of six trading strategies as simple directional experts. This online learning approach guarantees sublinear regret relative to the best expert in hindsight, ensuring that average regret vanishes over time. To empirically validate this framework, we backtest the algorithm on daily SPY data from April 2024 to April 2026. Over this two-year period, the online learner closely tracks the best-performing fixed expert, achieving a final expected regret well below the theoretical worst-case bound. These results demonstrate that the randomized weighted majority algorithm provides a simple and grounded approach for market prediction, while also highlighting the distinction between minimizing directional classification errors and real-world trading performance.

Natural Gradient Descent: Theoretical Derivation, Practical Approximation, and Empirical Evaluation

(Leo Rasmussen)

A second-order Taylor expansion around $\delta\theta = 0$ gives the local quadratic approximation

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta+\delta\theta}) \approx \frac{1}{2}\delta\theta^\top F(\theta)\delta\theta,$$

where the Fisher information matrix

$$F(\theta) = E_{x\sim p_\theta}\left[\left(\nabla_\theta \log p_\theta(x)\right)\left(\nabla_\theta \log p_\theta(x)\right)^\top\right]$$

is the "metric tensor" of the statistical manifold of distributions. $F(\theta)$ is positive semi-definite and is the covariance of the "score" function $\nabla_\theta \log p_\theta$. This (Riemannian) geometry is the foundation of natural gradient methods.

Sample Complexity of Support Vector Machines

(Evan Poon)

$$\mathbb{E}[L_{0/1}(A(S))] \leq \min_{\|w\| \leq B} L_{\text{hinge}}(w) + \sqrt{\frac{8\rho^2 B^2}{m}}.$$

Nonuniform Learnability and Generalization via Compression

(Siraaj Sandhu)

Theorem 2.1 (Compression bound [1])**.** *Let $G_s = \{g_{A,s} \mid A \in \mathcal{A}\}$, where $\mathcal{A}$ is all sets of q parameters, each of which can take on r discrete values, and s is a helper string. Let S be a training set of size m. If a classifier f is (γ, S)-compressible via G_s, with helper string s, then $\exists A \in \mathcal{A}$ s.t. with high probability over S,*

$$L_0(g_{A,s}) \leq \hat{L}_\gamma(f) + O\left(\sqrt{\frac{q \log r}{m}}\right)$$

Proof. The justification provided by the paper is split across the main text and appendix, with either different symbols used for the same things or the same symbols used to mean different things across sections, making it hard to read. This theorem also constitutes the "main idea" of the paper, so it is worth restating more cleanly and with most of the missing steps filled in.

For fixed A, $\hat{L}_{0,S}(g_{A,s})$ is an average of iid bounded random variables with mean $L_0(g_{A,s})$. The authors use a "Chernoff bound," which turns out to be Hoeffding's inequality (B.6 in [2]). So, for some compressed model $g_{A,s}$ (fixed A), we have

The power of deeper networks for expressing natural functions

(Ian Soukup)

Theorem 2. *Let $p(\boldsymbol{x})$ denote the monomial $x_1^{r_1} x_2^{r_2} \ldots x_n^{r_n}$, with $d = \sum_{i=0}^{n} r_i$. Suppose that the nonlinearity σ has nonzero Taylor coefficients up to degree $2d$. Then we have:*

1. $m_1^{uniform}(p) = \prod_{i=1}^{n}(r_i + 1)$
2. $m^{uniform}(p) \leq \sum_{i=1}^{n}(7\lceil \log_2(r_i) + 4)\rceil)$

RKHS Methods and Applications to Kernel-Learned Transformations

(Jerry Wang)

This project aims to discuss the theoretical framework of reproducing kernel Hilbert space (RKHS) Transformations with regards to Kernel Methods, with a primary focus on explaining the theoretical framework of [1]. In particular, we establish a proof of the Generalized Representer Theorem by building on techniques discussed in class. Following that, this framework is applied to a particular problem of learning the Cole–Hopf transformation, as examined by [1], including both the theoretical formulation of the problem and some preliminary results.

The Non-Convergence of Adam

(David Zhou)

Theorem 3.1. *For any constant $\beta_1, \beta_2 \in [0, 1)$ such that $\beta_1 < \sqrt{\beta_2}$, there is a stochastic convex optimization problem for which Adam does not converge to the optimal solution. (Theorem 3 in [Reddi et al., 2019])*

Proof. Let $\delta > 0$, and C be a large enough constant chosen based on β_1, β_2, δ. Consider the following one dimensional stochastic optimization setting on domain [-1,1]. At each time step t, the function $f_t(x)$ is chosen i.i.d:

$$f_t(x) = \begin{cases} Cx & \text{with probability } p = \frac{1+\delta}{C+1} \\ -x & \text{with probability } 1-p \end{cases}.$$

Then we have $\mathbb{E}[f_t(x)] = F(x) = \delta x$. Notice here $F(x)$ is convex and its