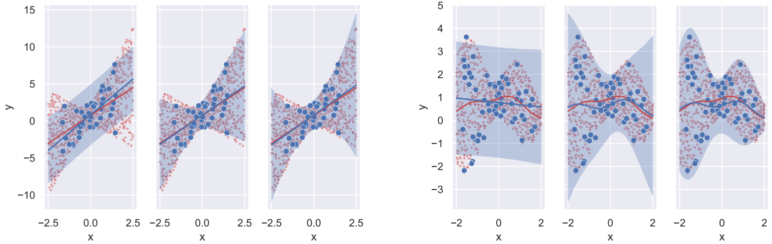
Universal Prediction Band via Semi-Definite Programming

Tengyuan Liang

The University of Chicago Booth School of Business

Be confident about (black-box) machine learning models, rigorously!



OUTLINE

- Motivation: uncertainty quantification dilemma in machine learning
- Semi-definite Programs (SDP): our approach
  - a numerical example
  - minimal implementation
- Rationale for Our SDP
  - sum-of-squares optimization
  - variance interpolation with confidence
  - connections to the literature
- Non-asymptotic Coverage Theory
  - assumptions and why universal
  - some insights: strong coverage, adaptivity
- Real Data Example: Fama-French

DILEMMA

A frequent criticism from the statistics community to modern machine learning (ML) is the lack of rigorous uncertainty quantification.

ML community would argue that conventional uncertainty quantification based on idealized distributional assumptions or asymptotics are too restrictive.

machine learning $\not\leftrightarrow$ statistical inference

DILEMMA

A frequent criticism from the statistics community to modern machine learning (ML) is the lack of rigorous uncertainty quantification.

ML community would argue that conventional uncertainty quantification based on idealized distributional assumptions or asymptotics are too restrictive.

machine learning $\nleftrightarrow$ statistical inference

A **dilemma**: uncertainty quantification for ML models that is

- rigorous with provable finite-sample properties
- universally applicable with little distributional assumptions

DILEMMA

Why important:

- available prediction intervals in scientific computing packages are merely **heuristics for visualization**
- **reliable decision making** based on complex ML models, such as deep neural networks and boosting machines

dilemma: general/universal $\nleftrightarrow$ rigorous/provable

DILEMMA

Why important:

- available prediction intervals in scientific computing packages are merely **heuristics for visualization**
- **reliable decision making** based on complex ML models, such as deep neural networks and boosting machines

dilemma: general/universal $\nleftrightarrow$ rigorous/provable

Some known approaches:

- conformal prediction
- (local) resampling method
- quantile regression

OUR CONTRIBUTION

We address the uncertainty quantification dilemma
via semi-definite programming (SDP).

OUR CONTRIBUTION

We address the uncertainty quantification dilemma
via semi-definite programming (SDP).

Our proposed method learns a data-adaptive, heteroskedastic prediction band

- **universally applicable** with mild distributional assumptions
- **strong non-asymptotic coverage** with/without user-specified predictive model
- **easy to implement** via standard convex optimization

OUR CONTRIBUTION

We address the uncertainty quantification dilemma
via semi-definite programming (SDP).

Our proposed method learns a data-adaptive, heteroskedastic prediction band

- **universally applicable** with mild distributional assumptions
- **strong non-asymptotic coverage** with/without user-specified predictive model
- **easy to implement** via standard convex optimization

machine learning $\overset{\text{SDP}}{\leftrightarrow}$ statistical inference

FORMULATION

Data: $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathbb{R}$ be the (covariates, response) drawn from an unknown dist. $\mathcal{P}$. $(x_i, y_i), i = 1, \ldots, n$ are n-i.i.d. samples.

Goal: given a regression or predictive ML model $\mathsf{m}_0(x)$, construct a prediction band $\widehat{\mathsf{PI}}(\mathbf{x})$ that covers $\mathbf{y}$.

FORMULATION

Data: $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathbb{R}$ be the (covariates, response) drawn from an unknown dist. $\mathcal{P}$. $(x_i, y_i), i = 1, \ldots, n$ are n-i.i.d. samples.

Goal: given a regression or predictive ML model $\mathsf{m}_0(x)$, construct a prediction band $\widehat{\mathsf{PI}}(\mathbf{x})$ that covers $\mathbf{y}$.

Kernel: $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous symmetric and positive-definite kernel function. Empirical kernel matrix $\mathbf{K} \in \mathbb{S}^{n \times n}$ with $\mathbf{K}_{ij} = K(x_i, x_j)$, with $\mathsf{K}_i \in \mathbb{R}^n$ denoting the i-th column.

SDP and prediction band

$$
\begin{aligned}
\widehat{\mathbf{B}} = \operatorname*{arg\,min}_{\mathbf{B}} \quad & \mathrm{Tr}(\mathbf{K}\mathbf{B}) \\
\text{s.t.} \quad & \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq (y_i - \mathsf{m}_0(x_i))^2, \; i = 1, \ldots, n \\
& \mathbf{B} \succeq 0
\end{aligned}
$$

optimization variable $\mathbf{B} \in \mathbb{S}^{n \times n}$ is a symmetric positive semi-definite (PSD) matrix.

SDP AND PREDICTION BAND

$$
\begin{aligned}
\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}} \quad & \mathrm{Tr}(\mathbf{K}\mathbf{B}) \\
\text{s.t.} \quad & \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq (y_i - \mathsf{m}_0(x_i))^2, \; i = 1, \ldots, n \\
& \mathbf{B} \succeq 0
\end{aligned}
$$

optimization variable $\mathbf{B} \in \mathbb{S}^{n \times n}$ is a symmetric positive semi-definite (PSD) matrix.

Prediction band

$$
\widehat{\mathsf{PI}}(x) := \left[\, \mathsf{m}_0(x) - \sqrt{\widehat{\mathsf{v}}(x)} \,,\; \mathsf{m}_0(x) + \sqrt{\widehat{\mathsf{v}}(x)} \,\right], \; \forall x \in \mathcal{X} \,,
$$

$$
\text{where} \quad \widehat{\mathsf{v}}(x) := \langle \mathsf{K}_x, \widehat{\mathbf{B}}\mathsf{K}_x \rangle \,,
$$

$$
\text{and} \quad \mathsf{K}_x := [K(x, x_1), \ldots, K(x, x_n)]^\top \in \mathbb{R}^n \,.
$$

SDP AND PREDICTION BAND

$\widehat{\mathsf{v}}(x)$ estimates the variability in the "deviations" $e_i := y_i - \mathsf{m}_0(x_i)$, computed based on any user-specified predictive model $\mathsf{m}_0(x)$

- absence of such a predictive model: set $\mathsf{m}_0(x) \equiv 0$,
  learn a conditional second-moment function to assess uncertainty.
- simultaneously learn the conditional mean and variance functions,
  using a variant of the aforementioned SDP.

pre-specified ML model $\mathsf{m}_0(x)$ is not required

WITHOUT USER-SPECIFIED PREDICTIVE MODEL

$$
\begin{aligned}
\min_{\alpha,\mathbf{B}} \quad & \gamma \cdot \left\langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \right\rangle + \mathrm{Tr}\left(\mathbf{K}^{\mathsf{v}} \mathbf{B}\right) \\
\text{s.t.} \quad & \left\langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \right\rangle \geq \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha \right\rangle\right)^2, \; i = 1, \ldots, n \\
& \mathbf{B} \succeq 0
\end{aligned}
$$

WITHOUT USER-SPECIFIED PREDICTIVE MODEL

$$
\begin{aligned}
\min_{\alpha, \mathbf{B}} \quad & \gamma \cdot \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle + \mathrm{Tr}\big(\mathbf{K}^{\mathsf{v}} \mathbf{B}\big) \\
\text{s.t.} \quad & \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \geq \big(y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle\big)^2, \; i = 1, \ldots, n \\
& \mathbf{B} \succeq 0
\end{aligned}
$$

Given the solution $\widehat{\mathbf{B}}$ and $\widehat{\alpha}$, the $\widehat{\mathsf{PI}}(x)$ is constructed as

$$
\widehat{\mathsf{PI}}(x) := \Big[\, \widehat{\mathsf{m}}(x) - \sqrt{\widehat{\mathsf{v}}(x)}\,,\; \widehat{\mathsf{m}}(x) + \sqrt{\widehat{\mathsf{v}}(x)}\, \Big], \; \forall x \in \mathcal{X}\,,
$$

where $\widehat{\mathsf{m}}(x) := \langle \mathsf{K}_x^{\mathsf{m}}, \widehat{\alpha} \rangle$ and $\widehat{\mathsf{v}}(x) := \langle \mathsf{K}_x^{\mathsf{v}}, \widehat{\mathbf{B}} \mathsf{K}_x^{\mathsf{v}} \rangle$.

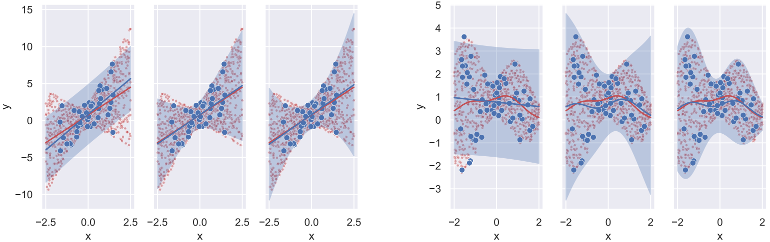
TEN-LINE IMPLEMENTATION

```
import cvxpy as cp

def sdpDual(K1, K2, Y, n, gamma = 1e1):
# K1 kernel for conditional mean, 1st moment
# K2 kernel for conditional variance, 2nd moment
# Define and solve the CVXPY problem.
    # Create a symmetric matrix variable \hat{B}
    hB = cp.Variable((n,n), symmetric=True)
    # Create a vector variable \hat{a}
    ha = cp.Variable(n)

  # PSD and inequality constraints
    constraints = [hB >> 0]
    constraints += [
        K2[i,:]@hB@K2[i,:] >=
        cp.square(Y[i] - K1[i,:]@ha) for i in range(n)
    ]
    prob = cp.Problem(cp.Minimize(
        gamma*cp.quad_form(ha, K1) + cp.trace(K2@hB)
    ), constraints)

    # Solve the SDP
    prob.solve()
    print("Optimal_Value", prob.value)

    return [ha.value, hB.value]
```

Listing 1: Minimal python code

A numerical example

A NUMERICAL EXAMPLE



A NUMERICAL EXAMPLE

Table 1: Simulated examples

	Coverage	Median Len	Average Len	MSE
Example 1: linear $\mathsf{m}(x)$, quadratic $\mathsf{v}(x)$				
SLR	85.88%	8.2057	8.2658	0.6294
SDP1	91.13%	7.4689	**7.7173**	**0.1146**
SDP2	**94.00%**	**7.2962**	8.3361	0.1720
Example 2: rbf $\mathsf{m}(x)$, rbf $\mathsf{v}(x)$				
SLR	96.13%	4.8048	4.8185	0.2556
SDP1	99.25%	4.4138	4.6196	0.1916
SDP2	**99.50%**	**3.3488**	**3.7506**	**0.1670**

Rationale behind the SDP

(1) sum-of-squares (SoS) optimization
(2) variance interpolation with confidence

REPRESENTATION THEOREM

finite-dim optimization:

$$\min_{\substack{\alpha\in\mathbb{R}^n \\ \mathbf{B}\in\mathbb{S}^{n\times n}}} \quad \gamma\cdot\left\langle \alpha, \mathbf{K}^{\mathsf{m}}\alpha\right\rangle + \mathrm{Tr}\left(\mathbf{K}^{\mathsf{v}}\mathbf{B}\right)$$

$$\text{s.t.} \quad \left\langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B}\mathsf{K}_i^{\mathsf{v}}\right\rangle \geq \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha\right\rangle\right)^2$$

$$\mathbf{B} \succeq 0$$

optimization of vector, matrix: α, $\mathbf{B}$

REPRESENTATION THEOREM

finite-dim optimization:

$$\min_{\substack{\alpha \in \mathbb{R}^n \\ \mathbf{B} \in \mathbb{S}^{n\times n}}} \quad \gamma \cdot \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle + \mathrm{Tr}(\mathbf{K}^{\mathsf{v}} \mathbf{B})$$

$$\text{s.t.} \quad \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \geq (y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle)^2$$

$$\mathbf{B} \succeq 0$$

optimization of vector, matrix: α, $\mathbf{B}$

infinite-dim optimization:

$$\min_{\substack{\beta \in \mathcal{H}^{\mathsf{m}} \\ \mathbf{A}: \mathcal{H}^{\mathsf{v}} \to \mathcal{H}^{\mathsf{v}}}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^{\mathsf{m}}} + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \langle \phi^{\mathsf{v}}_{x_i}, \mathbf{A} \phi^{\mathsf{v}}_{x_i} \rangle_{\mathcal{H}^{\mathsf{v}}} \geq (y_i - \langle \phi^{\mathsf{m}}_{x_i}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}})^2$$

$$\mathbf{A} \succeq 0$$

$\mathcal{H}^{\mathsf{m}}, \mathcal{H}^{\mathsf{v}}$ are the RKHSs $\mathsf{m}(x), \mathsf{v}(x)$ reside

REPRESENTATION THEOREM

finite-dim optimization:

$$\min_{\substack{\alpha \in \mathbb{R}^n \\ \mathbf{B} \in \mathbb{S}^{n\times n}}} \quad \gamma \cdot \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle + \mathrm{Tr}(\mathbf{K}^{\mathsf{v}} \mathbf{B})$$

$$\text{s.t.} \quad \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \geq (y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle)^2$$

$$\mathbf{B} \succeq 0$$

optimization of vector, matrix: α, $\mathbf{B}$

infinite-dim optimization:

$$\min_{\substack{\beta \in \mathcal{H}^{\mathsf{m}} \\ \mathbf{A}: \mathcal{H}^{\mathsf{v}} \to \mathcal{H}^{\mathsf{v}}}} \quad \gamma \cdot \|\beta\|_{\mathcal{H}^{\mathsf{m}}}^2 + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \langle \phi_{x_i}^{\mathsf{v}}, \mathbf{A} \phi_{x_i}^{\mathsf{v}} \rangle_{\mathcal{H}^{\mathsf{v}}} \geq (y_i - \langle \phi_{x_i}^{\mathsf{m}}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}})^2$$

$$\mathbf{A} \succeq 0$$

$\mathcal{H}^{\mathsf{m}}, \mathcal{H}^{\mathsf{v}}$ are the RKHSs $\mathsf{m}(x), \mathsf{v}(x)$ reside

Theorem (L.'21, representation)**.**

Above two optimizations are equivalent.

(1) sum-of-squares (SoS) optimization

SUM-OF-SQUARES OPTIMIZATION

Attempt 1:

infinite-dim optimization:

$$\min_{\substack{\beta \in \mathcal{H}^m \\ \mathbf{A}: \mathcal{H}^V \to \mathcal{H}^V}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^m} + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \boxed{? = \overbrace{\left(y_i - \underbrace{\langle \phi^m_{x_i}, \beta \rangle_{\mathcal{H}^m}}_{m(x)}\right)^2}^{v(x)}}$$

$$\mathbf{A} \succeq 0$$

What do we know about $v(x)$? Non-negative function! Yet, optimization over non-negative functions are NP-hard.

SUM-OF-SQUARES OPTIMIZATION

Attempt 1:

infinite-dim optimization:

$$\min_{\substack{\beta \in \mathcal{H}^m \\ \mathbf{A}: \mathcal{H}^v \to \mathcal{H}^v}} \quad \gamma \cdot \|\beta\|_{\mathcal{H}^m}^2 + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \boxed{? = \overbrace{\Big(y_i - \underbrace{\langle \phi_{x_i}^m, \beta \rangle_{\mathcal{H}^m}}_{m(x)}\Big)^2}^{v(x)}}$$

$$\mathbf{A} \succeq 0$$

What do we know about $v(x)$? Non-negative function! Yet, optimization over non-negative functions are NP-hard.

$$\text{sum-of-squares function} \overset{\text{relaxation}}{\Leftarrow} \text{non-negative function}$$

Lasserre (2001)

$$0 \le \langle \phi_x^v, \mathbf{A}\phi_x^v \rangle_{\mathcal{H}^v} = \overbrace{\big(y - m(x)\big)^2}^{v(x)}, \text{ for some } \mathbf{A} \succeq 0 .$$

when K^v is universal, the above sum-of-squares function can approximate all smooth, positive functions

Fefferman and Phong (1978); Bagnell and Farahmand (2015); Marteau-Ferey et al. (2020)

SUM-OF-SQUARES OPTIMIZATION

Attempt 1:

$$\min_{\beta, \mathbf{A}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^{\mathsf{m}}} + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \boxed{\langle \phi^{\mathsf{v}}_x, \mathbf{A}\phi^{\mathsf{v}}_x \rangle_{\mathcal{H}^{\mathsf{v}}} = \overbrace{\big(y_i - \underbrace{\langle \phi^{\mathsf{m}}_{x_i}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}}}_{\mathsf{m}(x)}\big)^2}^{\mathsf{v}(x)}}$$

$$\mathbf{A} \succeq 0$$

Problem: non-convex in $\mathbf{A}$, β!

SUM-OF-SQUARES OPTIMIZATION

Attempt 1:

$$\min_{\beta, \mathbf{A}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^{\mathrm{m}}} + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \boxed{\langle \phi^{\mathrm{v}}_x, \mathbf{A}\phi^{\mathrm{v}}_x \rangle_{\mathcal{H}^{\mathrm{v}}} = \overbrace{\big(y_i - \underbrace{\langle \phi^{\mathrm{m}}_{x_i}, \beta \rangle_{\mathcal{H}^{\mathrm{m}}}}_{\mathrm{m}(x)}\big)^2}^{\mathrm{v}(x)}}$$

$$\mathbf{A} \succeq 0$$

Problem: non-convex in $\mathbf{A}$, β!

Attempt 2:

$$\min_{\beta, \mathbf{A}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^{\mathrm{m}}} + \boxed{\|\mathbf{A}\|_\star}$$

$$\text{s.t.} \quad \boxed{\langle \phi^{\mathrm{v}}_x, \mathbf{A}\phi^{\mathrm{v}}_x \rangle_{\mathcal{H}^{\mathrm{v}}} \geq \overbrace{\big(y_i - \underbrace{\langle \phi^{\mathrm{m}}_{x_i}, \beta \rangle_{\mathcal{H}^{\mathrm{m}}}}_{\mathrm{m}(x)}\big)^2}^{\mathrm{v}(x)}}$$

$$\mathbf{A} \succeq 0$$

Solution: among the SoS functions that shelter the variance, find the minimum complexity one.

Now a convex program in $\mathbf{A}$, β!

SUM-OF-SQUARES OPTIMIZATION

Attempt 1:

$$\min_{\beta, \mathbf{A}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^{\mathsf{m}}} + \|\mathbf{A}\|_\star$$

$$\text{s.t.} \quad \boxed{\langle \phi^{\mathsf{v}}_x, \mathbf{A}\phi^{\mathsf{v}}_x \rangle_{\mathcal{H}^{\mathsf{v}}} = \overbrace{\Big(y_i - \underbrace{\langle \phi^{\mathsf{m}}_{x_i}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}}}_{\mathsf{m}(x)}\Big)^2}^{\mathsf{v}(x)}}$$

$$\mathbf{A} \succeq 0$$

Problem: non-convex in $\mathbf{A}$, β!

Attempt 2:

$$\min_{\beta, \mathbf{A}} \quad \gamma \cdot \|\beta\|^2_{\mathcal{H}^{\mathsf{m}}} + \boxed{\|\mathbf{A}\|_\star}$$

$$\text{s.t.} \quad \boxed{\langle \phi^{\mathsf{v}}_x, \mathbf{A}\phi^{\mathsf{v}}_x \rangle_{\mathcal{H}^{\mathsf{v}}} \geq \overbrace{\Big(y_i - \underbrace{\langle \phi^{\mathsf{m}}_{x_i}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}}}_{\mathsf{m}(x)}\Big)^2}^{\mathsf{v}(x)}}$$

$$\mathbf{A} \succeq 0$$

Solution: among the SoS functions that shelter the variance, find the minimum complexity one.

Now a convex program in $\mathbf{A}$, β!

minimum nuclear-norm $\Rightarrow$ small rank $\Rightarrow$ few factors realizing the conditional variance function

a particular form of **minimal prediction bandwidth**!

(2) variance interpolation with confidence

VARIANCE INTERPOLATION W. CONFIDENCE

finite-dim optimization:

$$\min_{\substack{\alpha\in\mathbb{R}^n \\ \mathbf{B}\in\mathbb{S}^{n\times n}}} \quad \gamma \cdot \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle + \mathrm{Tr}(\mathbf{K}^{\mathsf{v}}\mathbf{B})$$

$$\text{s.t.} \quad \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B}\mathsf{K}_i^{\mathsf{v}} \rangle \geq \left(y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle\right)^2$$

$$\mathbf{B} \succeq 0$$

$\gamma \to 0$:

$$\min_{\alpha} \quad \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle$$

$$\text{s.t.} \quad 0 = \left(y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle\right)^2, \ \forall i \, .$$

min-norm interpolation with kernel $\mathbf{K}^{\mathsf{m}}$

Bartlett et al. (2020, 2021)
Ghorbani et al. (2020); Montanari et al. (2020)
Liang and Rakhlin (2018); Liang and Recht (2021)

VARIANCE INTERPOLATION W. CONFIDENCE

$\gamma \to 0$:

$$\min_{\alpha} \quad \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle$$
$$\text{s.t.} \quad 0 = \left(y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle\right)^2, \ \forall i .$$

min-norm interpolation with kernel $\mathbf{K}^{\mathsf{m}}$

Bartlett et al. (2020, 2021)
Ghorbani et al. (2020); Montanari et al. (2020)
Liang and Rakhlin (2018); Liang and Recht (2021)

$\gamma \to \infty$:

$$\min_{\mathbf{B}} \quad \mathrm{Tr}(\mathbf{K}^{\mathsf{v}} \mathbf{B})$$
$$\text{s.t.} \quad \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \overset{\text{interpolate}}{=} y_i^2, \ \forall i .$$
$$\mathbf{B} \succeq 0$$

min-norm <u>variance interpolation</u>

VARIANCE INTERPOLATION W. CONFIDENCE

$\gamma \to 0$:

$$\min_{\alpha} \quad \left\langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \right\rangle$$

$$\text{s.t.} \quad 0 = \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha \right\rangle \right)^2, \ \forall i \, .$$

min-norm interpolation with kernel $\mathbf{K}^{\mathsf{m}}$

Bartlett et al. (2020, 2021)
Ghorbani et al. (2020); Montanari et al. (2020)
Liang and Rakhlin (2018); Liang and Recht (2021)

$\gamma \to \infty$:

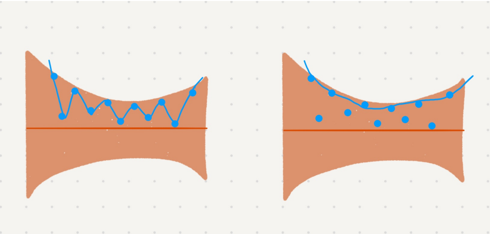
$$\min_{\mathbf{B}} \quad \mathrm{Tr}(\mathbf{K}^{\mathsf{v}} \mathbf{B})$$

$$\text{s.t.} \quad \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \overset{\text{confidence}}{\geq} y_i^2, \ \forall i \, .$$

$$\mathbf{B} \succeq 0$$

min-norm variance interpolation with confidence

not all realizations have large variability in y



VARIANCE INTERPOLATION W. CONFIDENCE

$\gamma \to 0$:

$$\min_{\alpha} \quad \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle$$
$$\text{s.t.} \quad 0 = \left(y_i - \langle \mathsf{K}_i^{\mathsf{m}}, \alpha \rangle\right)^2, \ \forall i \, .$$

min-norm interpolation with kernel $\mathbf{K}^{\mathsf{m}}$

Bartlett et al. (2020, 2021)
Ghorbani et al. (2020); Montanari et al. (2020)
Liang and Rakhlin (2018); Liang and Recht (2021)

$\gamma \to \infty$:

$$\min_{\mathbf{B}} \quad \mathrm{Tr}(\mathbf{K}^{\mathsf{v}} \mathbf{B})$$
$$\text{s.t.} \quad \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \overset{\text{confidence}}{\geq} y_i^2, \ \forall i \, .$$
$$\mathbf{B} \succeq 0$$

min-norm variance interpolation with confidence

not all realizations have large variability in y

role of the *tuning parameter* γ: trades off the conditional mean $\mathsf{m}(x)$ and variance $\mathsf{v}(x)$.

A small γ: a complex mean $\mathsf{m}(x)$, a parsimonious variance $\mathsf{v}(x)$ to explain the overall variability, and vice versa.

RELATED LITERATURE

Conformal Prediction:

Vovk et al. (2005); Shafer and Vovk (2008)

Residual Resampling:

Quantile Regression:

Koenker and Bassett Jr (1978); Koenker and Hallock (2001)
Belloni and Chernozhukov (2011); Belloni et al. (2019)

RELATED LITERATURE

Conformal Prediction: elegant theory based on exchangeability

Vovk et al. (2005); Shafer and Vovk (2008)

- motivated from online learning/sequential prediction
- user specify a **nonconformity measure** $A(B, z)$ with $z = (x, y)$
- conformal prediction alg.: enumerate all possibilities of $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, for each possibility, calculate $n+1$ nonconformity measures via leave-one-out

$$\alpha_i = A\big(\{z_1, \ldots, z_n, z\} \backslash \{z_i\}, z_i\big)$$

- include $y \in \widehat{\mathsf{PI}}(x)$ iff $\frac{\sum_i \mathbf{1}(\alpha_i \geq \alpha_{n+1})}{n+1} > 0.05$

exchangeability of α_i's

RELATED LITERATURE

Conformal Prediction: elegant theory based on exchangeability

Vovk et al. (2005); Shafer and Vovk (2008)

- motivated from online learning/sequential prediction
- user specify a **nonconformity measure** $A(B, z)$ with $z = (x, y)$
- conformal prediction alg.: enumerate all possibilities of $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, for each possibility, calculate $n + 1$ nonconformity measures via leave-one-out

$$\alpha_i = A\big(\{z_1, \ldots, z_n, z\} \backslash \{z_i\}, z_i\big)$$

- include $y \in \widehat{\mathsf{PI}}(x)$ iff $\frac{\sum_i \mathbf{1}(\alpha_i \geq \alpha_{n+1})}{n+1} > 0.05$

exchangeability of α_i's

Comparison:

- computation budget $n \times |\mathcal{Y}| \times |\mathcal{X}| \times \overbrace{\text{Oracle}(A)}^{\text{LOO refit of ML model}}$
- metric structure on $\mathcal{X}$ is not leveraged
- coverage guarantee is over the $\mathbb{P}_{\{(x_i, y_i)\}_{i=1}^n, (\mathbf{x}, \mathbf{y})}\big[\widehat{\mathsf{PI}}(\mathbf{x}) \text{ cover } \mathbf{y}\big] \geq 0.95$

RELATED LITERATURE

Conformal Prediction: elegant theory based on exchangeability

Vovk et al. (2005); Shafer and Vovk (2008)

- motivated from online learning/sequential prediction
- user specify a **nonconformity measure** $A(B, z)$ with $z = (x, y)$
- conformal prediction alg.: enumerate all possibilities of $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, for each possibility, calculate $n + 1$ nonconformity measures via leave-one-out

$$\alpha_i = A\big(\{z_1, \ldots, z_n, z\} \backslash \{z_i\}, z_i\big)$$

- include $y \in \widehat{\mathsf{PI}}(x)$ iff $\frac{\sum_i \mathbf{1}(\alpha_i \geq \alpha_{n+1})}{n+1} > 0.05$

exchangeability of α_i's

Comparison:

- computation budget $n \times |\mathcal{Y}| \times |\mathcal{X}| \times \overbrace{\mathrm{Oracle}(A)}^{\text{LOO refit of ML model}}$ vs. our SDP: n^2
- metric structure on $\mathcal{X}$ is not leveraged vs. our SDP: leverages metric structure in $\mathcal{X}$
- coverage guarantee is over the $\mathbb{P}_{\{(x_i,y_i)\}_{i=1}^n, (\mathbf{x},\mathbf{y})}\left[\widehat{\mathsf{PI}}(\mathbf{x}) \text{ cover } \mathbf{y}\right] \geq 0.95$ vs. our SDP: $\mathbb{P}_{(\mathbf{x},\mathbf{y})}\left[\widehat{\mathsf{PI}}(\mathbf{x}) \text{ cover } \mathbf{y} \mid \{(x_i, y_i)\}_{i=1}^n\right] \geq 0.95$ given 99.9999% of $\{(x_i, y_i)\}_{i=1}^n$

SDP: better computational complexity and potentially stronger coverage

RELATED LITERATURE

Residual Resampling: how to effectively pool local residuals in high dimensions? rigorous?

RELATED LITERATURE

Residual Resampling: how to effectively pool local residuals in high dimensions? rigorous?

Quantile Regression: estimate conditional quantile function $\widehat{\xi}^{\tau}(\cdot)$

$$\widehat{\xi}^{\tau}(\cdot) = \operatorname*{arg\,min}_{\xi} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}(y_i - \xi(x_i))$$

where $\tau \in (0,1)$ is a quantile parameter, $\rho_{\tau} : \mathbb{R} \to \mathbb{R}_+$ tilted absolute value function

not guaranteed $\tau_1 < \tau_2$, for all $x \in \mathcal{X}$, the estimated conditional quantile satisfies

$$\widehat{\xi}^{\tau_1}(x) < \widehat{\xi}^{\tau_2}(x)$$

$\Rightarrow$ empty conditional prediction intervals for several x

Koenker and Bassett Jr (1978); Koenker and Hallock (2001)
Belloni and Chernozhukov (2011); Belloni et al. (2019)

Theory of Non-asymptotic Coverage

ASSUMPTIONS

[S1] (Kernel and RKHS)
Kernel K is continuous, PSD and satisfies $\sup_{x \in \mathcal{X}} K(x, x) \leq C$.
Eigenvalues of the associated integral operator $\mathcal{T}$ satisfy $\lambda_j(\mathcal{T}) \leq C j^{-\tau}, j \in \mathbb{N}$ for some constant $\tau > 1$.

[S2] (Non-trivial uncertainty)
There exist constants $\eta \in (0, 1), \xi > 0$ such that $\Pr\left[\mathbf{y}^2 > \xi \cdot K(\mathbf{x}, \mathbf{x}) \mid \mathbf{x} = x\right] > \eta$ holds for all $x \in \mathcal{X}$.

[S3] (Non-wild uncertainty)
There exists a constant $\omega > 0$ such that $\Pr\left[\mathbf{y}^2 > t \cdot K(\mathbf{x}, \mathbf{x})\right] < \exp(-Ct^{\omega})$ for all $t \geq 1$.

ASSUMPTIONS

[S1] (Kernel and RKHS)
Kernel K is continuous, PSD and satisfies $\sup_{x \in \mathcal{X}} K(x, x) \le C$.
Eigenvalues of the associated integral operator $\mathcal{T}$ satisfy $\lambda_j(\mathcal{T}) \le Cj^{-\tau}, j \in \mathbb{N}$ for some constant $\tau > 1$.

[S2] (Non-trivial uncertainty)
There exist constants $\eta \in (0, 1)$, $\xi > 0$ such that $\Pr\left[\mathbf{y}^2 > \xi \cdot K(\mathbf{x}, \mathbf{x}) \mid \mathbf{x} = x\right] > \eta$ holds for all $x \in \mathcal{X}$.

[S3] (Non-wild uncertainty)
There exists a constant $\omega > 0$ such that $\Pr\left[\mathbf{y}^2 > t \cdot K(\mathbf{x}, \mathbf{x})\right] < \exp(-Ct^{\omega})$ for all $t \ge 1$.

Mild assumptions compared to strong distributional assumptions on $\mathbf{y}|\mathbf{x} = x$.

NON-ASYMPTOTIC COVERAGE

Define the objective value of the SDP

$$\widehat{\mathsf{Opt}}_n := \min_{\mathbf{B}} \quad \mathrm{Tr}(\mathbf{KB})$$
$$\text{s.t.} \quad \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq y_i^2, \; i = 1, \ldots, n \, .$$
$$\mathbf{B} \succeq 0$$

and the constructed prediction band with a confidence parameter $\delta \in (0, 1]$

$$\widehat{\mathsf{PI}}(x, \delta) = \left[\pm \sqrt{1+\delta} \cdot \sqrt{\widehat{\mathsf{v}}(x)} \right] \, .$$

Here $\widehat{\mathsf{v}}(x) := \langle \mathsf{K}_x, \widehat{\mathbf{B}}\mathsf{K}_x \rangle$ with $\mathsf{K}_x := [K(x, x_1), \ldots, K(x, x_n)]^\top \in \mathbb{R}^n$.

NON-ASYMPTOTIC COVERAGE

Define the objective value of the SDP

$$\widehat{\mathsf{Opt}}_n := \min_{\mathbf{B}} \quad \mathrm{Tr}(\mathbf{K}\mathbf{B})$$
$$\text{s.t.} \quad \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq y_i^2, \; i = 1, \ldots, n\,.$$
$$\mathbf{B} \succeq 0$$

$$\widehat{\mathsf{PI}}(x, \delta) = \left[\pm\sqrt{1+\delta} \cdot \sqrt{\widehat{\mathrm{v}}(x)}\right]\,.$$

Theorem (L.'21, non-asymptotic coverage).

Let [S1]-[S3] hold. For any $\delta \in (0, 1]$, the following non-asymptotic, data-dependent coverage guarantee holds,

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta)\right] \leq \delta^{-1}(\widehat{\mathsf{Opt}}_n \vee 1)\sqrt{C_{\tau,\xi,\eta,\omega} \cdot \frac{\log(n)}{n}}\,,$$

with prob. $1 - n^{-10}$ on $\{(x_i, y_i)\}_{i=1}^n$.

Here the constants $C_{\tau,\xi,\eta,\omega}, c_\omega$ only depend on parameters in [S1]-[S3].

NON-ASYMPTOTIC COVERAGE

Define the objective value of the SDP

$$\widehat{\mathsf{Opt}}_n := \min_{\mathbf{B}} \quad \mathrm{Tr}(\mathbf{KB})$$
$$\text{s.t.} \quad \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq y_i^2, \; i = 1, \ldots, n \,.$$
$$\mathbf{B} \succeq 0$$

$$\widehat{\mathsf{PI}}(x, \delta) = \left[\pm\sqrt{1+\delta} \cdot \sqrt{\widehat{\mathsf{v}}(x)} \right] .$$

Theorem (L.'21, non-asymptotic coverage).

Let [S1]-[S3] hold. For any $\delta \in (0,1]$, the following non-asymptotic, data-dependent coverage guarantee holds,

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}} \left[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta) \right] \leq \delta^{-1} (\widehat{\mathsf{Opt}}_n \vee 1) \sqrt{C_{\tau,\xi,\eta,\omega} \cdot \frac{\log(n)}{n}} ,$$

$$\text{and} \quad \widehat{\mathsf{Opt}}_n \leq \left[\log(n) \right]^{c_\omega} ,$$

with prob. $1 - n^{-10}$ on $\{(x_i, y_i)\}_{i=1}^n$.

Here the constants $C_{\tau,\xi,\eta,\omega}, c_\omega$ only depend on parameters in [S1]-[S3].

Some Remarks on the Coverage Theory

STRONG COVERAGE

- SDP prediction band will correctly cover a fresh data point $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$, with a **non-asymptotic** coverage probability (on the new data $\mathbf{x}, \mathbf{y}$)

$$1 - \delta^{-1} \frac{\mathrm{polylog}(n)}{\sqrt{n}} .$$

- With $\delta = 0.5$, the bandwidth $\mathrm{Length}\left[\widehat{\mathsf{PI}}(x)\right] = 2.45\sqrt{\widehat{\mathsf{v}}(x)}$ is at a heteroskedastic level adaptive to x.

STRONG COVERAGE

- SDP prediction band will correctly cover a fresh data point $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$, with a **non-asymptotic** coverage probability (on the new data $\mathbf{x}, \mathbf{y}$)

$$1 - \delta^{-1} \frac{\mathrm{polylog}(n)}{\sqrt{n}} \, .$$

- With $\delta = 0.5$, the bandwidth $\mathrm{Length}\big[\widehat{\mathsf{PI}}(x)\big] = 2.45\sqrt{\widehat{\mathsf{v}}(x)}$ is at a heteroskedastic level adaptive to x.

coverage can be arbitrary close to 1 with $n \uparrow \infty$ with a fixed confidence δ

STRONG COVERAGE

- SDP prediction band will correctly cover a fresh data point $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$, with a **non-asymptotic** coverage probability (on the new data $\mathbf{x}, \mathbf{y}$)

$$1 - \delta^{-1} \frac{\mathrm{polylog}(n)}{\sqrt{n}} .$$

- With $\delta = 0.5$, the bandwidth $\mathrm{Length}\left[\widehat{\mathsf{PI}}(x)\right] = 2.45\sqrt{\widehat{\mathsf{v}}(x)}$ is at a heteroskedastic level adaptive to x.

coverage can be arbitrary close to 1 with $n \uparrow \infty$ with a fixed confidence δ

holds essentially on $99.9999\% \le 1 - n^{-10}$ of the datasets $\{(x_i, y_i\}_{i=1}^n$

OPTIMALITY

Fix a 95% coverage

classic simple linear regression

$$\mathrm{Len}[\widehat{\mathsf{PI}}(x)] = \left(1 + \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}\right) \cdot 3.92\hat{s}$$

with $\hat{s} = \sqrt{\frac{\sum_i \hat{e}_i^2}{n-2}}$ being the estimated residual standard error.

our universal prediction interval

$$\mathrm{Len}[\widehat{\mathsf{PI}}(x)] = \left(1 + O^\star\left(\sqrt{\tfrac{1}{n}}\right)\right) \cdot \sqrt{\widehat{\mathsf{v}}(x)}$$

with the choice $\delta = O^\star(\sqrt{\frac{1}{n}})$.

OPTIMALITY

Fix a 95% coverage

classic simple linear regression

$$\mathrm{Len}[\widehat{\mathsf{PI}}(x)] = \left(1 + \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}}\right) \cdot 3.92\hat{s}$$

with $\hat{s} = \sqrt{\frac{\sum_i \hat{e}_i^2}{n-2}}$ being the estimated residual standard error.

our universal prediction interval

$$\mathrm{Len}[\widehat{\mathsf{PI}}(x)] = \left(1 + O^\star(\sqrt{\tfrac{1}{n}})\right) \cdot \sqrt{\widehat{\mathsf{v}}(x)}$$

with the choice $\delta = O^\star(\sqrt{\frac{1}{n}})$.

$\sqrt{\frac{1}{n}}$ fluctuation seems to indicate the optimality of our theory

DATA ADAPTIVITY

$\widehat{\mathsf{Opt}}_n$ of the convex optimization quantifies the uncertainty of the prediction band

DATA ADAPTIVITY

$\widehat{\mathsf{Opt}}_n$ of the convex optimization quantifies the uncertainty of the prediction band

A smaller $\widehat{\mathsf{Opt}}_n$ (computed based on the dataset)

$\Rightarrow$ a better confidence/coverage guarantee

$\Rightarrow$ a narrower prediction band overall

DATA ADAPTIVITY

$\widehat{\mathsf{Opt}}_n$ of the convex optimization quantifies the uncertainty of the prediction band

A smaller $\widehat{\mathsf{Opt}}_n$ (computed based on the dataset)

$\Rightarrow$ a better confidence/coverage guarantee

$\Rightarrow$ a narrower prediction band overall

SDP constructs the prediction band via its solution, and at the same time, reveals the confidence via its objective value.

DATA ADAPTIVITY

Our Theorem: Convex Optimization $\overset{\text{interface}}{\leftrightarrow}$ Uncertainty Quantification

DATA ADAPTIVITY

Our Theorem: Convex Optimization $\overset{\text{interface}}{\leftrightarrow}$ Uncertainty Quantification

- $\widehat{\mathsf{Opt}}_n$ is adaptive to the dataset
  $\Rightarrow$ our Theorem reveals which dataset allows for a better prediction band
- $\widehat{\mathsf{Opt}}_n = \|\widehat{\mathsf{v}}(\cdot)\|_\star^2$ is also a particular norm of the heteroskedastic variance function
  $\Rightarrow$ curiously, a simpler variance func. $\widehat{\mathsf{v}}(x)$ will simultaneously result in a narrower band and better coverage.

DATA ADAPTIVITY

Our Theorem: Convex Optimization $\overset{\text{interface}}{\leftrightarrow}$ Uncertainty Quantification

- $\widehat{\mathsf{Opt}}_n$ is adaptive to the dataset
  $\Rightarrow$ our Theorem reveals which dataset allows for a better prediction band
- $\widehat{\mathsf{Opt}}_n = \|\widehat{\mathsf{v}}(\cdot)\|_\star^2$ is also a particular norm of the heteroskedastic variance function
  $\Rightarrow$ curiously, a simpler variance func. $\widehat{\mathsf{v}}(x)$ will simultaneously result in a narrower band and better coverage.

Conventional wisdom: narrow band leads to poor coverage

Real Data Example

Fama-French 1993

	Value	Neutral	Growth
Small	Small Value	Small Neutral	Small Growth
Big	Big Value	Big Neutral	Big Growth

Fama and French (1993)

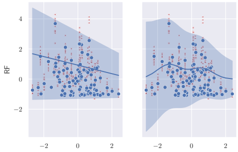
FAMA-FRENCH 1993

	Value	Neutral	Growth
Small	Small Value	Small Neutral	Small Growth
Big	Big Value	Big Neutral	Big Growth

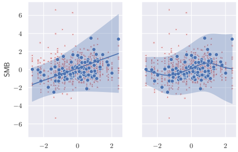
Fama and French (1993)

- Size: SMB = 1/3 (Small Value + Small Neutral + Small Growth)
  - 1/3 (Big Value + Big Neutral + Big Growth).
- Value: HML =1/2 (Small Value + Big Value)
  - 1/2 (Small Growth + Big Growth).
- Interest: RF, Market: Mkt - RF

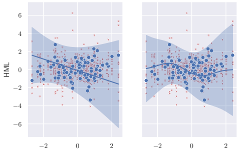
FAMA-FRENCH



(a) RF



(b) SMB



(c) HML

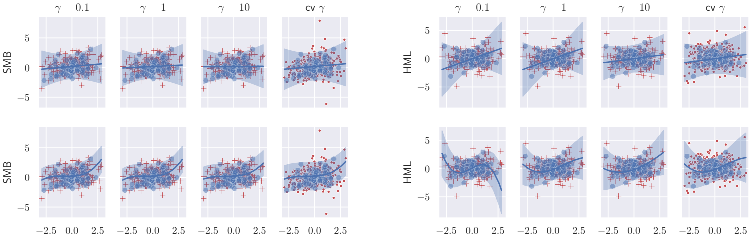
Fama and French (1993)

Fama-French

Table 2: Real data: Fama-French

	Kernel	Coverage	Median Len	Average Len
RF	lin $\mathsf{m}(x)$, quad $\mathsf{v}(x)$	**98.68%**	**4.3616**	**4.4358**
RF	rbf $\mathsf{m}(x)$, quad $\mathsf{v}(x)$	98.59%	4.5693	4.6847
SMB	lin $\mathsf{m}(x)$, quad $\mathsf{v}(x)$	95.77%	**5.2560**	**5.2798**
SMB	rbf $\mathsf{m}(x)$, quad $\mathsf{v}(x)$	**97.53%**	5.5407	5.4290
HML	lin $\mathsf{m}(x)$, quad $\mathsf{v}(x)$	96.56%	5.2822	5.5556
HML	rbf $\mathsf{m}(x)$, quad $\mathsf{v}(x)$	**97.27%**	**4.9180**	**5.3640**

TUNING PARAMETER γ



CONCLUSION

We address the uncertainty quantification dilemma
via semi-definite programming (SDP).

general/universal $\leftrightarrow$ rigorous/provable

machine learning $\overset{\text{SDP}}{\leftrightarrow}$ statistical inference

Enlarge the toolbox of applied researchers

CONCLUSION

We address the uncertainty quantification dilemma
via semi-definite programming (SDP).

general/universal $\leftrightarrow$ rigorous/provable

machine learning $\overset{\text{SDP}}{\leftrightarrow}$ statistical inference

Enlarge the toolbox of applied researchers

Be confident about (black-box) machine learning models, rigorously!

Thank you!

- **Liang, T.** (2021). — Universal Prediction Band via Semi-Definite Programming. *arXiv:2103.17203*

- **Liang, T.** & Recht, B. (2021). — Interpolating Classifiers Make Few Mistakes.
  arXiv:2101.11815
- **Liang, T.** & Sur, P. (2020). — A Precise High-Dimensional Asymptotic Theory for Boosting and Min-L1-Norm Interpolated Classifiers.
  arXiv:2002.01586
- **Liang, T.**, Rakhlin, A. & Zhai, X. (2019). — On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels.
  Conference on Learning Theory (COLT), 2020
- **Liang, T.**, Tran-Bach, H. (2020). — Mehler's Formula, Branching Process, and Compositional Kernels of Deep Neural Networks.
  Journal of the American Statistical Association, 2021
- **Liang, T.** & Rakhlin, A. (2018). — Just Interpolate: Kernel "Ridgeless" Regression Can Generalize.
  The Annals of Statistics, 2020
- Farrell, M., **Liang, T.** & Misra, S. (2018). —– Deep Neural Networks for Estimation and Inference.
  Econometrica, 2021