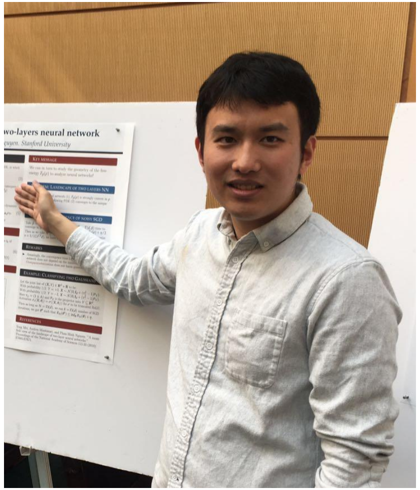
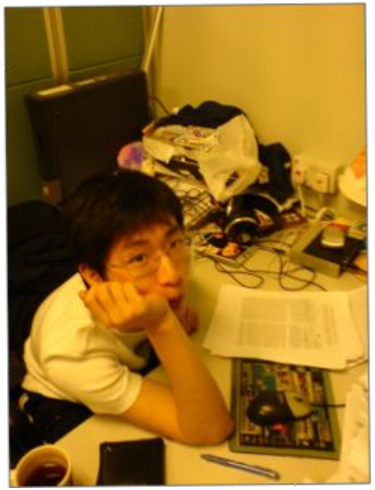
Understanding the Under-Coverage Bias in Uncertainty Estimation and Calibration

Yu Bai
Salesforce Research

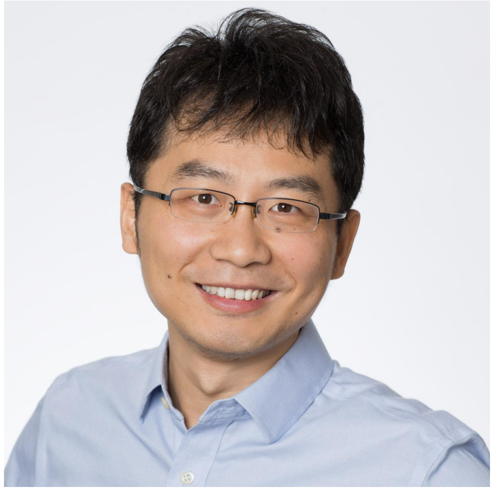
Collaborators



Song Mei (UC Berkeley)



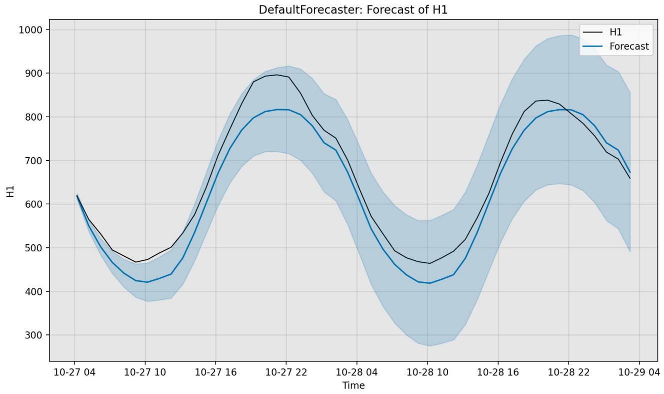
Huan Wang (Salesforce)



Caiming Xiong (Salesforce)

Uncertainty quantification for prediction problems

Enhance point prediction with a quantification of the associated uncertainty.



time series forecasting

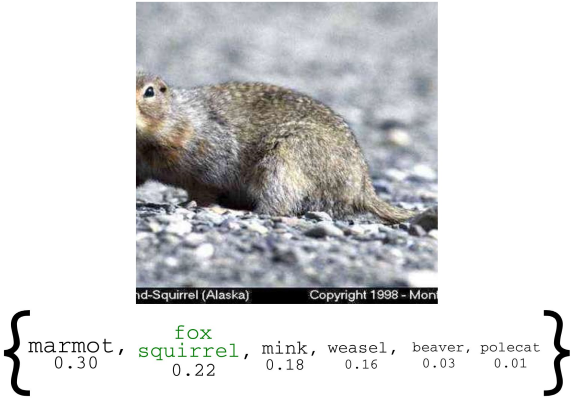


image classification

Image source:
Left: Merlion library, Salesforce.
Right: Uncertainty Sets for Image Classification using Conformal Prediction, Angelopoulos et al. 2021.

Notions of uncertainty quantification

Many existing notions of uncertainty quantification:

- Regression: variance estimation, **quantiles** / prediction intervals
- Classification: **calibration**, label prediction sets
- Others: OOD detection, ...

Quantiles / prediction intervals

salesforce

High-probability upper / lower bounds of y|x with good (marginal) **coverage**

$$\mathrm{Coverage}(\widehat{f}) = \mathbb{P}_{(X,Y)}(Y \le \widehat{f}(X)) \ge \underline{\alpha}$$

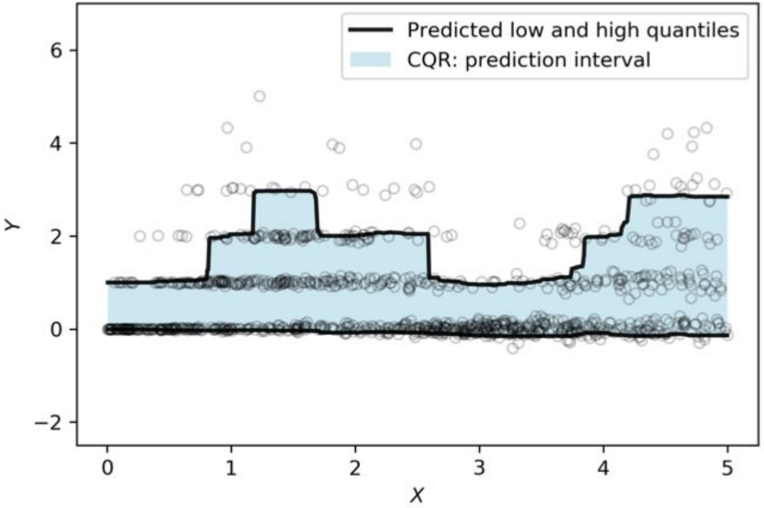
e.g. 0.9, 0.95

Quantiles / prediction intervals

High-probability upper / lower bounds of y|x with good (marginal) **coverage**

$$\text{Coverage}(\hat{f}) = \mathbb{P}_{(X,Y)}(Y \leq \hat{f}(X)) \geq \underline{\alpha}$$

e.g. 0.9, 0.95



One-sided: quantiles
Two-sided: prediction intervals

Image source: Conformalized Quantile Regression, Romano et al. 2019.

Classical methods for learning quantiles

- **Parametric estimation** (Cox 1975, Lawless & Fredette 2005, ...)

  Assume parametric family $\{p_\theta(y|x)\}_{\theta\in\Theta}$, get estimate $\widehat{\theta}$ from observed data
  Then take

$$\widehat{f}(x) := \alpha \text{ upper quantile of } p_{\widehat{\theta}}(\cdot|x)$$

Classical methods for learning quantiles

- **Parametric estimation** (Cox 1975, Lawless & Fredette 2005, ...)

  Assume parametric family $\{p_\theta(y|x)\}_{\theta\in\Theta}$, get estimate $\widehat{\theta}$ from observed data
  Then take

$$\widehat{f}(x) := \alpha \text{ upper quantile of } p_{\widehat{\theta}}(\cdot|x)$$

Approximate coverage when family is correct + sample size large enough so that $\widehat{\theta} \approx \theta_\star$

Classical methods for learning quantiles

- **Quantile regression** (Koenker & Bassett 1978, ...)

  Directly learn a quantile function $f_{\widehat{\theta}}$ by minimizing the *pinball loss* on the data:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell^{\alpha}(y_i - f_{\theta}(\mathbf{x}_i)).$$

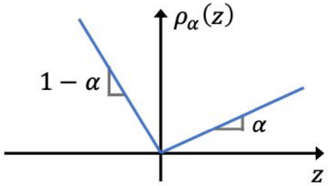


Figure 1: Visualization of the pinball loss function in (6), where $z = y - \hat{y}$.

Classical methods for learning quantiles

- **Quantile regression** (Koenker & Bassett 1978, ...)

  Directly learn a quantile function $f_{\widehat{\theta}}$ by minimizing the *pinball loss* on the data:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell^{\alpha}(y_i - f_\theta(\mathbf{x}_i)).$$

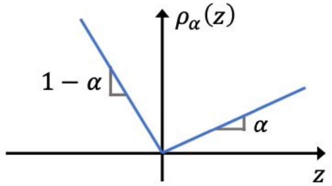


Figure 1: Visualization of the pinball loss function in (6), where $z = y - \hat{y}$.

Approximate coverage if family $\{f_\theta\}$ contains true α-quantile of $Y|X$ + large enough sample size

Over-coverage vs. under-coverage

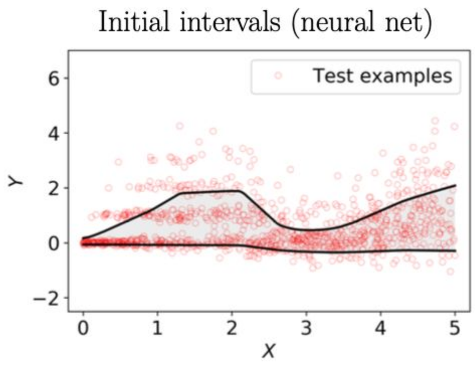
Sign of the coverage bias $\mathrm{Coverage}(\widehat{f}) - \alpha$ matters.

Over-coverage: $\mathrm{Coverage}(\widehat{f}) > \alpha$ 😃 (just over-conservative, but achieves desired coverage)

Under-coverage: $\mathrm{Coverage}(\widehat{f}) < \alpha$ 😟 (does not achieve desired coverage)

Quantile regression exhibits under-cover bias

Empirically, quantile regression is often found to **under-cover** (esp. with neural nets).



Target coverage level: 90%
Actual coverage: 66.77%

Image source: Conformalized Quantile Regression, Romano et al. 2019.

Quantile regression exhibits under-cover bias

Empirically, quantile regression is often found to **under-cover** (esp. with neural nets).

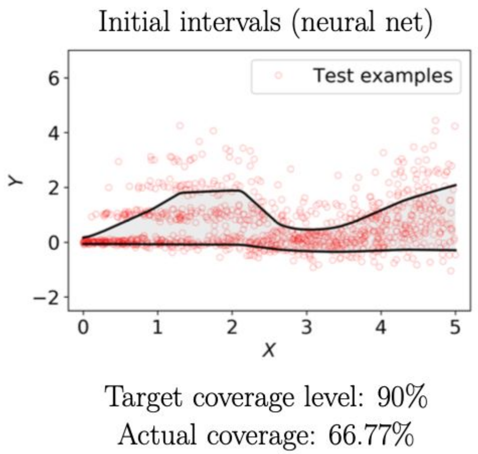


Image source: Conformalized Quantile Regression, Romano et al. 2019.

😃 Recent approaches such as conformal prediction can fix this (Vovk et al. 2005, Lei et al. 2018, …).

🤔 **Existing "approximate coverage" theories do not explain this under-coverage bias**.

Existing theories cannot tell under- or over-coverage

- Asymptotic guarantees (Koencker & Bassett, 1978):

Existing theories cannot tell under- or over-coverage

- Asymptotic guarantees (Koencker & Bassett, 1978):

  Fix num parameters d, sample size $n \to \infty$:

$$\sqrt{n}(\widehat{\theta} - \theta_\star) \xrightarrow{d} \mathsf{N}(0, V) \quad \Longrightarrow \quad \sqrt{n}(\mathrm{Coverage}(f_{\widehat{\theta}}) - \alpha) \xrightarrow{d} \mathsf{N}(0, \tau^2)$$

  Coverage bias has equal chance to be >0 or <0 in asymptotic regime.

Existing theories cannot tell under- or over-coverage

- Finite-sample bounds via *self-calibration inequalities* (Steinwart & Christmann, 2011):

Existing theories cannot tell under- or over-coverage

- Finite-sample bounds via *self-calibration inequalities* (Steinwart & Christmann, 2011):

Any fixed n, d:

$$\|\widehat{\theta} - \theta_\star\|_2 \le C\sqrt{\underline{R(f_{\widehat{\theta}})} - \underline{R(f_{\theta_\star})}}$$

Population (expected) pinball loss

$$\Longrightarrow \left|\mathrm{Coverage}(f_{\widehat{\theta}}) - \alpha\right| \le C\sqrt{\frac{\mathrm{Comp}(\{f_\theta\})}{n}}$$

Capacity of function class (e.g. Rademacher complexity)

Cannot tell the sign of the coverage bias.

Linear Quantile Regression Exhibits Under-Coverage

Data follows **linear model**:

$$y = \mathbf{w}_\star^\top \mathbf{x} + z, \quad \text{where} \quad \mathbf{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d), \quad z \sim P_z.$$

Use quantile regression to learn a **linear quantile function** (with bias) at target level $\alpha \in (0.5, 1)$:

$$\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}$$

Linear Quantile Regression Exhibits Under-Coverage

Data follows **linear model**:

$$y = \mathbf{w}_\star^\top \mathbf{x} + z, \quad \text{where} \quad \mathbf{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d), \quad z \sim P_z.$$

Use quantile regression to learn a **linear quantile function** (with bias) at target level $\alpha \in (0.5, 1)$:

$$\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}$$

Main Theorem: In the above setup, suppose $n, d \to \infty, d/n \to \kappa \in (0, \kappa_0]$, then we have

$$\text{Coverage}(\widehat{f}) \xrightarrow{p} C_{\alpha,\kappa} < \alpha,$$

that is, well-specified linear quantile regression has an **under-coverage** bias.

Further, for small κ we have the local expansion

$$C_{\alpha,\kappa} = \alpha - (\alpha - 1/2)\kappa + o(\kappa).$$

i.e. under-coverage bias has order $\Theta(\kappa) = \Theta(d/n)$.

Linear Quantile Regression Exhibits Under-Coverage

Main Theorem: In the above setup, suppose $n, d \to \infty, d/n \to \kappa \in (0, \kappa_0]$, then we have

$$\mathrm{Coverage}(\widehat{f}) \xrightarrow{p} C_{\alpha,\kappa} < \alpha,$$

that is, well-specified linear quantile regression has an **under-coverage** bias.

Further, for small κ we have the local expansion

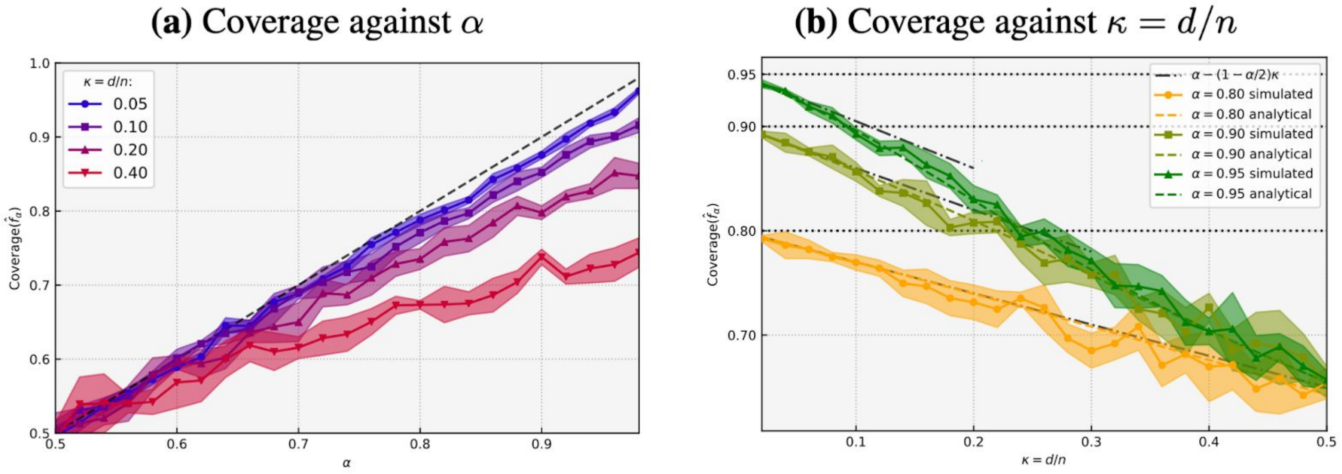
$$C_{\alpha,\kappa} = \alpha - (\alpha - 1/2)\kappa + o(\kappa).$$

$$\boxed{\begin{array}{l} \alpha = 0.9,\ n = 10d\ (\kappa = d/n = 0.1) \\ \Longrightarrow C_{\alpha,\kappa} \approx 0.86 \end{array}}$$

i.e. under-coverage bias has order $\Theta(\kappa) = \Theta(d/n)$.

Simulations

On Gaussian linear model (d=100), under-coverage bias matches our theoretical prediction.



Real data experiments

Quantile regression with {linear model, NNs} on real data

Table 1: Coverage (%) of quantile regression on real data at nominal level $\alpha = 0.9$. Each entry reports the test-set coverage with mean and std over 8 random seeds. (d, n) denotes the {feature dim, # training examples}.

Dataset	Linear	MLP-3-64	MLP-3-512	MLP-freeze-3-512	d	n
Community	88.63±1.53	76.46±1.41	63.09±2.91	87.85±1.30	100	1599
Bike	89.64±0.44	88.75±0.91	87.67±0.49	89.27±0.57	18	8708
Star	89.48±2.56	83.14±1.76	69.71±1.82	88.05±2.42	39	1728
MEPS_19	90.09±0.72	85.46±0.96	78.55±0.93	89.03±0.51	139	12628
MEPS_20	90.06±0.57	86.52±0.65	80.77±0.72	89.60±0.28	139	14032
MEPS_21	89.99±0.39	83.79±0.52	73.09±0.82	89.15±0.36	139	12524
Nominal (α)	90.00	90.00	90.00	90.00	-	-

Overview of techniques

Step 1: Express coverage as function of parameter estimation errors

$$\mathrm{Coverage}(\widehat{f}) = \mathbb{E}_{G\sim \mathsf{N}(0,1)}[\Phi_z(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b})].$$

Overview of techniques

Step 1: Express coverage as function of parameter estimation errors

$$\mathrm{Coverage}(\widehat{f}) = \mathbb{E}_{G\sim \mathsf{N}(0,1)}[\Phi_z(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b})].$$

Step 2: *High-dimensional proportional limit* analysis of estimation error

(Donoho & Montanari 2013, Thrampoulidis et al. 2016, Sur & Candes 2019, ...):

As $n, d \to \infty, d/n \to \kappa \in (0, \infty)$,

$$\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 \xrightarrow{p} \tau_\star(\kappa), \quad \text{and} \quad \widehat{b} \xrightarrow{p} b_\star(\kappa).$$

Above, quantities $(\tau_\star(\kappa), b_\star(\kappa), \lambda_\star(\kappa))$ are the solution to a 3x3 system of nonlinear equations.

😐 solutions no closed-form.

Overview of techniques

Step 1: Express coverage as function of parameter estimation errors

$$\mathrm{Coverage}(\widehat{f}) = \mathbb{E}_{G\sim \mathsf{N}(0,1)}[\Phi_z(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b})].$$

Step 2: *High-dimensional proportional limit* analysis of estimation error (Donoho & Montanari 2013, Thrampoulidis et al. 2016, Sur & Candes 2019, ...):

As $n, d \to \infty, d/n \to \kappa \in (0, \infty)$,

$$\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 \xrightarrow{p} \tau_\star(\kappa), \quad \text{and} \quad \widehat{b} \xrightarrow{p} b_\star(\kappa).$$

Above, quantities $(\tau_\star(\kappa), b_\star(\kappa), \lambda_\star(\kappa))$ are the solution to a 3x3 system of nonlinear equations.

😐 solutions no closed-form.

Step 3: *Local linear analysis* of solutions at small κ:

3x3 nonlinear system ≈ (Linearized) 3x3 linear system with closed-form solutions

Classification calibration

Calibration: A commonly used notion of uncertainty in classification.

Calibration Error = |Confidence - Accuracy|

Of all the days where the model predicted rain with 80% probability, what fraction did we observe rain?

- 80% implies perfect calibration
- Less than 80% implies model is overconfident
- Greater than 80% implies model is under-confident

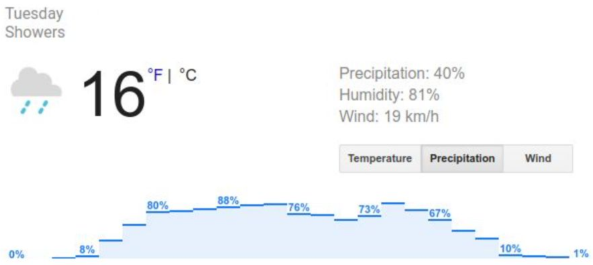


Image source: Practical Uncertainty Estimation & Out-of-Distribution Robustness in Deep Learning, NeurIPS 2020 tutorial

Over- and under-confidence

For any binary classifier, its calibration error at level $p \in (0.5, 1)$ is defined as

$$\Delta_p^{\mathsf{cal}}(\widehat{f}) := p - \mathbb{P}_{(X,Y)\sim P}\Big(Y = 1 \mid \widehat{f}(\mathbf{X}) = p\Big)$$

Over-confident: $\Delta_p^{\mathsf{cal}}(\widehat{f}) > 0$ (when model predicts 80% raining, actually 70% chance of raining)

$\Rightarrow$ *under-estimates* uncertainty in the data.

Well-specified logistic regression is over-confident

Solve binary linear logistic regression on realizable data:

$$P: \quad \mathbf{X} \sim \mathsf{N}(0, \mathbf{I}_d), \quad \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \sigma(\mathbf{w}_\star^\top \mathbf{x}),$$

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\left[\log(1 + \exp(\mathbf{w}^\top \mathbf{x}_i)) - y_i \mathbf{w}^\top \mathbf{x}_i\right].$$

Well-specified logistic regression is over-confident

Solve binary linear logistic regression on realizable data:

$$P: \quad \mathbf{X} \sim \mathsf{N}(0, \mathbf{I}_d), \quad \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \sigma(\mathbf{w}_\star^\top \mathbf{x}),$$

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \Big[\log(1 + \exp(\mathbf{w}^\top \mathbf{x}_i)) - y_i \mathbf{w}^\top \mathbf{x}_i \Big].$$

Theorem: In the above setting, suppose $n, d \to \infty, d/n \to \kappa \leq \kappa_0$

For any $p \in (0.5, 1)$, its calibration error at p converges to the following limit

$$\Delta_p^{\mathrm{cal}}(\widehat{f}) = p - \mathbb{P}(Y = 1 | \widehat{f}(X) = p)$$

$$\xrightarrow{a.s.} C_{p,\kappa} = C_p \kappa + o(\kappa), \quad C_p > 0$$

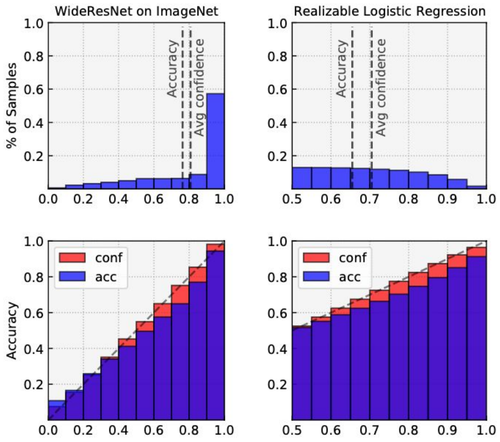
That is, logistic regression is over-confident by an amount of $\Theta(\kappa) = \Theta(d/n)$.

Similar techniques (proportional limit theory + local linear analysis of non-linear system).

Over-confidence in classification

Large neural nets are over-confident (Guo et al. 2017);

Realizable logistic regression (n=2000, d=100) also exhibits over-confidence, agreeing with our theory.



x-axis: confidence (predicted top probability) of the learned classifier

Conclusion & future directions

- First precise theoretical characterization of under-coverage bias in uncertainty quantification
  - Coverage in linear quantile regression
  - Calibration in binary classification w/ linear logistic regression
- Take-away: Under-estimation of data uncertainty is quite prevalent
  - Further theories? (e.g. non-linear models)
- How can we inspire new correction methods for practitioners

Thank you!

[References]

- *Understanding the Under-Coverage Bias in Uncertainty Estimation*. Yu Bai, Song Mei, Huan Wang, Caiming Xiong. NeurIPS 2021.
- *Don't Just Blame Over-Parametrization for Over-Confidence: Theoretical Analysis of Calibration in Binary Classification*. Yu Bai, Song Mei, Huan Wang, Caiming Xiong. ICML 2021.

Backup Slides

Nonlinear system for the coverage result

$$\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 \xrightarrow{p} \tau_\star(\kappa), \quad \text{and} \quad \widehat{b} \xrightarrow{p} b_\star(\kappa).$$

$$\begin{cases} \tau^2\kappa = \lambda^2 \cdot \mathbb{E}_{(G,Z)\sim\mathsf{N}(0,1)\times P_z}[e'_{\ell_b^\alpha}(\tau G + Z; \lambda)^2], \\ \tau\kappa = \lambda \cdot \mathbb{E}_{(G,Z)\sim\mathsf{N}(0,1)\times P_z}[e'_{\ell_b^\alpha}(\tau G + Z; \lambda)G], \\ 0 = \mathbb{E}_{(G,Z)\sim\mathsf{N}(0,1)\times P_z}[e'_{\ell_b^\alpha}(\tau G + Z; \lambda)], \end{cases}$$

$$e_\ell(x;\tau) \;=\; \min_v \tfrac{1}{2\tau}(x-v)^2 + \ell(v) \qquad\qquad \ell_b^\alpha \;=\; \ell^\alpha(t-b)$$