Approximate Conditional Coverage via Neural Model Approximations

Allen Schmaltz and Danielle Rasooly

re.express

Overview

We construct prediction sets over Transformer networks, via KNN-based approximations and constrained sampling, obtaining reliable assumption- and parameter-light approximate conditional coverage even in the presence of distribution shifts.

Split-conformal prediction sets for classification

- Computationally expensive blackbox: F
- Training dataset: $\mathcal{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^{I}$ with $Y_i \in \mathcal{Y} = \{1, \ldots, C\}$
- Held-out labeled calibration dataset: $\mathcal{D}_{\text{ca}} = \{(X_j, Y_j)\}_{j=I+1}^{N=I+J}$
- Seek: A prediction set $\hat{\mathcal{C}}(X_{N+1}) \in 2^C$ for a new, unseen test instance X_{N+1} from $\mathcal{D}_{\text{te}}$
    - Contains the true label with coverage level $1-\alpha \in (0,1)$ *on average*
- Finite-sample *marginal* guarantee:
    - $\mathbb{P}\left\{Y_{N+1} \in \hat{\mathcal{C}}(X_{N+1})\right\} \geq 1-\alpha$

      Quantile threshold
        - Via $\hat{\mathcal{C}}(x_{N+1}) = \left\{c \in \mathcal{Y} : \hat{\pi}^c(x_{N+1}) \geq \hat{\tau}^{\alpha}\right\}$, where $\hat{\tau}^{\alpha} = 1 - \hat{l}^{\alpha}$
- Finite-sample *conditional* coverage:

  Not possible without additional assumptions

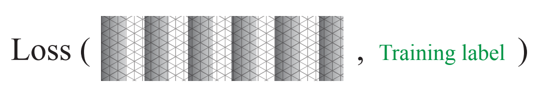
$$\mathbb{P}\left\{Y_{N+1} \in \hat{\mathcal{C}}(X_{N+1}) \mid X_{N+1} = x\right\} \geq 1-\alpha$$

- Finite-sample *approximate conditional* coverage:

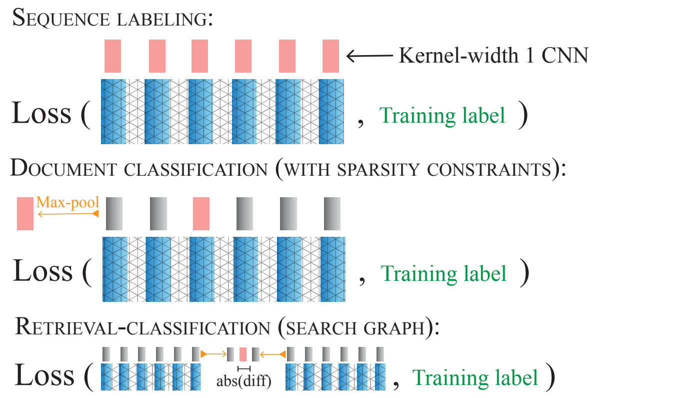
$$\mathbb{P}\left\{Y_{N+1} \in \hat{\mathcal{C}}(X_{N+1}) \mid X_{N+1} \in \mathcal{B}(x), Y_{N+1} = y\right\} \geq 1-\alpha, \text{ with } P_X(\mathcal{B}(x)) \geq \xi$$

ADMIT: A general framework for constructing, constraining, and analyzing point predictions and distribution-free prediction sets for deep neural networks.

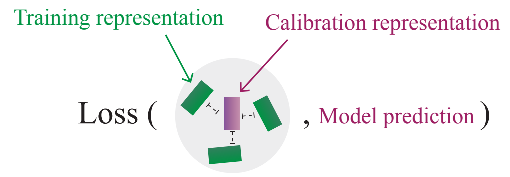
1. (Pre-) Train (& fine-tune) deep network, as usual.



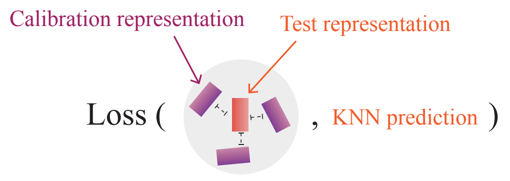
2. Freeze network. Add & train a memory layer for TASK. Extract exemplar representations.



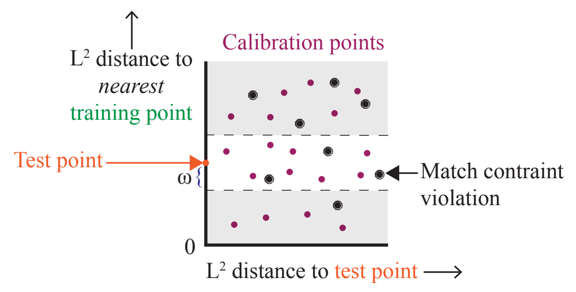
3. Train a KNN-based model approximation over exemplar representations from the memory layer, relating a new instance to training instances (predictions and ground-truth labels): $f(x)_{\text{tr}}^{\text{KNN}}$



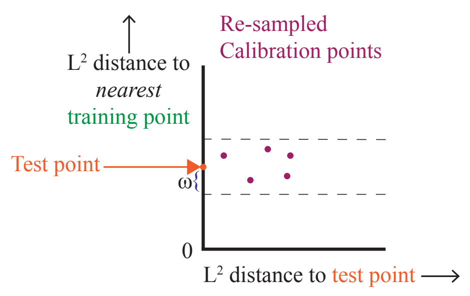
4. Train another KNN-based model approximation, relating a new test instance to representations and *KNN predictions* over the calibration set: $f(x)_{\text{ca}}^{\text{KNN}}$



5. Calculate unique quantile thresholds *for each label for each test point* from the constrained set of calibration points within the distance band.



6. Optionally, re-sample the calibration set to be more similar to the test distribution. Repeat Step 5.

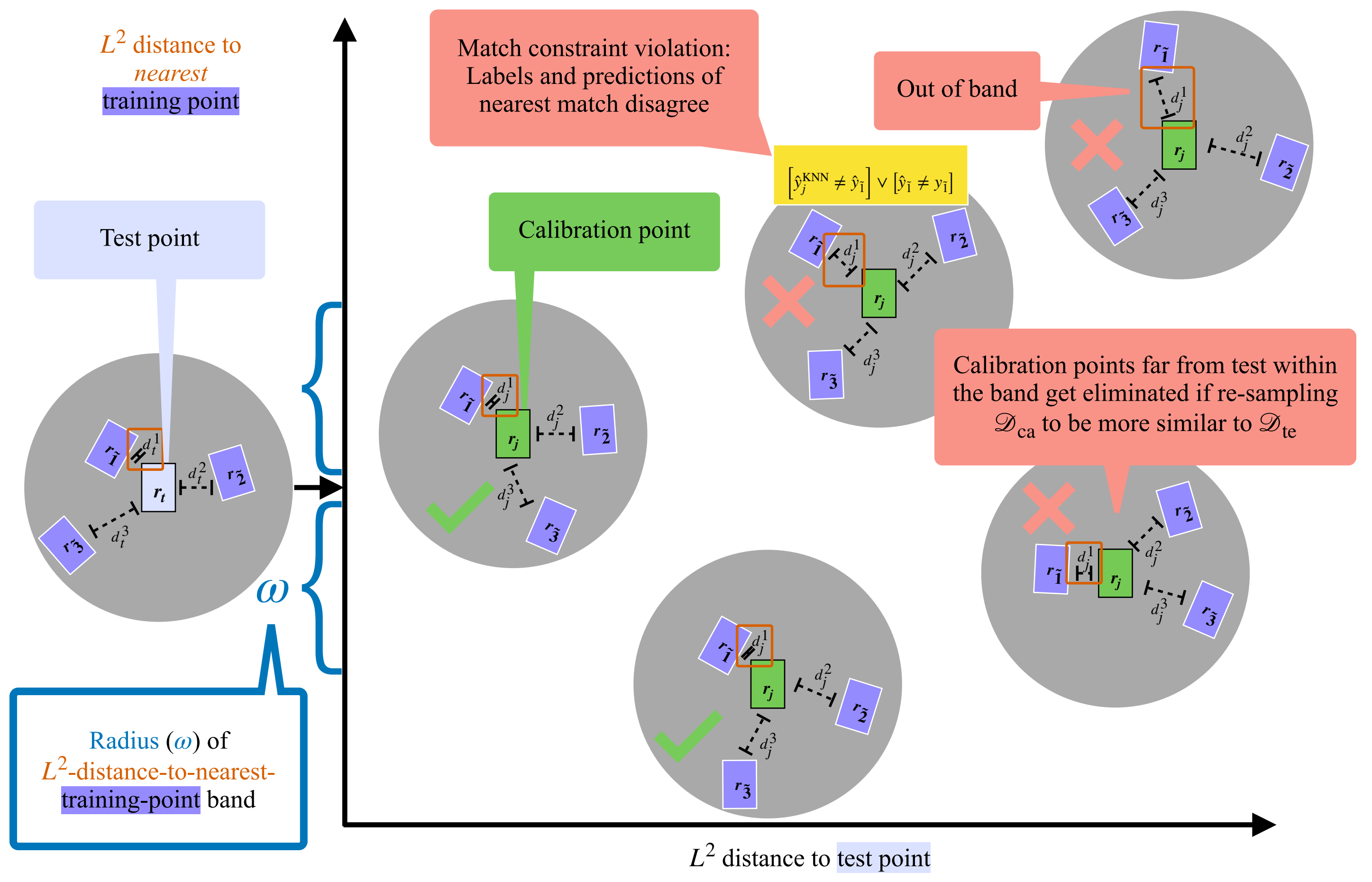


7. Optionally, condition on prediction set membership. Additional heuristics screen unreliable cases. (See text.)

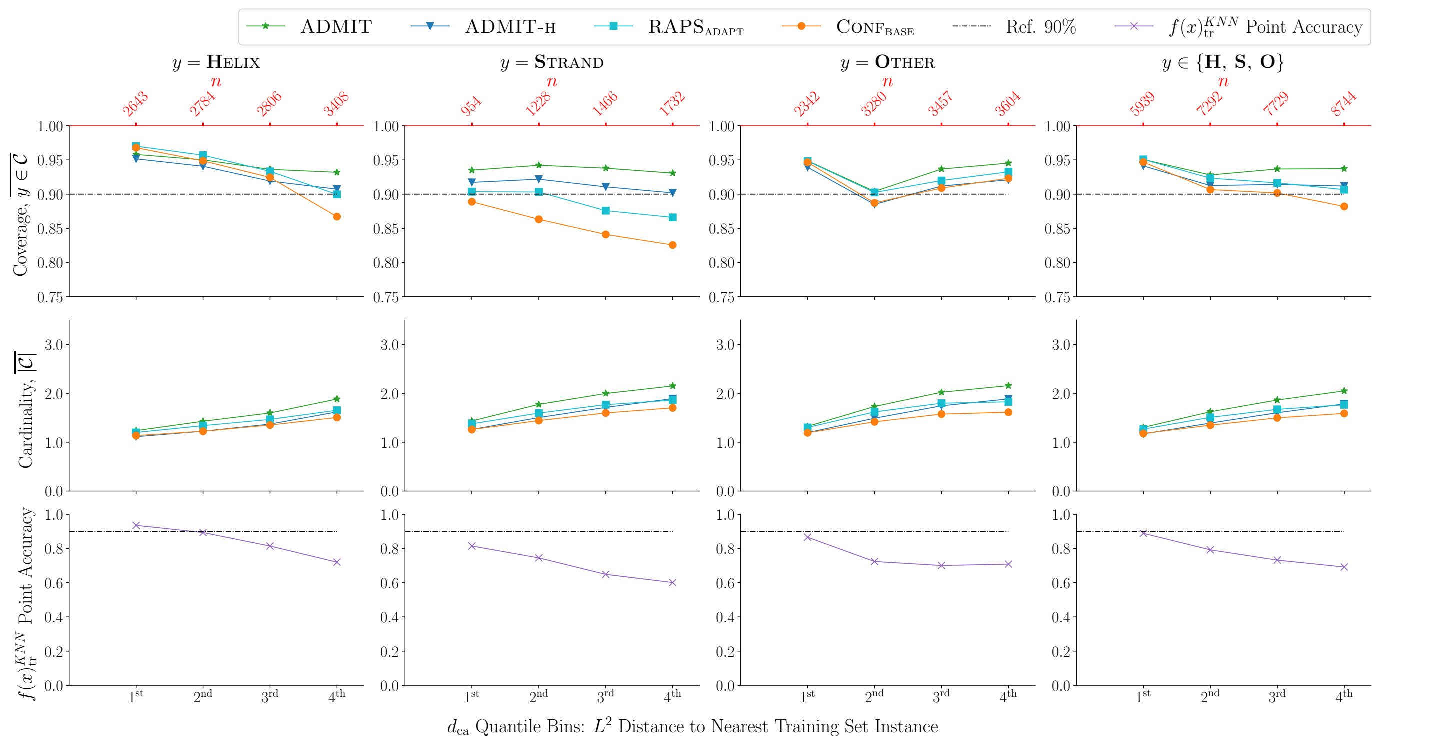
Weighted KNN approximations of the deep network encode strong signals for prediction reliability:

Predictions become less reliable at distances farther from the training set and with increased label and prediction mismatches among the nearest matches.

Key: Construct a distance band around the test point containing a constrained set of calibration points (✓), excluding dissimilar points (✗)



Empirical behavior (*see paper for additional results*)



Coverage, cardinality, and point accuracy for the TS115 test set from the PROTEIN task.

Contact: allen@re.express, danielle@re.express