Minimum complexity interpolation in random features models

Theodor Misiakiewicz

Stanford University

June 15th, 2021

Youth in High Dimensions 2021

Joint work with Michael Celentano and Andrea Montanari (Stanford)

RKHS methods (NNs in the 'lazy training' regime)

- **Supervised learning setting:** $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$, $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \mathbb{P})$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$.
- **Kernel machines:** starting from a weight space (Ω, μ)
  - Featurization map: $\phi(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathbb{R}$, e.g., $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$,

$$\mathcal{F}_2 = \Big\{ f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) : \|a\|_{L^2}^2 = \int_\Omega |a(\boldsymbol{w})|^2 \mu(\mathrm{d}\boldsymbol{w}) \Big\}.$$

  Associated kernel: $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_\Omega \phi(\boldsymbol{x}_1; \boldsymbol{w})\phi(\boldsymbol{x}_2; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.

  - Convex loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n = \arg\min_{f\in\mathcal{F}_2} \Big\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \Big\}.$$

  - Can be solved efficiently despite $\mathcal{F}_2$ being infinite-dimensional.
    The 'representer theorem':

$$\hat{a}(\boldsymbol{w}) = \sum_{i=1}^n c_i \phi(\boldsymbol{x}_i; \boldsymbol{w}).$$

RKHS methods (NNs in the 'lazy training' regime)

- **Supervised learning setting:** $\{(y_i, \boldsymbol{x}_i)\}_{i \in [n]}$, $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \mathbb{P})$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$.
- **Kernel machines:** starting from a weight space (Ω, μ)
  - Featurization map: $\phi(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathbb{R}$, e.g., $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)$,

$$\mathcal{F}_2 = \Big\{ f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w}) \phi(\boldsymbol{x}; \boldsymbol{w}) \mu(\mathrm{d}\boldsymbol{w}) : \|a\|_{L^2}^2 = \int_\Omega |a(\boldsymbol{w})|^2 \mu(\mathrm{d}\boldsymbol{w}) \Big\}.$$

    Associated kernel: $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_\Omega \phi(\boldsymbol{x}_1; \boldsymbol{w}) \phi(\boldsymbol{x}_2; \boldsymbol{w}) \mu(\mathrm{d}\boldsymbol{w})$.

  - Convex loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}_2} \Big\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \Big\}.$$

  - Can be solved efficiently despite $\mathcal{F}_2$ being infinite-dimensional.
    The 'representer theorem':

$$\hat{a}(\boldsymbol{w}) = \sum_{i=1}^n c_i \phi(\boldsymbol{x}_i; \boldsymbol{w}).$$

RKHS methods (NNs in the 'lazy training' regime)

- **Supervised learning setting:** $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$, $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \mathbb{P})$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$.
- **Kernel machines:** starting from a weight space (Ω, μ)
  - Featurization map: $\phi(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathbb{R}$, e.g., $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$,

$$\mathcal{F}_2 = \left\{ f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) : \|a\|_{L^2}^2 = \int_\Omega |a(\boldsymbol{w})|^2 \mu(\mathrm{d}\boldsymbol{w}) \right\}.$$

    Associated kernel: $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_\Omega \phi(\boldsymbol{x}_1; \boldsymbol{w})\phi(\boldsymbol{x}_2; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.

  - Convex loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}_2} \left\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \right\}.$$

  - Can be solved efficiently despite $\mathcal{F}_2$ being infinite-dimensional. The 'representer theorem':

$$\hat{a}(\boldsymbol{w}) = \sum_{i=1}^n c_i \phi(\boldsymbol{x}_i; \boldsymbol{w}).$$

RKHS methods (NNs in the 'lazy training' regime)

- **Supervised learning setting:** $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$, $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \mathbb{P})$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$.
- **Kernel machines:** starting from a weight space (Ω, μ)
  - Featurization map: $\phi(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathbb{R}$, e.g., $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$,

$$\mathcal{F}_2 = \left\{ f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) : \|a\|_{L^2}^2 = \int_\Omega |a(\boldsymbol{w})|^2 \mu(\mathrm{d}\boldsymbol{w}) \right\}.$$

  Associated kernel: $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_\Omega \phi(\boldsymbol{x}_1; \boldsymbol{w})\phi(\boldsymbol{x}_2; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.

  - Convex loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}_2} \left\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \right\}.$$

  - Can be solved efficiently despite $\mathcal{F}_2$ being infinite-dimensional.
    The 'representer theorem':

$$\hat{a}(\boldsymbol{w}) = \sum_{i=1}^n c_i \phi(\boldsymbol{x}_i; \boldsymbol{w}).$$

RKHS methods (NNs in the 'lazy training' regime)

- **Supervised learning setting:** $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$, $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \mathbb{P})$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$.

- **Kernel machines:** starting from a weight space (Ω, μ)

  - Featurization map: $\phi(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathbb{R}$, e.g., $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$,

$$\mathcal{F}_2 = \left\{ f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) : \|a\|_{L^2}^2 = \int_\Omega |a(\boldsymbol{w})|^2 \mu(\mathrm{d}\boldsymbol{w}) \right\}.$$

    Associated kernel: $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_\Omega \phi(\boldsymbol{x}_1; \boldsymbol{w})\phi(\boldsymbol{x}_2; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.

  - Convex loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}_2} \left\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \right\}.$$

  - Can be solved efficiently despite $\mathcal{F}_2$ being infinite-dimensional.
    The 'representer theorem':

$$\hat{a}(\boldsymbol{w}) = \sum_{i=1}^n c_i \phi(\boldsymbol{x}_i; \boldsymbol{w}).$$

RKHS methods (NNs in the 'lazy training' regime)

- **Supervised learning setting:** $\{(y_i, \boldsymbol{x}_i)\}_{i \in [n]}$, $\boldsymbol{x}_i \sim_{iid} (\mathcal{X}, \mathbb{P})$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$.
- **Kernel machines:** starting from a weight space (Ω, μ)
  - Featurization map: $\phi(\cdot; \boldsymbol{w}) : \mathcal{X} \to \mathbb{R}$, e.g., $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)$,

$$\mathcal{F}_2 = \left\{ f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w}) \phi(\boldsymbol{x}; \boldsymbol{w}) \mu(\mathrm{d}\boldsymbol{w}) : \|a\|_{L^2}^2 = \int_\Omega |a(\boldsymbol{w})|^2 \mu(\mathrm{d}\boldsymbol{w}) \right\}.$$

    Associated kernel: $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_\Omega \phi(\boldsymbol{x}_1; \boldsymbol{w}) \phi(\boldsymbol{x}_2; \boldsymbol{w}) \mu(\mathrm{d}\boldsymbol{w})$.

  - Convex loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}_2} \left\{ \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i; a)) + \lambda \|a\|_{L^2}^2 \right\}.$$

  - Can be solved efficiently despite $\mathcal{F}_2$ being infinite-dimensional.
    The 'representer theorem':

$$\hat{a}(\boldsymbol{w}) = \sum_{i=1}^n c_i \phi(\boldsymbol{x}_i; \boldsymbol{w}).$$

Curse of dimensionality

- $\mathcal{F}_2(R) = \left\{ f(\boldsymbol{x}; a) : \|a\|_{L^2} \leq R \right\}$, $f_\star \in \mathcal{F}_2(R)$,

$$\text{Generalization error of learning } f_\star \text{ from } n \text{ samples } \leq C \frac{R}{\sqrt{n}}.$$

- Kernel methods suffer from the **curse of dimensionality**.

- $\boldsymbol{x}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$, $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)$.

  Any target function $f_\star \in L^2(\mathcal{X})$, number of samples: $d^k \ll n \ll d^{k+1}$ [GMMM,'19]

$$\text{Test error with squared loss:} \quad \|f_\star - \hat{f}_n\|_{L^2}^2 \geq \|\mathrm{P}_{>k} f_\star\|_{L^2}^2 + o_d(1).$$

  $\mathcal{F}_2(\sqrt{n}) \approx \{\text{degree-}k \text{ polynomials}\}$.

Curse of dimensionality

- $\mathcal{F}_2(R) = \Big\{ f(\boldsymbol{x}; a) : \|a\|_{L^2} \le R \Big\}$, $f_\star \in \mathcal{F}_2(R)$,

$$\text{Generalization error of learning } f_\star \text{ from } n \text{ samples } \le C\frac{R}{\sqrt{n}}.$$

- Kernel methods suffer from the **curse of dimensionality**.

- $\boldsymbol{x}_i \sim_{iid} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim \text{Unif}(\mathbb{S}^{d-1}(1))$, $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)$.

  Any target function $f_\star \in L^2(\mathcal{X})$, number of samples: $d^k \ll n \ll d^{k+1}$ [GMMM,'19]

$$\text{Test error with squared loss:} \quad \|f_\star - \hat{f}_n\|_{L^2}^2 \ge \|\mathrm{P}_{>k} f_\star\|_{L^2}^2 + o_d(1).$$

  $\mathcal{F}_2(\sqrt{n}) \approx \{\text{degree-}k \text{ polynomials}\}$.

- $\mathcal{F}_2(R) = \Big\{ f(\boldsymbol{x}; a) : \|a\|_{L^2} \le R \Big\}$, $f_\star \in \mathcal{F}_2(R)$,

$$\text{Generalization error of learning } f_\star \text{ from } n \text{ samples } \le C\frac{R}{\sqrt{n}}.$$

- Kernel methods suffer from the **curse of dimensionality**.

- $\boldsymbol{x}_i \sim_{iid} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim \text{Unif}(\mathbb{S}^{d-1}(1))$, $\phi(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)$.

  Any target function $f_\star \in L^2(\mathcal{X})$, number of samples: $d^k \ll n \ll d^{k+1}$ [GMMM,'19]

$$\text{Test error with squared loss:} \quad \|f_\star - \hat{f}_n\|_{L^2}^2 \ge \|\mathsf{P}_{>k} f_\star\|_{L^2}^2 + o_d(1).$$

  $\mathcal{F}_2(\sqrt{n}) \approx \{\text{degree-}k \text{ polynomials}\}$.

'Convex neural networks'

- $f_\star(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{w}_*)$. We have $a_\star(\boldsymbol{w}) = \delta_{\boldsymbol{w},\boldsymbol{w}_\star}$ and $\|a_\star\|_{L^2} = \infty$ (hence $f_\star \notin \mathcal{F}_2$).

  $f_\star$ is not learned efficiently by Kernel methods: $f_\star$ cannot be approximated by $f(\cdot; a)$ with $\|a\|_{L^2}$ bounded.

- 'Convex neural network' [Bengio et al.,'06]: for $1 \leq p < 2$,

$$\mathcal{F}_p(R) = \Big\{ f(\cdot; a) : \|a\|_{L^p} = \Big(\int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \Big)^{1/p} \leq R \Big\}.$$

- By Jensen's inequality $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$.

- $a(\boldsymbol{w})$ may tend to singular distribution with $\|a\|_{L^1}$ bounded (not true for L^2-norm).

  [Bach,'17] $\mathcal{F}_1$ is adaptive and beat the curse of dimensionality for functions that depends on a low-dimensional projection of the covariates.

  $\mathcal{F}_2$ is not adaptive.

'Convex neural networks'

- $f_\star(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{w}_*)$. We have $a_\star(\boldsymbol{w}) = \delta_{\boldsymbol{w},\boldsymbol{w}_\star}$ and $\|a_\star\|_{L^2} = \infty$ (hence $f_\star \notin \mathcal{F}_2$).

  $f_\star$ is not learned efficiently by Kernel methods: $f_\star$ cannot be approximated by $f(\cdot; a)$ with $\|a\|_{L^2}$ bounded.

- 'Convex neural network' [Bengio et al.,'06]: for $1 \le p < 2$,

$$\mathcal{F}_p(R) = \left\{ f(\cdot; a) : \|a\|_{L^p} = \left(\int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \right)^{1/p} \le R \right\}.$$

- By Jensen's inequality $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$.

- $a(\boldsymbol{w})$ may tend to singular distribution with $\|a\|_{L^1}$ bounded (not true for L^2-norm).

  [Bach,'17] $\mathcal{F}_1$ is adaptive and beat the curse of dimensionality for functions that depends on a low-dimensional projection of the covariates.

  $\mathcal{F}_2$ is not adaptive.

'Convex neural networks'

- $f_\star(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{w}_*)$. We have $a_\star(\boldsymbol{w}) = \delta_{\boldsymbol{w},\boldsymbol{w}_\star}$ and $\|a_\star\|_{L^2} = \infty$ (hence $f_\star \notin \mathcal{F}_2$).

  $f_\star$ is not learned efficiently by Kernel methods: $f_\star$ cannot be approximated by $f(\cdot; a)$ with $\|a\|_{L^2}$ bounded.

- 'Convex neural network' [Bengio et al.,'06]: for $1 \leq p < 2$,

$$\mathcal{F}_p(R) = \left\{ f(\cdot; a) : \|a\|_{L^p} = \left(\int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \right)^{1/p} \leq R \right\}.$$

- By Jensen's inequality $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$.

- $a(\boldsymbol{w})$ may tend to singular distribution with $\|a\|_{L^1}$ bounded (not true for L^2-norm).

  [Bach,'17] $\mathcal{F}_1$ is adaptive and beat the curse of dimensionality for functions that depends on a low-dimensional projection of the covariates.

  $\mathcal{F}_2$ is not adaptive.

'Convex neural networks'

- $f_\star(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{w}_*)$. We have $a_\star(\boldsymbol{w}) = \delta_{\boldsymbol{w},\boldsymbol{w}_\star}$ and $\|a_\star\|_{L^2} = \infty$ (hence $f_\star \notin \mathcal{F}_2$).

  $f_\star$ is not learned efficiently by Kernel methods: $f_\star$ cannot be approximated by $f(\cdot; a)$ with $\|a\|_{L^2}$ bounded.

- 'Convex neural network' [Bengio et al.,'06]: for $1 \le p < 2$,

$$\mathcal{F}_p(R) = \left\{ f(\cdot; a) : \|a\|_{L^p} = \left(\int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \right)^{1/p} \le R \right\}.$$

- By Jensen's inequality $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$.

- $a(\boldsymbol{w})$ may tend to singular distribution with $\|a\|_{L^1}$ bounded (not true for L^2-norm).

  [Bach,'17] $\mathcal{F}_1$ is adaptive and beat the curse of dimensionality for functions that depends on a low-dimensional projection of the covariates.

  $\mathcal{F}_2$ is not adaptive.

$\mathcal{F}_p$-interpolation solution

- **Minimum-norm interpolating solution:** ($\lambda \to 0^+$ in ERM)

$$\begin{aligned} \text{minimize} \quad & \int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \,, \\ \text{subj. to} \quad & f(\boldsymbol{x}_i; a) = y_i, \quad \forall i \leq n \,. \end{aligned}$$

- Correspond to modern practice of training until interpolation.
- Infinite dimensional convex problem. Not clear if it is tractable for $p \neq 2$.
- For $p = 1$:

  [Bengio et al.,'06] incremental algorithm but no global optimality guarantees.

  [Bach,'17] conditional gradient algorithm, but each step potentially hard.

$\mathcal{F}_p$-interpolation solution

- **Minimum-norm interpolating solution:** ($\lambda \to 0^+$ in ERM)

$$\begin{aligned} \text{minimize} \quad & \int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}), \\ \text{subj. to} \quad & f(\boldsymbol{x}_i; a) = y_i, \quad \forall i \le n. \end{aligned}$$

- Correspond to modern practice of training until interpolation.

- Infinite dimensional convex problem. Not clear if it is tractable for $p \neq 2$.

- For $p = 1$:

  [Bengio et al.,'06] incremental algorithm but no global optimality guarantees.

  [Bach,'17] conditional gradient algorithm, but each step potentially hard.

- **Minimum-norm interpolating solution:** ($\lambda \to 0^+$ in ERM)

$$\begin{aligned} \text{minimize} \quad & \int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \,, \\ \text{subj. to} \quad & f(\boldsymbol{x}_i; a) = y_i, \quad \forall i \le n \,. \end{aligned}$$

- Correspond to modern practice of training until interpolation.
- Infinite dimensional convex problem. Not clear if it is tractable for $p \neq 2$.
- For $p = 1$:

  [Bengio et al.,'06] incremental algorithm but no global optimality guarantees.

  [Bach,'17] conditional gradient algorithm, but each step potentially hard.

$\mathcal{F}_p$-interpolation solution

- **Minimum-norm interpolating solution:** ($\lambda \to 0^+$ in ERM)

$$\begin{aligned} \text{minimize} \quad & \int_\Omega |a(\boldsymbol{w})|^p \mu(\mathrm{d}\boldsymbol{w}) \,, \\ \text{subj. to} \quad & f(\boldsymbol{x}_i; a) = y_i, \quad \forall i \leq n \,. \end{aligned}$$

- Correspond to modern practice of training until interpolation.
- Infinite dimensional convex problem. Not clear if it is tractable for $p \neq 2$.
- For $p = 1$:

  [Bengio et al.,'06] incremental algorithm but no global optimality guarantees.

  [Bach,'17] conditional gradient algorithm, but each step potentially hard.

Random Features approximation

- Approximate μ by finitely supported $\hat{\mu}_M$: sample M weights $\boldsymbol{w}_j \sim_{iid} \mu$,

$$f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) \longrightarrow f_M(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{M}\sum_{j=1}^{M} a_j \phi(\boldsymbol{x}; \boldsymbol{w}_j).$$

- 'Finite-width' problem is easy to solve: $\boldsymbol{a} \in \mathbb{R}^M$,

$$\text{minimize} \quad \frac{1}{M}\sum_{j=1}^{M} |a_j|^p \,,$$
$$\text{subj. to} \quad f_M(\boldsymbol{x}_i; \boldsymbol{a}) = y_i, \quad \forall i \leq n \,.$$

How large M needs to be for the finite-width solution $\hat{f}_{\mathrm{RF},M,n}$ to approximate the infinite-width solution $\hat{f}_n$?

- For $p = 2$, we have $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} \approx 0$ when $M \geq n^{1+\delta}$ and $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} > 0$ when $M \leq n^{1-\delta}$ [MMM,'21].

Random Features approximation

- Approximate μ by finitely supported $\hat{\mu}_M$: sample M weights $\boldsymbol{w}_j \sim_{iid} \mu$,

$$f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) \longrightarrow f_M(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{M}\sum_{j=1}^{M} a_j \phi(\boldsymbol{x}; \boldsymbol{w}_j).$$

- 'Finite-width' problem is easy to solve: $\boldsymbol{a} \in \mathbb{R}^M$,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{M}\sum_{j=1}^{M} |a_j|^p \,, \\ \text{subj. to} \quad & f_M(\boldsymbol{x}_i; \boldsymbol{a}) = y_i, \quad \forall i \le n \,. \end{aligned}$$

How large M needs to be for the finite-width solution $\hat{f}_{\mathrm{RF},M,n}$ to approximate the infinite-width solution $\hat{f}_n$?

- For $p = 2$, we have $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} \approx 0$ when $M \ge n^{1+\delta}$ and $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} > 0$ when $M \le n^{1-\delta}$ [MMM,'21].

Random Features approximation

- Approximate μ by finitely supported $\hat{\mu}_M$: sample M weights $\boldsymbol{w}_j \sim_{iid} \mu$,

$$f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) \longrightarrow f_M(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{M}\sum_{j=1}^{M} a_j \phi(\boldsymbol{x}; \boldsymbol{w}_j).$$

- 'Finite-width' problem is easy to solve: $\boldsymbol{a} \in \mathbb{R}^M$,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{M}\sum_{j=1}^{M} |a_j|^p \,, \\ \text{subj. to} \quad & f_M(\boldsymbol{x}_i; \boldsymbol{a}) = y_i, \quad \forall i \leq n \,. \end{aligned}$$

How large M needs to be for the finite-width solution $\hat{f}_{\mathrm{RF},M,n}$ to approximate the infinite-width solution $\hat{f}_n$?

- For $p = 2$, we have $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} \approx 0$ when $M \geq n^{1+\delta}$ and $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} > 0$ when $M \leq n^{1-\delta}$ [MMM,'21].

Random Features approximation

- Approximate μ by finitely supported $\hat{\mu}_M$: sample M weights $\boldsymbol{w}_j \sim_{iid} \mu$,

$$f(\boldsymbol{x}; a) = \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) \longrightarrow f_M(\boldsymbol{x}; \boldsymbol{a}) = \frac{1}{M}\sum_{j=1}^{M} a_j \phi(\boldsymbol{x}; \boldsymbol{w}_j).$$

- 'Finite-width' problem is easy to solve: $\boldsymbol{a} \in \mathbb{R}^M$,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{M}\sum_{j=1}^{M} |a_j|^p \,, \\ \text{subj. to} \quad & f_M(\boldsymbol{x}_i; \boldsymbol{a}) = y_i, \quad \forall i \le n\,. \end{aligned}$$

How large M needs to be for the finite-width solution $\hat{f}_{\mathrm{RF},M,n}$ to approximate the infinite-width solution $\hat{f}_n$?

- For $p = 2$, we have $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} \approx 0$ when $M \ge n^{1+\delta}$ and $\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2} > 0$ when $M \le n^{1-\delta}$ [MMM,'21].

Setting

- **Data:** $\{y_i, \boldsymbol{x}_i\}_{i\leq n}$, bounded support: $\|\boldsymbol{x}\|_2 \leq C\sqrt{d}$.
- **Kernel matrix:** $K_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j\leq n}$, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int_\Omega \phi(\boldsymbol{x}_i; \boldsymbol{w})\phi(\boldsymbol{x}_j; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.
- **Random Features vectors:** $\{\boldsymbol{w}_j\}_{j\leq M}$ fixed iid from μ,

$$\boldsymbol{\phi}_{n,j} = \boldsymbol{\phi}_n(\boldsymbol{w}_j) := [\phi(\boldsymbol{x}_1; \boldsymbol{w}_j), \ldots, \phi(\boldsymbol{x}_n; \boldsymbol{w}_j)] \in \mathbb{R}^n.$$

- **Whitened features:** $\boldsymbol{\psi}_n(\boldsymbol{w}) := \boldsymbol{K}_n^{-1/2}\boldsymbol{\phi}_n(\boldsymbol{w})$ (so that $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{\psi}_n(\boldsymbol{w})\boldsymbol{\psi}_n(\boldsymbol{w})^{\mathsf{T}}] = \mathrm{I}_n$).

Result will hold conditionally on realization of the data $\{y_i, \boldsymbol{x}_i\}$, and exploit randomness of weights $\{\boldsymbol{w}_j\}$ to show concentration.

E.g., $\boldsymbol{\phi}_{n,j}$ conditional on $\boldsymbol{X}$: iid random vectors in $\mathbb{R}^n$.

Setting

- **Data:** $\{y_i, \boldsymbol{x}_i\}_{i \le n}$, bounded support: $\|\boldsymbol{x}\|_2 \le C\sqrt{d}$.

- **Kernel matrix:** $\boldsymbol{K}_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \le n}$, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int_\Omega \phi(\boldsymbol{x}_i; \boldsymbol{w})\phi(\boldsymbol{x}_j; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.

- **Random Features vectors:** $\{\boldsymbol{w}_j\}_{j \le M}$ fixed iid from μ,

$$\boldsymbol{\phi}_{n,j} = \boldsymbol{\phi}_n(\boldsymbol{w}_j) := [\phi(\boldsymbol{x}_1; \boldsymbol{w}_j), \ldots, \phi(\boldsymbol{x}_n; \boldsymbol{w}_j)] \in \mathbb{R}^n.$$

- **Whitened features:** $\boldsymbol{\psi}_n(\boldsymbol{w}) := \boldsymbol{K}_n^{-1/2}\boldsymbol{\phi}_n(\boldsymbol{w})$ (so that $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{\psi}_n(\boldsymbol{w})\boldsymbol{\psi}_n(\boldsymbol{w})^{\mathsf{T}}] = \mathrm{I}_n$).

Result will hold conditionally on realization of the data $\{y_i, \boldsymbol{x}_i\}$, and exploit randomness of weights $\{\boldsymbol{w}_j\}$ to show concentration.

E.g., $\boldsymbol{\phi}_{n,j}$ conditional on $\boldsymbol{X}$: iid random vectors in $\mathbb{R}^n$.

Setting

- **Data:** $\{y_i, \boldsymbol{x}_i\}_{i \le n}$, bounded support: $\|\boldsymbol{x}\|_2 \le C\sqrt{d}$.

- **Kernel matrix:** $\boldsymbol{K}_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \le n}$, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int_\Omega \phi(\boldsymbol{x}_i; \boldsymbol{w})\phi(\boldsymbol{x}_j; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.

- **Random Features vectors:** $\{\boldsymbol{w}_j\}_{j \le M}$ fixed iid from μ,

$$\boldsymbol{\phi}_{n,j} = \boldsymbol{\phi}_n(\boldsymbol{w}_j) := [\phi(\boldsymbol{x}_1; \boldsymbol{w}_j), \ldots, \phi(\boldsymbol{x}_n; \boldsymbol{w}_j)] \in \mathbb{R}^n.$$

- **Whitened features:** $\boldsymbol{\psi}_n(\boldsymbol{w}) := \boldsymbol{K}_n^{-1/2}\boldsymbol{\phi}_n(\boldsymbol{w})$ (so that $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{\psi}_n(\boldsymbol{w})\boldsymbol{\psi}_n(\boldsymbol{w})^{\mathsf{T}}] = \mathrm{I}_n$).

Result will hold conditionally on realization of the data $\{y_i, \boldsymbol{x}_i\}$, and exploit randomness of weights $\{\boldsymbol{w}_j\}$ to show concentration.

E.g., $\boldsymbol{\phi}_{n,j}$ conditional on $\boldsymbol{X}$: iid random vectors in $\mathbb{R}^n$.

Setting

- **Data:** $\{y_i, \boldsymbol{x}_i\}_{i \le n}$, bounded support: $\|\boldsymbol{x}\|_2 \le C\sqrt{d}$.
- **Kernel matrix:** $\boldsymbol{K}_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \le n}$, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int_\Omega \phi(\boldsymbol{x}_i; \boldsymbol{w}) \phi(\boldsymbol{x}_j; \boldsymbol{w}) \mu(\mathrm{d}\boldsymbol{w})$.
- **Random Features vectors:** $\{\boldsymbol{w}_j\}_{j \le M}$ fixed iid from μ,

$$\boldsymbol{\phi}_{n,j} = \boldsymbol{\phi}_n(\boldsymbol{w}_j) := [\phi(\boldsymbol{x}_1; \boldsymbol{w}_j), \ldots, \phi(\boldsymbol{x}_n; \boldsymbol{w}_j)] \in \mathbb{R}^n.$$

- **Whitened features:** $\boldsymbol{\psi}_n(\boldsymbol{w}) := \boldsymbol{K}_n^{-1/2} \boldsymbol{\phi}_n(\boldsymbol{w})$ (so that $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{\psi}_n(\boldsymbol{w}) \boldsymbol{\psi}_n(\boldsymbol{w})^\mathsf{T}] = \mathrm{I}_n$).

Result will hold conditionally on realization of the data $\{y_i, \boldsymbol{x}_i\}$, and exploit randomness of weights $\{\boldsymbol{w}_j\}$ to show concentration.

E.g., $\boldsymbol{\phi}_{n,j}$ conditional on $\boldsymbol{X}$: iid random vectors in $\mathbb{R}^n$.

Setting

- **Data:** $\{y_i, \boldsymbol{x}_i\}_{i \le n}$, bounded support: $\|\boldsymbol{x}\|_2 \le C\sqrt{d}$.
- **Kernel matrix:** $\boldsymbol{K}_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \le n}$, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \int_\Omega \phi(\boldsymbol{x}_i; \boldsymbol{w})\phi(\boldsymbol{x}_j; \boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w})$.
- **Random Features vectors:** $\{\boldsymbol{w}_j\}_{j \le M}$ fixed iid from μ,

$$\boldsymbol{\phi}_{n,j} = \boldsymbol{\phi}_n(\boldsymbol{w}_j) := [\phi(\boldsymbol{x}_1; \boldsymbol{w}_j), \dots, \phi(\boldsymbol{x}_n; \boldsymbol{w}_j)] \in \mathbb{R}^n.$$

- **Whitened features:** $\boldsymbol{\psi}_n(\boldsymbol{w}) := \boldsymbol{K}_n^{-1/2}\boldsymbol{\phi}_n(\boldsymbol{w})$ (so that $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{\psi}_n(\boldsymbol{w})\boldsymbol{\psi}_n(\boldsymbol{w})^\mathsf{T}] = \mathrm{I}_n$).

Result will hold conditionally on realization of the data $\{y_i, \boldsymbol{x}_i\}$, and exploit randomness of weights $\{\boldsymbol{w}_j\}$ to show concentration.

E.g., $\boldsymbol{\phi}_{n,j}$ conditional on $\boldsymbol{X}$: iid random vectors in $\mathbb{R}^n$.

Conditionally on $\boldsymbol{y}, \boldsymbol{X}$:

A1 **[Sub-gaussianity]** For any $\|\boldsymbol{x}\|_2 \leq C\sqrt{d}$, $\phi(\boldsymbol{x}; \boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.
Whitened feature vector $\psi_n(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A2 **[Lipschitz continuity]** Feature $\phi(\boldsymbol{x}; \boldsymbol{w})$ is $L(\boldsymbol{w})$-Lipschitz w.r.t $\boldsymbol{x}$ and $L(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A3 **[Small ball property]** There exist $\eta, c > 0$ such that

$$\inf_{\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1} \mathbb{P}(|\langle \boldsymbol{u}, \psi_n(\boldsymbol{w}) \rangle| \geq \eta, |\langle \boldsymbol{v}, \psi_n(\boldsymbol{w}) \rangle|\rangle| \geq \eta) \geq c.$$

Assumptions A1 and A3 can be hard to check in practice.

Assumptions

Conditionally on $\boldsymbol{y}, \boldsymbol{X}$:

A1 **[Sub-gaussianity]** For any $\|\boldsymbol{x}\|_2 \leq C\sqrt{d}$, $\phi(\boldsymbol{x}; \boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.
Whitened feature vector $\psi_n(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A2 **[Lipschitz continuity]** Feature $\phi(\boldsymbol{x}; \boldsymbol{w})$ is $L(\boldsymbol{w})$-Lipschitz w.r.t $\boldsymbol{x}$ and $L(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A3 **[Small ball property]** There exist $\eta, c > 0$ such that

$$\inf_{\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1} \mathbb{P}(|\langle \boldsymbol{u}, \psi_n(\boldsymbol{w})\rangle| \geq \eta, |\langle \boldsymbol{v}, \psi_n(\boldsymbol{w})\rangle|\rangle| \geq \eta) \geq c.$$

Assumptions A1 and A3 can be hard to check in practice.

Conditionally on $\boldsymbol{y}, \boldsymbol{X}$:

A1 **[Sub-gaussianity]** For any $\|\boldsymbol{x}\|_2 \le C\sqrt{d}$, $\phi(\boldsymbol{x}; \boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.
Whitened feature vector $\boldsymbol{\psi}_n(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A2 **[Lipschitz continuity]** Feature $\phi(\boldsymbol{x}; \boldsymbol{w})$ is $L(\boldsymbol{w})$-Lipschitz w.r.t $\boldsymbol{x}$ and $L(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A3 **[Small ball property]** There exist $\eta, c > 0$ such that

$$\inf_{\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1} \mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{\psi}_n(\boldsymbol{w})\rangle| \ge \eta, |\langle \boldsymbol{v}, \boldsymbol{\psi}_n(\boldsymbol{w})\rangle|\rangle| \ge \eta) \ge c.$$

Assumptions A1 and A3 can be hard to check in practice.

Conditionally on $\boldsymbol{y}, \boldsymbol{X}$:

A1 **[Sub-gaussianity]** For any $\|\boldsymbol{x}\|_2 \leq C\sqrt{d}$, $\phi(\boldsymbol{x}; \boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.
Whitened feature vector $\boldsymbol{\psi}_n(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A2 **[Lipschitz continuity]** Feature $\phi(\boldsymbol{x}; \boldsymbol{w})$ is $L(\boldsymbol{w})$-Lipschitz w.r.t $\boldsymbol{x}$ and $L(\boldsymbol{w})$ is τ^2-sub-Gaussian when $\boldsymbol{w} \sim \mu$.

A3 **[Small ball property]** There exist $\eta, c > 0$ such that

$$\inf_{\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1} \mathbb{P}(|\langle \boldsymbol{u}, \boldsymbol{\psi}_n(\boldsymbol{w})\rangle| \geq \eta, |\langle \boldsymbol{v}, \boldsymbol{\psi}_n(\boldsymbol{w})\rangle|\rangle| \geq \eta) \geq c.$$

Assumptions A1 and A3 can be hard to check in practice.

Main result

Theorem (Celentano, Misiakiewicz, Montanari, 2021)

For $p \in (1, 2]$. Assume A1, A2 and A3 hold conditionally on $\boldsymbol{y}, \boldsymbol{X}$. Then with probability at least $1 - CM^{-cn}$,

$$\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2}^2 \leq C \left(\frac{n \log M}{M} \vee \frac{(n \log M)^{p/(p-1)}}{M^2} \right) \|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2^2.$$

- Typically, $\|\boldsymbol{K}_n^{-1/2}\boldsymbol{y}\|_2 \leq C\sqrt{n}$, hence we need $M \geq (n \log(n))^{2 \vee \left(\frac{p - \frac{1}{2}}{p-1}\right)}$.
- Bound is not optimal: e.g., $p = 2$, we expect $M \geq n \log(n)$ to be sufficient.
- Bound diverges as $p \to 1$: we can't solve efficiently the infinite width problem $\mathcal{F}_1$ with the RF approach.

  E.g., learning a single neuron:

  [Bach,'17] can be learned with $\mathcal{F}_1$ with $n \leq d^{O(1)}$.
  [GMMM,'19] need $M \geq e^{\Theta(d)}$ features to be approximated.

Main result

Theorem (Celentano, Misiakiewicz, Montanari, 2021)

For $p \in (1, 2]$. Assume A1, A2 and A3 hold conditionally on $\boldsymbol{y}, \boldsymbol{X}$. Then with probability at least $1 - CM^{-cn}$,

$$\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2}^2 \leq C \left(\frac{n \log M}{M} \vee \frac{(n \log M)^{p/(p-1)}}{M^2} \right) \|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2^2.$$

- Typically, $\|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2 \leq C\sqrt{n}$, hence we need $M \geq (n \log(n))^{2 \vee \left(\frac{p - \frac{1}{2}}{p-1} \right)}$.
- Bound is not optimal: e.g., $p = 2$, we expect $M \geq n \log(n)$ to be sufficient.
- Bound diverges as $p \to 1$: we can't solve efficiently the infinite width problem $\mathcal{F}_1$ with the RF approach.

  E.g., learning a single neuron:

  [Bach,'17] can be learned with $\mathcal{F}_1$ with $n \leq d^{O(1)}$.

  [GMMM,'19] need $M \geq e^{\Theta(d)}$ features to be approximated.

Main result

Theorem (Celentano, Misiakiewicz, Montanari, 2021)

For $p \in (1, 2]$. Assume A1, A2 and A3 hold conditionally on $\boldsymbol{y}, \boldsymbol{X}$. Then with probability at least $1 - CM^{-cn}$,

$$\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2}^2 \le C \left(\frac{n \log M}{M} \vee \frac{(n \log M)^{p/(p-1)}}{M^2} \right) \|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2^2.$$

- Typically, $\|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2 \le C\sqrt{n}$, hence we need $M \ge (n \log(n))^{2 \vee \left(\frac{p - \frac{1}{2}}{p-1}\right)}$.
- Bound is not optimal: e.g., $p = 2$, we expect $M \ge n \log(n)$ to be sufficient.
- Bound diverges as $p \to 1$: we can't solve efficiently the infinite width problem $\mathcal{F}_1$ with the RF approach.

  E.g., learning a single neuron:

  [Bach,'17] can be learned with $\mathcal{F}_1$ with $n \le d^{O(1)}$.

  [GMMM,'19] need $M \ge e^{\Theta(d)}$ features to be approximated.

Main result

Theorem (Celentano, Misiakiewicz, Montanari, 2021)

For $p \in (1, 2]$. Assume A1, A2 and A3 hold conditionally on $\boldsymbol{y}, \boldsymbol{X}$. Then with probability at least $1 - CM^{-cn}$,

$$\|\hat{f}_{\mathrm{RF},M,n} - \hat{f}_n\|_{L^2}^2 \leq C \left(\frac{n \log M}{M} \vee \frac{(n \log M)^{p/(p-1)}}{M^2} \right) \|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2^2.$$

- Typically, $\|\boldsymbol{K}_n^{-1/2} \boldsymbol{y}\|_2 \leq C\sqrt{n}$, hence we need $M \geq (n \log(n))^{2 \vee \left(\frac{p - \frac{1}{2}}{p-1} \right)}$.
- Bound is not optimal: e.g., $p = 2$, we expect $M \geq n \log(n)$ to be sufficient.
- Bound diverges as $p \to 1$: we can't solve efficiently the infinite width problem $\mathcal{F}_1$ with the RF approach.

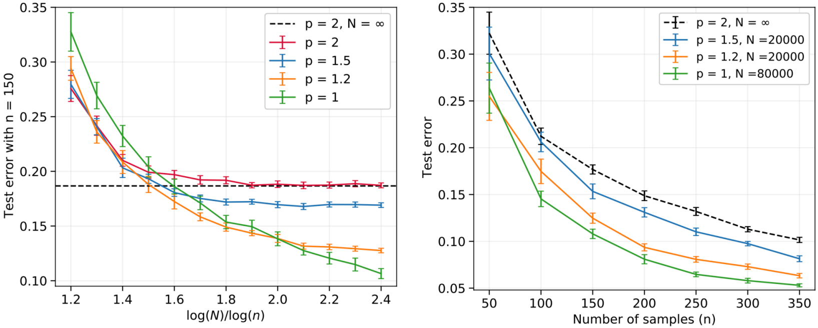
  E.g., learning a single neuron:

  [Bach,'17] can be learned with $\mathcal{F}_1$ with $n \leq d^{O(1)}$.

  [GMMM,'19] need $M \geq e^{\Theta(d)}$ features to be approximated.

Illustration

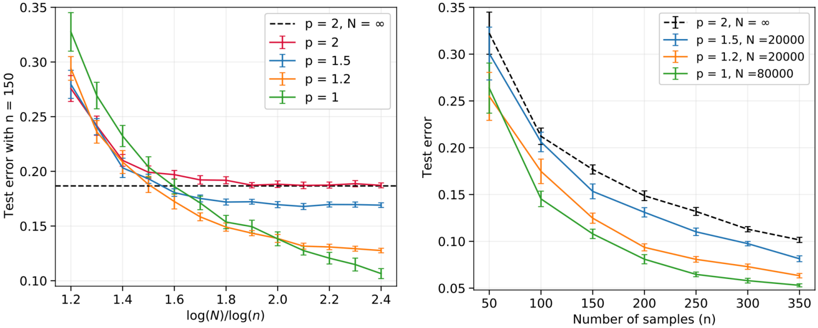
Learning single neuron: $\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_* \rangle)$, $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim N(0, \mathrm{I}_d/d)$, $d = 30$.



- **Left:** test error settles on 'infinite-width' solution error when $M \geq M_*(n,p)$.
- **Left:** $M_*(n,p)$ increases as p decreases ($p = 1$, unable to reach $M = \infty$ error).
- **Right:** test error decreases when p decreases. $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$ capture better and better functions highly dependent on low-dimensional projection of $\boldsymbol{x}$.

Illustration

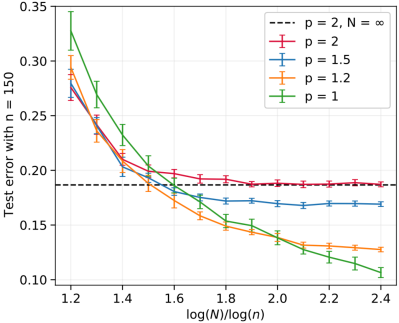
Learning single neuron: $\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_* \rangle)$, $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim N(0, \mathrm{I}_d/d)$, $d = 30$.

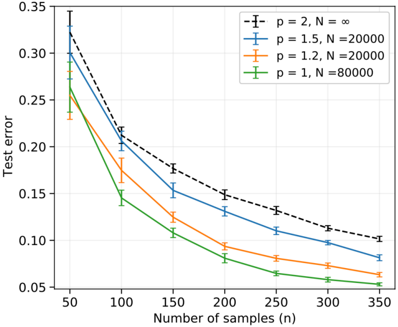


- **Left:** test error settles on 'infinite-width' solution error when $M \geq M_*(n, p)$.
- **Left:** $M_*(n, p)$ increases as p decreases ($p = 1$, unable to reach $M = \infty$ error).
- **Right:** test error decreases when p decreases. $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$ capture better and better functions highly dependent on low-dimensional projection of $\boldsymbol{x}$.

Illustration

Learning single neuron: $\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_* \rangle)$, $\boldsymbol{x} \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim N(0, \mathrm{I}_d/d)$, $d = 30$.

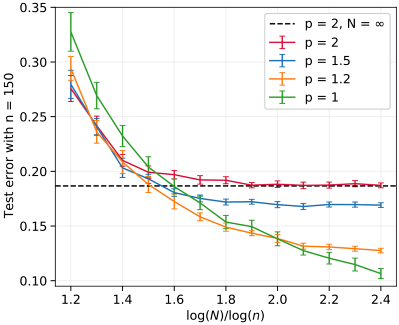


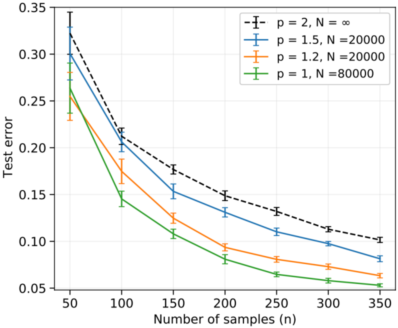


- **Left:** test error settles on 'infinite-width' solution error when $M \geq M_*(n, p)$.
- **Left:** $M_*(n, p)$ increases as p decreases ($p = 1$, unable to reach $M = \infty$ error).
- **Right:** test error decreases when p decreases. $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$ capture better and better functions highly dependent on low-dimensional projection of $\boldsymbol{x}$.

Illustration

Learning single neuron: $\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_* \rangle)$, $\boldsymbol{x} \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\boldsymbol{w} \sim N(0, \mathrm{I}_d/d)$, $d = 30$.





- **Left:** test error settles on 'infinite-width' solution error when $M \geq M_*(n, p)$.
- **Left:** $M_*(n, p)$ increases as p decreases ($p = 1$, unable to reach $M = \infty$ error).
- **Right:** test error decreases when p decreases. $\mathcal{F}_2(R) \subset \mathcal{F}_p(R) \subset \mathcal{F}_1(R)$ capture better and better functions highly dependent on low-dimensional projection of $\boldsymbol{x}$.

Idea of the proof (I): dual problem

Infinite-width: $$\min_{a:\Omega\to\mathbb{R}} \left\{ \int_\Omega \rho(a(\boldsymbol{w}))\mu(\mathrm{d}\boldsymbol{w}) : \forall i \le n, \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}_i;\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) = y_i \right\},$$

Finite-width: $$\min_{\boldsymbol{a}\in\mathbb{R}^M} \left\{ \frac{1}{M}\sum_{j\le M} \rho(a_j) : \forall i \le n, \frac{1}{M}\sum_{j\le M} a_j\phi(\boldsymbol{x}_i;\boldsymbol{w}_j) = y_i \right\}.$$

Dual problems: $\rho(x) = \frac{1}{p}|x|^p$, $\rho^* = \frac{1}{q}|x|^q$, $q = p/(p-1)$.

Infinite-width: $$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\mu(\mathrm{d}\boldsymbol{w}),$$

Finite-width: $$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \frac{1}{M}\sum_{j\le M} \rho^*(\langle \boldsymbol{\phi}_{n,j}, \boldsymbol{\lambda}\rangle).$$

- **Primal:** overparametrized, difficult to control.
  Dual: underparametrized, can use uniform convergence argument.

- 'Representer' theorem for $p \in (1,2]$: with $s(x) = (\rho^*)'(x) = \mathrm{sign}(x)|x|^{q-1}$, $\boldsymbol{\lambda} \in \mathbb{R}^n$

$$\hat{f}_n(\boldsymbol{x};\boldsymbol{\lambda}) = \int_\Omega s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}),$$

$$\hat{f}_{\mathrm{RF},M,n}(\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{M}\sum_{j\le M} s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}_j), \boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w}_j).$$

Idea of the proof (I): dual problem

Infinite-width: $$\min_{a:\Omega\to\mathbb{R}} \left\{ \int_\Omega \rho(a(\boldsymbol{w}))\mu(\mathrm{d}\boldsymbol{w}) : \forall i \le n, \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}_i;\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) = y_i \right\},$$

Finite-width: $$\min_{\boldsymbol{a}\in\mathbb{R}^M} \left\{ \frac{1}{M}\sum_{j\le M} \rho(a_j) : \forall i \le n, \frac{1}{M}\sum_{j\le M} a_j\phi(\boldsymbol{x}_i;\boldsymbol{w}_j) = y_i \right\}.$$

Dual problems: $\rho(x) = \frac{1}{p}|x|^p$, $\rho^* = \frac{1}{q}|x|^q$, $q = p/(p-1)$.

Infinite-width: $$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\mu(\mathrm{d}\boldsymbol{w}),$$

Finite-width: $$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \frac{1}{M}\sum_{j\le M} \rho^*(\langle \boldsymbol{\phi}_{n,j}, \boldsymbol{\lambda}\rangle).$$

- **Primal:** overparametrized, difficult to control.
  Dual: underparametrized, can use uniform convergence argument.

- 'Representer' theorem for $p \in (1,2]$: with $s(x) = (\rho^*)'(x) = \mathrm{sign}(x)|x|^{q-1}$, $\boldsymbol{\lambda} \in \mathbb{R}^n$

$$\hat{f}_n(\boldsymbol{x};\boldsymbol{\lambda}) = \int_\Omega s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}),$$

$$\hat{f}_{\mathrm{RF},M,n}(\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{M}\sum_{j\le M} s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}_j), \boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w}_j).$$

Idea of the proof (I): dual problem

Infinite-width: $$\min_{a:\Omega\to\mathbb{R}} \Big\{ \int_\Omega \rho(a(\boldsymbol{w}))\mu(\mathrm{d}\boldsymbol{w}) : \forall i \le n, \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}_i;\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) = y_i \Big\},$$

Finite-width: $$\min_{\boldsymbol{a}\in\mathbb{R}^M} \Big\{ \frac{1}{M}\sum_{j\le M} \rho(a_j) : \forall i \le n, \frac{1}{M}\sum_{j\le M} a_j\phi(\boldsymbol{x}_i;\boldsymbol{w}_j) = y_i \Big\}.$$

Dual problems: $\rho(x) = \frac{1}{p}|x|^p$, $\rho^* = \frac{1}{q}|x|^q$, $q = p/(p-1)$.

Infinite-width: $$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\mu(\mathrm{d}\boldsymbol{w}),$$

Finite-width: $$\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \frac{1}{M}\sum_{j\le M} \rho^*(\langle \boldsymbol{\phi}_{n,j}, \boldsymbol{\lambda}\rangle).$$

- **Primal:** overparametrized, difficult to control.
  Dual: underparametrized, can use uniform convergence argument.

- 'Representer' theorem for $p \in (1,2]$: with $s(x) = (\rho^*)'(x) = \mathrm{sign}(x)|x|^{q-1}$, $\boldsymbol{\lambda}\in\mathbb{R}^n$

$$\hat{f}_n(\boldsymbol{x};\boldsymbol{\lambda}) = \int_\Omega s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}),$$

$$\hat{f}_{\mathrm{RF},M,n}(\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{M}\sum_{j\le M} s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}_j), \boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w}_j).$$

Idea of the proof (I): dual problem

Infinite-width: $\min_{a:\Omega\to\mathbb{R}} \left\{ \int_\Omega \rho(a(\boldsymbol{w}))\mu(\mathrm{d}\boldsymbol{w}) : \forall i \le n, \int_\Omega a(\boldsymbol{w})\phi(\boldsymbol{x}_i;\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}) = y_i \right\}$,

Finite-width: $\min_{\boldsymbol{a}\in\mathbb{R}^M} \left\{ \frac{1}{M}\sum_{j\le M} \rho(a_j) : \forall i \le n, \frac{1}{M}\sum_{j\le M} a_j\phi(\boldsymbol{x}_i;\boldsymbol{w}_j) = y_i \right\}$.

Dual problems: $\rho(x) = \frac{1}{p}|x|^p$, $\rho^* = \frac{1}{q}|x|^q$, $q = p/(p-1)$.

Infinite-width: $\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y},\boldsymbol{\lambda}\rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}),\boldsymbol{\lambda}\rangle)\mu(\mathrm{d}\boldsymbol{w})$,

Finite-width: $\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y},\boldsymbol{\lambda}\rangle - \frac{1}{M}\sum_{j\le M} \rho^*(\langle \boldsymbol{\phi}_{n,j},\boldsymbol{\lambda}\rangle)$.

- **Primal:** overparametrized, difficult to control.
  Dual: underparametrized, can use uniform convergence argument.

- 'Representer' theorem for $p \in (1,2]$: with $s(x) = (\rho^*)'(x) = \mathrm{sign}(x)|x|^{q-1}$, $\boldsymbol{\lambda}\in\mathbb{R}^n$

$$\hat{f}_n(\boldsymbol{x};\boldsymbol{\lambda}) = \int_\Omega s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}),\boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w})\mu(\mathrm{d}\boldsymbol{w}),$$

$$\hat{f}_{\mathrm{RF},M,n}(\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{M}\sum_{j\le M} s(\langle \boldsymbol{\phi}_n(\boldsymbol{w}_j),\boldsymbol{\lambda}\rangle)\phi(\boldsymbol{x};\boldsymbol{w}_j).$$

Idea of the proof (II): uniform convergence

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\mu(\mathrm{d}\boldsymbol{w}),$$
$$\hat{\boldsymbol{\lambda}}_M = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \frac{1}{M}\sum_{j\le M} \rho^*(\langle \boldsymbol{\phi}_{n,j}, \boldsymbol{\lambda}\rangle).$$

- Concentration of the landscape uniformly over balls in $\boldsymbol{\lambda}$:

$$\nabla F_M(\boldsymbol{\lambda}) \to \nabla F(\boldsymbol{\lambda}), \qquad \lambda_{\max}(\nabla^2 F_M(\boldsymbol{\lambda})) \to \lambda_{\max}(\nabla^2 F(\boldsymbol{\lambda})),$$

which shows $\|\hat{\boldsymbol{\lambda}}_M - \hat{\boldsymbol{\lambda}}\|_2 \le \varepsilon_1(n, M, p)$.

- Uniform concentration of the predictor over balls in λ:

$$\max_{\boldsymbol{\lambda}\in B} \|\hat{f}_{\mathrm{RF},M,n}(\cdot;\boldsymbol{\lambda}) - \hat{f}_n(\cdot;\boldsymbol{\lambda})\|_{L^2} \le \varepsilon_2(n, M, p).$$

- Combining the two + Lipschitzness of $\hat{f}_n$ w.r.t λ:

$$\begin{aligned}\|\hat{f}_{\mathrm{RF},M,n}(\cdot;\hat{\boldsymbol{\lambda}}_M) - \hat{f}_n(\cdot;\hat{\boldsymbol{\lambda}})\|_{L^2} &\le \|\hat{f}_{\mathrm{RF},M,n}(\cdot;\hat{\boldsymbol{\lambda}}_M) - \hat{f}_n(\cdot;\hat{\boldsymbol{\lambda}}_M)\|_{L^2} + \|\hat{f}_n(\cdot;\hat{\boldsymbol{\lambda}}_M) - \hat{f}_n(\cdot;\hat{\boldsymbol{\lambda}})\|_{L^2}\\ &\le \varepsilon_2(n, M, p) + \varepsilon_1(n, M, p).\end{aligned}$$

Idea of the proof (II): uniform convergence

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda}\rangle)\mu(\mathrm{d}\boldsymbol{w}),$$

$$\hat{\boldsymbol{\lambda}}_M = \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda}\rangle - \frac{1}{M}\sum_{j\le M} \rho^*(\langle \boldsymbol{\phi}_{n,j}, \boldsymbol{\lambda}\rangle).$$

- Concentration of the landscape uniformly over balls in $\boldsymbol{\lambda}$:

$$\nabla F_M(\boldsymbol{\lambda}) \to \nabla F(\boldsymbol{\lambda}), \qquad \lambda_{\max}(\nabla^2 F_M(\boldsymbol{\lambda})) \to \lambda_{\max}(\nabla^2 F(\boldsymbol{\lambda})),$$

which shows $\|\hat{\boldsymbol{\lambda}}_M - \hat{\boldsymbol{\lambda}}\|_2 \le \varepsilon_1(n, M, p)$.

- Uniform concentration of the predictor over balls in $\boldsymbol{\lambda}$:

$$\max_{\boldsymbol{\lambda}\in B} \|\hat{f}_{\mathrm{RF},M,n}(\cdot;\boldsymbol{\lambda}) - \hat{f}_n(\cdot;\boldsymbol{\lambda})\|_{L^2} \le \varepsilon_2(n, M, p).$$

- Combining the two + Lipschitzness of $\hat{f}_n$ w.r.t λ:

$$\begin{aligned}\|\hat{f}_{\mathrm{RF},M,n}(\cdot;\hat{\lambda}_M) - \hat{f}_n(\cdot;\hat{\lambda})\|_{L^2} &\le \|\hat{f}_{\mathrm{RF},M,n}(\cdot;\hat{\lambda}_M) - \hat{f}_n(\cdot;\hat{\lambda}_M)\|_{L^2} + \|\hat{f}_n(\cdot;\hat{\lambda}_M) - \hat{f}_n(\cdot;\hat{\lambda})\|_{L^2} \\ &\le \varepsilon_2(n, M, p) + \varepsilon_1(n, M, p).\end{aligned}$$

Idea of the proof (II): uniform convergence

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} F(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda} \rangle - \int_\Omega \rho^*(\langle \boldsymbol{\phi}_n(\boldsymbol{w}), \boldsymbol{\lambda} \rangle) \mu(\mathrm{d}\boldsymbol{w}),$$

$$\hat{\boldsymbol{\lambda}}_M = \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} F_M(\boldsymbol{\lambda}) := \langle \boldsymbol{y}, \boldsymbol{\lambda} \rangle - \frac{1}{M} \sum_{j \leq M} \rho^*(\langle \boldsymbol{\phi}_{n,j}, \boldsymbol{\lambda} \rangle).$$

- Concentration of the landscape uniformly over balls in $\boldsymbol{\lambda}$:

$$\nabla F_M(\boldsymbol{\lambda}) \to \nabla F(\boldsymbol{\lambda}), \qquad \lambda_{\max}(\nabla^2 F_M(\boldsymbol{\lambda})) \to \lambda_{\max}(\nabla^2 F(\boldsymbol{\lambda})),$$

which shows $\|\hat{\boldsymbol{\lambda}}_M - \hat{\boldsymbol{\lambda}}\|_2 \leq \varepsilon_1(n, M, p)$.

- Uniform concentration of the predictor over balls in $\boldsymbol{\lambda}$:

$$\max_{\boldsymbol{\lambda} \in B} \|\hat{f}_{\mathrm{RF},M,n}(\cdot; \boldsymbol{\lambda}) - \hat{f}_n(\cdot; \boldsymbol{\lambda})\|_{L^2} \leq \varepsilon_2(n, M, p).$$

- Combining the two + Lipschitzness of $\hat{f}_n$ w.r.t λ:

$$\begin{aligned}\|\hat{f}_{\mathrm{RF},M,n}(\cdot; \hat{\boldsymbol{\lambda}}_M) - \hat{f}_n(\cdot; \hat{\boldsymbol{\lambda}})\|_{L^2} &\leq \|\hat{f}_{\mathrm{RF},M,n}(\cdot; \hat{\boldsymbol{\lambda}}_M) - \hat{f}_n(\cdot; \hat{\boldsymbol{\lambda}}_M)\|_{L^2} + \|\hat{f}_n(\cdot; \hat{\boldsymbol{\lambda}}_M) - \hat{f}_n(\cdot; \hat{\boldsymbol{\lambda}})\|_{L^2} \\ &\leq \varepsilon_2(n, M, p) + \varepsilon_1(n, M, p).\end{aligned}$$

Conclusion

- $\mathcal{F}_p$ $(1 < p < 2)$: function spaces of 2-layers NNs that are tractable and not RKHS.
- Studying effect of regularization in RF models: reduces to studying the corresponding infinite width model when sufficiently overparametrized.
- New proof of double descent phenomenon, that does not rely on strong assumptions. Simple mechanism: uniform concentration of the dual to an infinite-width problem.

Future directions:

- Generalization properties of $\mathcal{F}_p$.
- Non-zero regularization (non-interpolating solution).
- The most interesting case $p = 1$. Is there an efficient algorithm?

Thank you!

Conclusion

- $\mathcal{F}_p$ $(1 < p < 2)$: function spaces of 2-layers NNs that are tractable and not RKHS.
- Studying effect of regularization in RF models: reduces to studying the corresponding infinite width model when sufficiently overparametrized.
- New proof of double descent phenomenon, that does not rely on strong assumptions. Simple mechanism: uniform concentration of the dual to an infinite-width problem.

Future directions:

- Generalization properties of $\mathcal{F}_p$.
- Non-zero regularization (non-interpolating solution).
- The most interesting case $p = 1$. Is there an efficient algorithm?

Thank you!

Conclusion

- $\mathcal{F}_p$ $(1 < p < 2)$: function spaces of 2-layers NNs that are tractable and not RKHS.
- Studying effect of regularization in RF models: reduces to studying the corresponding infinite width model when sufficiently overparametrized.
- New proof of double descent phenomenon, that does not rely on strong assumptions. Simple mechanism: uniform concentration of the dual to an infinite-width problem.

Future directions:

- Generalization properties of $\mathcal{F}_p$.
- Non-zero regularization (non-interpolating solution).
- The most interesting case $p = 1$. Is there an efficient algorithm?

Thank you!

Conclusion

- $\mathcal{F}_p$ $(1 < p < 2)$: function spaces of 2-layers NNs that are tractable and not RKHS.

- Studying effect of regularization in RF models: reduces to studying the corresponding infinite width model when sufficiently overparametrized.

- New proof of double descent phenomenon, that does not rely on strong assumptions. Simple mechanism: uniform concentration of the dual to an infinite-width problem.

Future directions:

- Generalization properties of $\mathcal{F}_p$.
- Non-zero regularization (non-interpolating solution).
- The most interesting case $p = 1$. Is there an efficient algorithm?

Thank you!

Conclusion

- $\mathcal{F}_p$ $(1 < p < 2)$: function spaces of 2-layers NNs that are tractable and not RKHS.
- Studying effect of regularization in RF models: reduces to studying the corresponding infinite width model when sufficiently overparametrized.
- New proof of double descent phenomenon, that does not rely on strong assumptions. Simple mechanism: uniform concentration of the dual to an infinite-width problem.

Future directions:

- Generalization properties of $\mathcal{F}_p$.
- Non-zero regularization (non-interpolating solution).
- The most interesting case $p = 1$. Is there an efficient algorithm?

Thank you!