Learning with invariances in random features and kernel models

Theodor Misiakiewicz

Stanford University

July 21st, 2021

Conference on Learning Theory 2021

Joint work with Song Mei (UC Berkeley) and Andrea Montanari (Stanford)

Learning with invariances

- In many learning tasks, the data present some **natural symmetries**.

  E.g., image recognition: labels are **invariant under translation** of the images.

- *Design predictive models that take advantage of these symmetries to make a more efficient use of data.*

- For example, **convolutional networks** are believed to owe their success to their ability to encode translation invariance.

- Empirically, models that exploit invariances perform better that models that do not.

Goal:

Quantify the performance gain achieved by invariant architectures over non-invariant ones.

- We focus on Random Features and kernel models.

Setting

- **Data:** $\boldsymbol{x} \sim \text{Unif}(\mathcal{A}_d)$, $\quad \mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$ or $\mathcal{A}_d = \{-1,+1\}^d$.

- **Invariance group:** $\mathcal{G}_d$ subgroup of orthogonal group $\mathcal{O}(d)$ (that preserves $\mathcal{A}_d$).

- **Goal:** learn a $\mathcal{G}_d$-invariant function $f_\star$ (i.e., $f_\star(g \cdot \boldsymbol{x}) = f_\star(\boldsymbol{x})$ for all $g \in \mathcal{G}_d$)

  Given iid samples $\{(y_i, \boldsymbol{x}_i)\}_{i \leq n}$:

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim_{iid} \text{Unif}(\mathcal{A}_d), \qquad \mathbb{E}[\varepsilon_i] = 0, \quad \mathbb{E}[\varepsilon_i^2] \leq \tau^2.$$

Example: the cyclic group $\mathcal{G}_d = \{g_0, g_1, \ldots, g_{d-1}\}$:

$$g_i \cdot \boldsymbol{x} = (x_{d-i+1}, x_{d-i+2}, \ldots, x_d, x_1, x_2, \ldots, x_{d-i}).$$

Target function: $f_\star(\boldsymbol{x}) = \sum_{i=1}^d x_i x_{i+1}$.

Stylized model for an image label $y = f_\star(\boldsymbol{x})$ invariant by translation of image $\boldsymbol{x}$.

Invariant random features and kernel models

- Random Features model: $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N)$ with $(\sqrt{d}\boldsymbol{w}_i) \sim_{iid} \mathrm{Unif}(\mathcal{A}_d)$ fixed,

$$\hat{f}_{\mathsf{RF}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle) \quad \to \quad \hat{f}_{\mathsf{RF}}^{\mathrm{inv}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \int_{\mathcal{G}_d} \sigma(\langle \boldsymbol{w}_j, g \cdot \boldsymbol{x} \rangle)\, \pi_d(\mathrm{d}g).$$

$$\hat{\boldsymbol{a}}^{\mathrm{inv}}(\lambda) = \arg\min_{\boldsymbol{a} \in \mathbb{R}^N} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{f}_{\mathsf{RF}}^{\mathrm{inv}}(\boldsymbol{x}_i; \boldsymbol{a}) \right)^2 + N\lambda \|\boldsymbol{a}\|_2^2 \right\}.$$

- Kernel Ridge regression:

$$H(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d) \quad \to \quad H^{\mathrm{inv}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathcal{G}_d} h(\langle \boldsymbol{x}_1, g \cdot \boldsymbol{x}_2 \rangle / d)\, \pi_d(\mathrm{d}g).$$

$$\hat{f}_{\lambda}^{\mathrm{inv}} = \arg\min_{\hat{f} \in \mathcal{H}^{\mathrm{inv}}} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{f}^{\mathrm{inv}}(\boldsymbol{x}_i) \right)^2 + \lambda \|\hat{f}^{\mathrm{inv}}\|_{\mathcal{H}^{\mathrm{inv}}}^2 \right\}.$$

Example of the cyclic group

- Cyclic group $\mathcal{G}_d = \{g_0, g_1, \dots, g_{d-1}\}$.
- **Random features models:**
  - Standard RF: $\hat{f}_{\rm RF}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^N a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)$.
  - Cyclic invariant RF model:

$$f_{\rm RF}^{\rm inv}(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{d} \sum_{j=1}^N a_j \sum_{k=1}^d \sigma(\langle \boldsymbol{w}_j, g_k \cdot \boldsymbol{x} \rangle).$$

    Two-layers CNN with global average pooling and patchsize d: non-linear convolution of N weights $\boldsymbol{w}_j \in \mathbb{R}^d$.
- **Kernel models:**
  - $H(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d)$: NTK of fully-connected NNs.
  - $H^{\rm inv}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{d} \sum_{k=1}^d h(\langle \boldsymbol{x}_1, g_k \cdot \boldsymbol{x}_2 \rangle / d)$: NTK of 2-layers CNN with global pooling.

Degeneracy of group $\mathcal{G}_d$

- Gain in approximation and generalization error characterized by **'degeneracy'** of $\mathcal{G}_d$.

Groups of degeneracy $\alpha \in \mathbb{R}_{\geq 0}$

- $V_{d,k}$: subspace of degree-k polynomials orthogonal to degree-$(k-1)$ polynomials in $L^2(\mathcal{A}_d)$.
- $V_{d,k}(\mathcal{G}_d)$: subspace of $V_{d,k}$ of $\mathcal{G}_d$-invariant polynomials.

$\mathcal{G}_d$ has degeneracy α if for any $k \geq \alpha$, we have $\dim(V_{d,k})/\dim(V_{d,k}(\mathcal{G}_d)) \asymp d^\alpha$.

- d^α: 'effective dimension' of the group seen through its action on polynomials.
- $\alpha = 1$ for cyclic group.
- Not necessarily equal to the size of the group:

  E.g., translation invariance on band-limited signals $\mathrm{Sft}_d = \{g_u, u \in [0,1]\}$

$$g_u \cdot \boldsymbol{x} = (x_1, \cos(2\pi u)x_2 + \sin(2\pi u)x_3, -\sin(2\pi u)x_2 + \cos(2\pi u)x_3, \ldots).$$

  Sft_d has degeneracy $\alpha = 1$.

Test error of learning with RF model (I)

- $\mathcal{G}_d$-invariant $f_\star$ with $\mathcal{G}_d$ of degeneracy α: given iid samples $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$,

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim_{iid} \mathrm{Unif}(\mathcal{A}_d), \qquad \mathbb{E}[\varepsilon_i] = 0, \quad \mathbb{E}[\varepsilon_i^2] \le \tau^2.$$

- Test error: $R_{\mathrm{RF}}(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \mathbb{E}_{\boldsymbol{x}}\Big\{ \Big(f_\star(\boldsymbol{x}) - \hat{f}_{\mathrm{RF}}(\boldsymbol{x}, \hat{\boldsymbol{a}}(\lambda))\Big)^2 \Big\}$.

Theorem (Mei, Misiakiewicz, Montanari, 2021)

For σ following some conditions. Then

- **Overparametrized regime:** $N \ge n \cdot d^\delta$, $\lambda = O_d(1)$,

$$d^{\ell+\delta} \le n \le d^{\ell+1-\delta}, \qquad R_{\mathrm{RF}}(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \|\mathrm{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot),$$

$$d^{\ell+\delta}/d^\alpha \le n \le d^{\ell+1-\delta}/d^\alpha, \qquad R_{\mathrm{RF}}^{\mathrm{inv}}(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda/d^\alpha) = \|\mathrm{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot).$$

- **Underparametrized regime:** $n \ge N \cdot d^\delta$, $\lambda = O_d(n/N)$,

$$d^{\ell+\delta} \le N \le d^{\ell+1-\delta}, \qquad R_{\mathrm{RF}}(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \|\mathrm{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot),$$

$$d^{\ell+\delta}/d^\alpha \le N \le d^{\ell+1-\delta}/d^\alpha, \qquad R_{\mathrm{RF}}^{\mathrm{inv}}(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda/d^\alpha) = \|\mathrm{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot).$$

$\mathrm{P}_{>\ell}$: projection orthogonal to the subspace of degree-ℓ polynomials.

(Note that for $\alpha > 1$, we need to add the condition $n, N \ge d^{O(\alpha)}$.)

Test error of learning with RF model (II)

- For $\mathcal{G}_d$ group of degeneracy α, we save a factor d^α in sample size and number of hidden units to achieve the same test error as for non-invariant model.
- For the cyclic group, we save a factor d in sample size and number of hidden units.
- **Conditions on σ:** the theorem is a consequence of a general framework in [Mei, **M.**, Montanari,'21]
  - For the cylcic group, we checked the assumptions for σ $(\ell + 1)$-differentiable.
  - For general groups of degeneracy α, we take σ to be a polynomial.

Deferred weaker conditions to future work.

- Test error: $R_{\rm KR}(f_\star, \boldsymbol{X}, \lambda) = \mathbb{E}_{\boldsymbol{x}}\Big\{ \Big(f_\star(\boldsymbol{x}) - \hat{f}_\lambda(\boldsymbol{x}) \Big)^2 \Big\}$.

Theorem (Mei, Misiakiewicz, Montanari, 2021)

For h following some conditions and $\lambda = O_d(1)$,

$$d^{\ell+\delta} \le n \le d^{\ell+1-\delta}, \qquad R_{\rm KR}(f_\star, \boldsymbol{X}, \lambda) = \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot),$$

$$d^{\ell+\delta}/d^{\alpha} \le n \le d^{\ell+1-\delta}/d^{\alpha}, \qquad R_{\rm KR}^{\rm inv}(f_\star, \boldsymbol{X}, \lambda/d^{\alpha}) = \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o_{d,\mathbb{P}}(\cdot).$$

- Gain of factor d^{α} in sample size to achieve the same test error as non-invariant KRR.

$$f_{\text{lin}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x_i, \qquad f_{\text{quad}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x_i x_{i+1}, \qquad f_{\text{cube}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x_i x_{i+1} x_{i+2}.$$

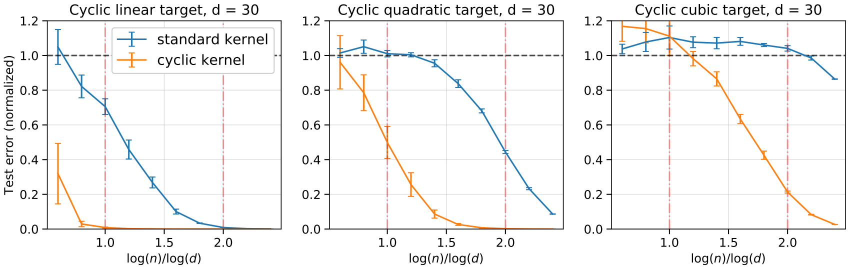
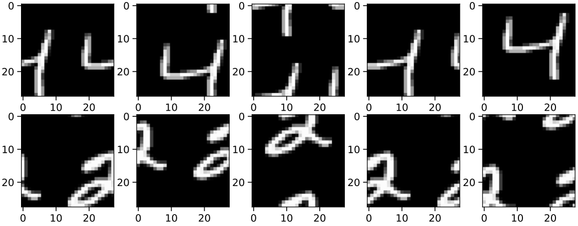


Figure: Test error of KRR with cyclic invariant kernel and inner product kernel.

Cyclic invariant MNIST



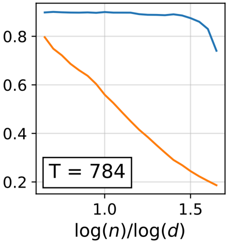


Figure: Test accuracy against number of samples (orange: cyclic kernel, blue: standard kernel).

Symmetrization and data augmentation

We compare 4 approaches: (a) Standard KRR. (b) Invariant KRR. (c) Output symmetrization of standard KRR. (d) Standard KRR with data augmentation.

(c) **Output symmetrization** of standard KRR $f_{K,n}$,

$$\mathcal{S}\hat{f}_{K,n}(\boldsymbol{x}) = \int_{\mathcal{G}_d} \hat{f}_{K,n}(g \cdot \boldsymbol{x}) \pi(\mathrm{d}g).$$

For $d^{\ell+\delta} \leq n \leq d^{\ell+1-\delta}$, $\|f_\star - \mathcal{S}\hat{f}_{K,n}\|_{L^2}^2 \approx \|f_\star - \hat{f}_{K,n}\|_{L^2}^2 = \|\mathrm{P}_{>\ell} f_\star\|_{L^2}^2 + o_d(\cdot)$.

Test error: $(c) \approx (a)$.

(d) **Data augmentation:** add to the training set $(y_i, g \cdot \boldsymbol{x}_i)$, $\forall g \in \mathcal{G}_d, \forall i \in [n]$.

Standard KRR with data augmentation $\iff$ invariant KRR [Li et al., 2019].

Test errors: $(b) = (d) \ll (c) \approx (a)$

Summary

- **Goal:** learn invariant function $f_\star$ with invariance group $\mathcal{G}_d$ subgroup of $\mathcal{O}(d)$.
- Standard RF and Kernel models and their invariant counterparts by group averaging.
- We identified the degeneracy α of $\mathcal{G}_d$ as the measure of performance gain:

$$\forall k \geq \alpha, \qquad \frac{\#\text{ degree k polynomials}}{\#\ \mathcal{G}_d\text{-invariant degree k polynomials}} \asymp d^\alpha.$$

  E.g., cyclic group $\alpha = 1$.

- Using invariant models leads to a factor d^α improvement in sample size and number of hidden units.
- Diagonalization of invariant kernels plus a representation lemma to count the number of invariant polynomials that might be of independent interest.

Thank you!