Certifying Out-of-Domain Generalization for Blackbox Functions

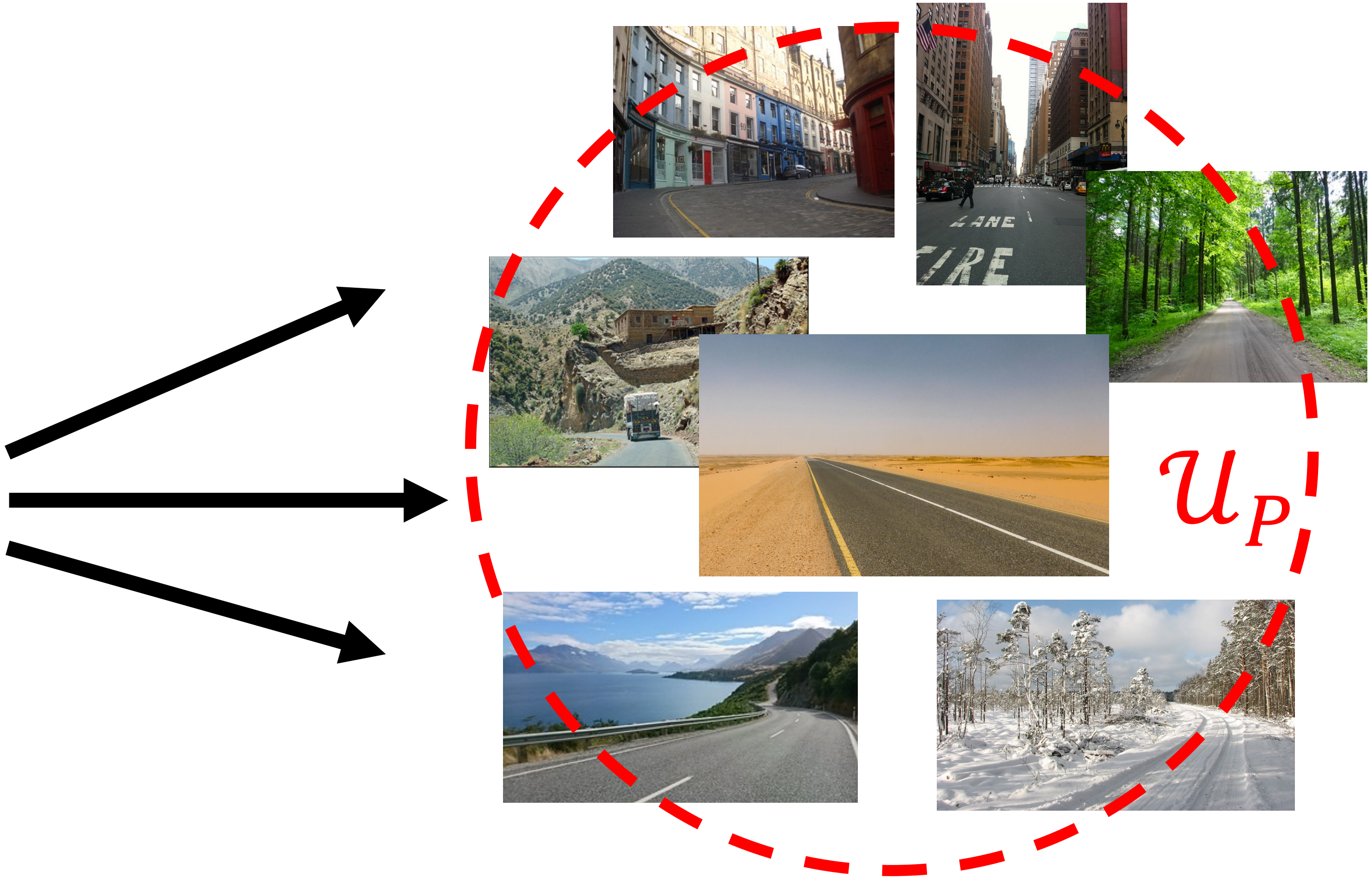
Maurice Weber[1] **Linyi Li[2]** **Boxin Wang[2]** **Zhikuan Zhao[1]** **Bo Li[1]** **Ce Zhang[1]**

[1]ETH Zürich

[2]University of Illinois Urbana-Champaign

Motivation

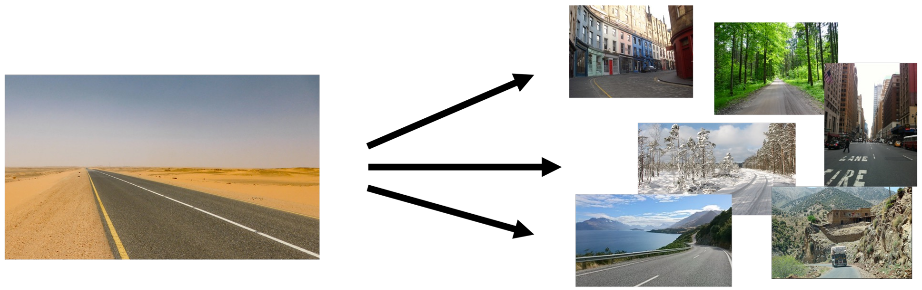
Train Distribution P



Test Distributions $Q \in \mathcal{U}_P$

Certificate: $\forall Q\colon \texttt{dist}(P, Q) \leq \rho \implies E_{Z\sim Q}[\ell(Z)] \leq C_\ell(\rho, P)$

Main Contributions



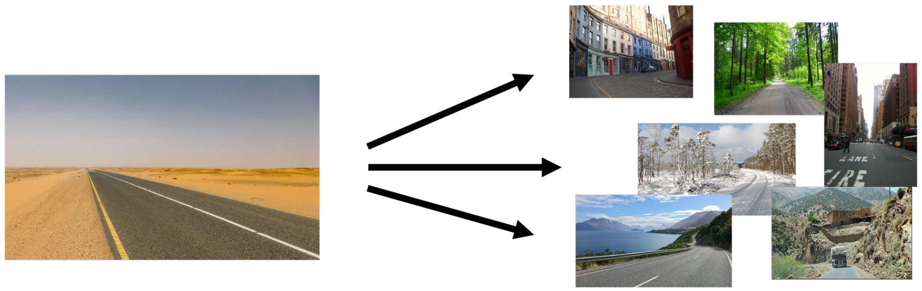
Train Distribution P Test Distributions $Q \in \mathcal{U}_P$

Provide a rigorous upper bound $C_\ell(\rho, P)$ with

$$\sup_{Q \in \mathcal{U}_P} E_{Z \sim Q}[\ell(Z)] \leq C_\ell(\rho, P)$$

- under minimal assumptions on the model ℓ ***(black-box)***
- that can be estimated from a finite, in-domain ***sample*** $\boldsymbol{Z_1, \ldots, Z_n \sim P}$
- which scales to ***large datasets*** and state-of-the-art ***neural networks***

Certifying Out-of-Domain Generalization



Train Distribution P Test Distributions $Q \in \mathcal{U}_P$

$$\sup_{Q \in \mathcal{U}_P} E_{Z \sim Q}[\ell(Z)] \leq C_\ell(\rho, P)$$

Uncertainty Set: $\mathcal{U}_P = \{Q \mid H(P, Q) \leq \rho\}$, (H = Hellinger distance)

Model / Loss function: $\forall z \in \mathcal{Z} : 0 \leq \ell(z) \leq M$, (Boundedness, Positivity)

Technique

1. Express Expectation values as inner products
2. Use Non-negativity of Gram matrices to get a robustness condition

Main Result

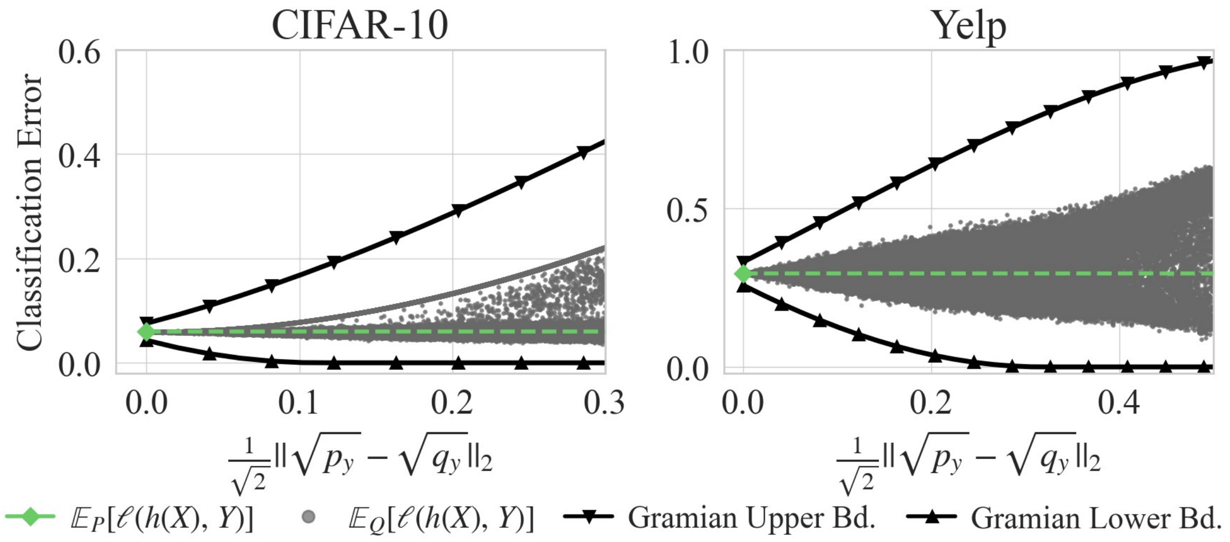
Let $\ell: \mathcal{Z} \to [0, M]$ be a loss function, P a probability measure on $\mathcal{Z}$. For $\rho > 0$, we have

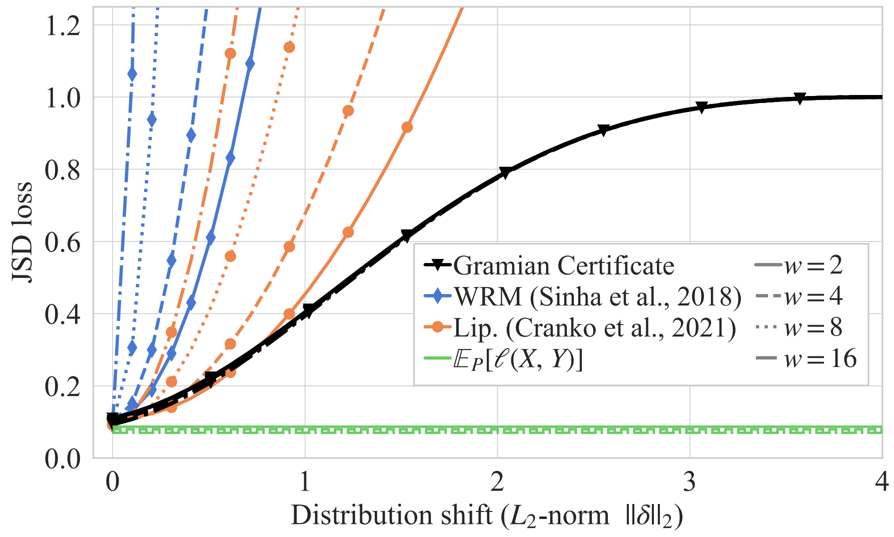
$$\sup_{Q:\, H(P,Q) \le \rho} E_Q[\ell] \le E_P[\ell] + 2C_\rho \sqrt{V_P[\ell]} + \phi_{M,\rho}(E_P[\ell], V_P[\ell])$$

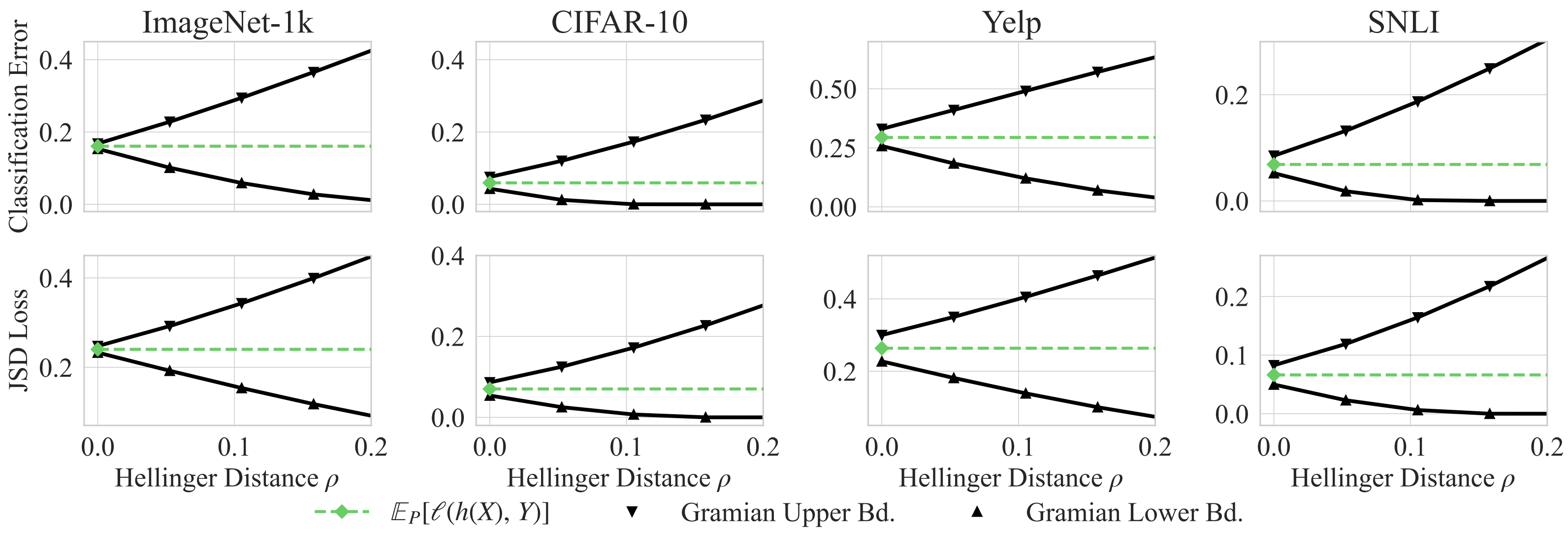
where ρ is required to satisfy $\rho^2 \le 1 - \left[1 + \frac{(M - E_P[\ell])^2}{V_P[\ell]}\right]^{-1/2}$.

- Bound only requires ***blackbox*** access to the model ℓ.
- Can be estimated from ***finite samples*** using concentration inequalities.
- ***Monotonically increasing*** in $E_P[\ell]$ and $V_P[\ell]$.

Experiments







Conclusion and Outlook

- We have presented a technique to ***certify out-of-domain generalization***
  - for uncertainty sets defined via the ***Hellinger*** distance
  - which only requires ***blackbox*** access to the model and loss function
  - and hence scales to ***large-scale*** models and datasets
- Future work
  - explore more ***specific distribution shifts*** to get tighter certificates
  - different ***applications***, beyond classification tasks