Learning with invariances in random features and kernel models

Song Mei [1] Theodor Misiakiewicz [2] Andrea Montanari [2, 3]

[1]Department of Statistics, UC Berkeley [2]Department of Statistics, Stanford University [3]Department of Electrial Engineering, Stanford University

Introduction

- In many learning tasks, the **data present some natural symmetries** (e.g., labels are invariant under translation of the images in image recognition tasks).
- One important goal of machine learning has been to design **predictive models that take advantage of these symmetries** to make a more efficient use of data.
- For instance, **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.

Focus of this work:

Quantifying the performance gain of using invariant architectures over non-invariant ones in random features and kernel models.

Setting and models

- **Data:** $\boldsymbol{x} \sim \mathsf{Unif}(\mathcal{A}_d)$, $\mathcal{A}_d = \mathbb{S}^{d-1}(\sqrt{d})$ or $\mathcal{A}_d = \{-1,+1\}^d$.
- **Invariance group:** $\mathcal{G}_d$ subgroup of orthogonal group $\mathcal{O}(d)$ (that preserves the hypercube if $\mathcal{A}_d = \{-1,+1\}^d$).
- **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 $\{y_i, \boldsymbol{x}_i\}_{i \le n}$ with $\boldsymbol{x}_i \sim_{iid} \mathsf{Unif}(\mathcal{A}_d)$ and

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0, \ \ \mathbb{E}[\varepsilon_i^2] \le \tau^2.$$

Models:

- **Random Features models:** $(\sqrt{d}\boldsymbol{w}_i) \sim_{iid} \mathsf{Unif}(\mathcal{A}_d)$ fixed,

$$\hat{f}_{\mathsf{RF}}(\boldsymbol{x};\boldsymbol{a}) = \sum_{j=1}^N a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)$$

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

where π_d is the Haar measure on the group $\mathcal{G}_d$.
Fit the coefficients with Ridge Regression (RFRR):

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

- **Kernel models:** inner-prod. kernel $H(\boldsymbol{x},\boldsymbol{z}) = h\left(\frac{\langle \boldsymbol{x},\boldsymbol{z}\rangle}{d}\right)$,

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

Fit the function with Kernel Ridge Regression (KRR):

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

Example: 2-layer CNN

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

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

- Cyclic invariant RF model:

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

  2-layers CNN with global average pooling & filters $\boldsymbol{w}_j \in \mathbb{R}^d$.
- Inner-prod. kernel: NTK of fully-connected NNs; vs Cyclic invariant kernel: NTK of 2-layer CNN with global pooling. → performance gap FC-NN vs. CNN in kernel regime.

Degeneracy of the invariance group

- Identify **degeneracy** of $\mathcal{G}_d$ as the measure of the approx. and generalization power gain of using invariant models.
- $V_{d,k}$: 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.

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

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

- d^α: **'effective dimension'** of the action of the group.
- $\alpha = 1$ for cyclic group.
- Not necessarily equal to the size of $\mathcal{G}_d$: e.g., translation invariance on band-limited signals $\mathrm{Sft}_d = \{g_u, u \in [0,1]\}$

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

  Sft_d has degeneracy $\alpha = 1$.

Counting invariant polynomials

- $\{Y_{ks}\}_{s \le B_{d,k}}$ orthonormal basis of $V_{d,k}$ ($B_{d,k} = \dim(V_{d,k})$).
- $\{\overline{Y}_{ks}\}_{s \le D_{d,k}}$ orth. basis of $V_{d,k}(\mathcal{G}_d)$ ($D_{d,k} = \dim(V_{d,k}(\mathcal{G}_d))$).
- Gegenbauer polynomial on $\mathcal{A}_d$ of degree-k:

$$Q_k(\langle \boldsymbol{x},\boldsymbol{z}\rangle) = \frac{1}{B_{d,k}} \sum_{s \le B_{d,k}} Y_{ks}(\boldsymbol{x}) Y_{ks}(\boldsymbol{z}).$$

Representation lemma

Lemma 1 ([3]) *We have*

$$\frac{1}{D_{d,k}} \sum_{s \le D_{d,k}} \overline{Y}_{ks}(\boldsymbol{x})\overline{Y}_{ks}(\boldsymbol{z}) = \frac{B_{d,k}}{D_{d,k}} \int_{\mathcal{G}_d} Q_k(\langle \boldsymbol{x}, g\cdot\boldsymbol{z}\rangle)\, \pi_d(\mathrm{d}g).$$

- To compute degeneracy, it is sufficient to show for all $k \ge \alpha$:

$$\mathbb{E}_{\boldsymbol{x}\sim\mathsf{Unif}(\mathcal{A}_d)}\Big[\int_{\mathcal{G}_d} Q_k(\langle \boldsymbol{x}, g\cdot\boldsymbol{z}\rangle)\, \pi_d(\mathrm{d}g)\Big] = \frac{D_{d,k}}{B_{d,k}} = \Theta_d(d^{-\alpha}).$$

Test error of invariant models

- Let $f_\star$ be $\mathcal{G}_d$-invariant with $\mathcal{G}_d$ group of degeneracy α.
- Test error with square loss:

$$R(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \mathbb{E}_{\boldsymbol{x}}\Big\{ \big(f_\star(\boldsymbol{x}) - \hat{f}_{\mathsf{RF}}(\boldsymbol{x}; \hat{\boldsymbol{a}}(\lambda))\big)^2 \Big\}.$$

Test error of RFRR

Theorem 1 ([3]) *Assume* $\max(N/n, n/N) \ge d^\delta$ *and* $\lambda = O_d(1 \vee (N/n))$, σ *stisfies some conditions, then:*

- *(Standard RF) If* $d^{\ell+\delta} \le \min(N, n) \le d^{\ell+1-\delta}$,

$$R(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \|\mathsf{P}_{>\ell} f_\star\|^2_{L^2} + o_{d,\mathbb{P}}(\cdot).$$

- *(Invariant RF) If* $d^{\ell+\delta}/d^\alpha \le \min(N, n) \le d^{\ell+1-\delta}/d^\alpha$,

$$R^{\mathrm{inv}}(f_\star, \boldsymbol{X}, \boldsymbol{W}, \lambda/d^\alpha) = \|\mathsf{P}_{>\ell} f_\star\|^2_{L^2} + o_{d,\mathbb{P}}(\cdot).$$

$\mathsf{P}_{>\ell}$*: projection orthogonal to the space of degree-*ℓ *polynomials.*

- RFRR learns the best degree-ℓ polynomial approx. to $f_\star$.
- Same result for KRR as above with $N = \infty$: invariant KRR saves a factor d^α in sample size compared to standard KRR.

Invariant RF saves a factor d^α in sample size and number of hidden units to achieve same test error as std. RF.

Assumptions on σ

- Results: **consequence of a general framework** in [2].
- Technical general conditions of [2] checked for
  - Cyclic group and σ assumed $(\ell+1)$-differentiable.
  - General groups of degeneracy α and σ polynomial.
- Deferred weaker conditions to future work (if σ diff., our proof techniques generalize to subgroups of permutations).

Symmetrization and data augmentation

Compare 4 approaches to learning invariant models:

(a) **Standard KRR:** with inner-prod. kernel.

(b) **Invariant KRR:** with invariant kernel ('intrinsic approach': invariance directly enforced in the model).

(c) **Output symmetrization:** take $\hat{f}_\lambda$ solution of standard KRR and symmetrize it:

$$\hat{f}^{\mathrm{inv}}_\lambda(\boldsymbol{x}) := \int_{\mathcal{G}_d} \hat{f}_\lambda(g\cdot\boldsymbol{x})\, \pi_d(\mathrm{d}g).$$

Does not significantly improve on standard KRR.

(d) **Full data augmentation:** add $\{(y_i, g\cdot\boldsymbol{x}_i)\}_{i \le n, g \in \mathcal{G}_d}$ to the training set with standard KRR.

⇒ this is equivalent to invariant KRR [1].

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

Sketch of the proof for KRR

- Inner-prod. kernels have eigenspaces $V_{d,k}$:

$$H_d(\boldsymbol{x},\boldsymbol{z}) = \sum_{k=0}^\infty \xi^2_{d,k} \sum_{s \le B_{d,k}} Y_{ks}(\boldsymbol{x})Y_{ks}(\boldsymbol{z}).$$

- Space $V_{d,k}$ is **preserved under the action of** $\mathcal{G}_d$:

$$H^{\mathrm{inv}}_d(\boldsymbol{x},\boldsymbol{z}) = \sum_{k=0}^\infty \xi^2_{d,k} \sum_{s \le D_{d,k}} \overline{Y}_{ks}(\boldsymbol{x})\overline{Y}_{ks}(\boldsymbol{z}).$$

- H^{inv}_d has the same eigenvalues $\xi^2_{d,k} = \Theta_d(d^{-k})$ as H_d but with degeneracy lower by a factor $B_{d,k}/D_{d,k} = \Theta_d(d^\alpha)$.

- **Theorem** [2]: kernel eigenval. $\{\lambda_{d,k}\}$ in decreas. order, eigenvect. $\{\psi_k\}$ + technical conditions. Let $m \in \mathbb{N}$ s.t.

$$\lambda_{d,m+1} \cdot n^{1+\delta} \le \sum_{k \ge m+1} \lambda_{d,k}, \qquad m \le n^{1-\delta}.$$

  Define $s^{\mathrm{eff}} = \lambda + \sum_{k \ge m+1} \lambda_k$.
  Then **KRR acts as shrinkage operator**, i.e.,

$$\hat{f}_\lambda(\boldsymbol{x}) \approx \sum_{k \ge 1} \frac{\lambda_{d,k}}{\lambda_{d,k} + s^{\mathrm{eff}}/n} \cdot \langle f_\star, \psi_k\rangle_{L^2} \cdot \psi_k(\boldsymbol{x}).$$

  Learn eigendirection if $\lambda_{d,k} \gg \frac{s^{\mathrm{eff}}}{n}$, not at all if $\lambda_{d,k} \ll \frac{s^{\mathrm{eff}}}{n}$.

- **Std KRR:** $m = \#\{Y_{ks}\}_{k \le \ell} = \Theta_d(d^\ell)$, $s^{\mathrm{eff}} = \Theta_d(1)$. Hence if $d^{\ell+\delta} \le n \le d^{\ell+1-\delta}$, learns degree-$\ell$ polynomial approx.

- **Inv KRR:** $m = \#\{\overline{Y}_{ks}\}_{k \le \ell} = \Theta_d(d^{\ell-\alpha})$, $s^{\mathrm{eff}} = \Theta_d(d^{-\alpha})$. Hence if $d^{\ell-\alpha+\delta} \le n \le d^{\ell+1-\alpha-\delta}$, learns degree-$\ell$ approx.

Numerical illustration

- Data $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ with $d = 30$.
- Target functions invariant w.r.t. cyclic group.
- Degeneracy $\alpha = 1$, hence save factor d in sample size.
- $f_{\mathrm{lin}} = \sum_{i \le d} x_i$, $f_{\mathrm{quad}} = \sum_{i \le d} x_i x_{i+1}$, $f_{\mathrm{cube}} = \sum_{i \le d} x_i x_{i+1} x_{i+2}$.

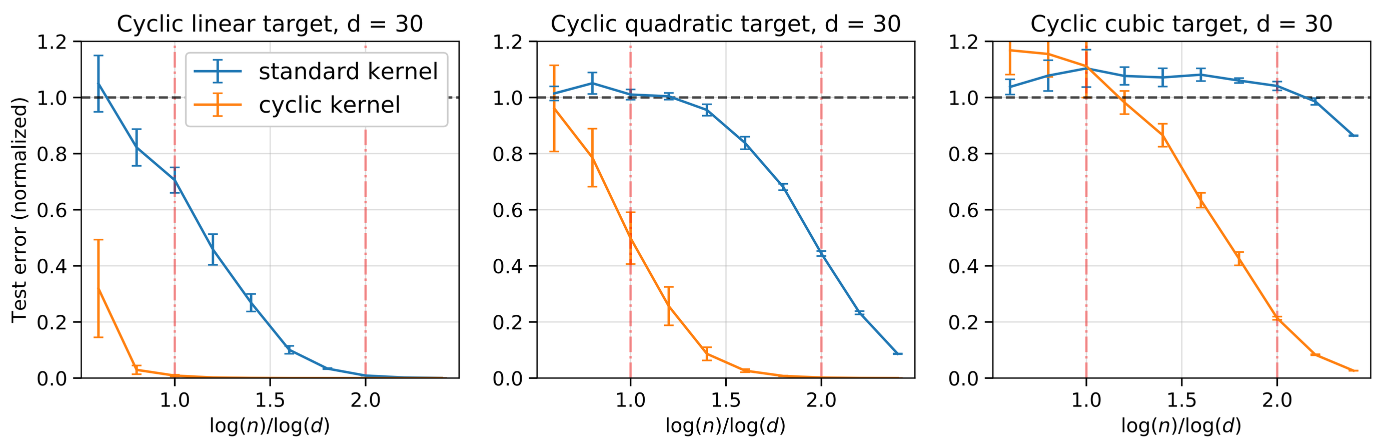

Figure 1: Normalized test error of KRR with cyclic vs standard kernels.

Bibliography

[1] Z. Li, R. Wang, D. Yu, S. S. Du, W. Hu, R. Salakhutdinov, and S. Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

[2] S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021.

[3] S. Mei, T. Misiakiewicz, and A. Montanari. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021.