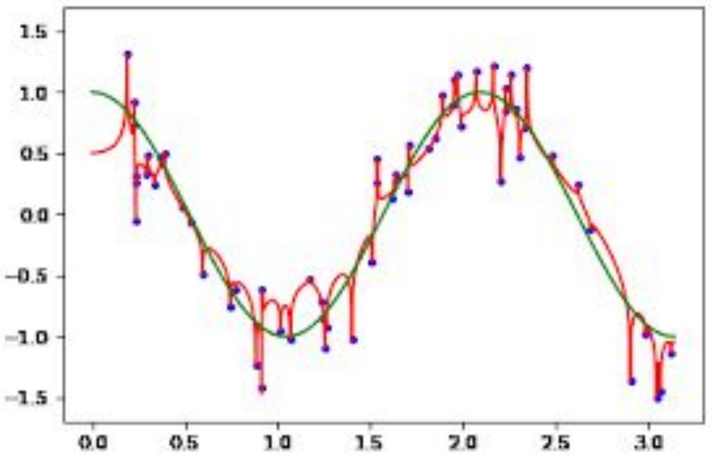
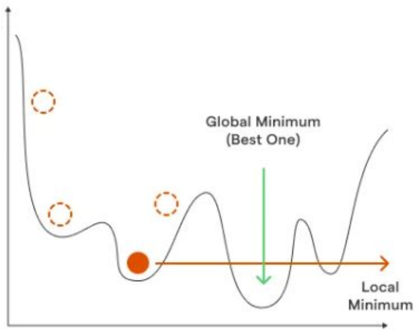
TPV: Parameter Perturbations Through the Lens of Test Prediction Variance

Devansh Arpit [1]

ICML 2026

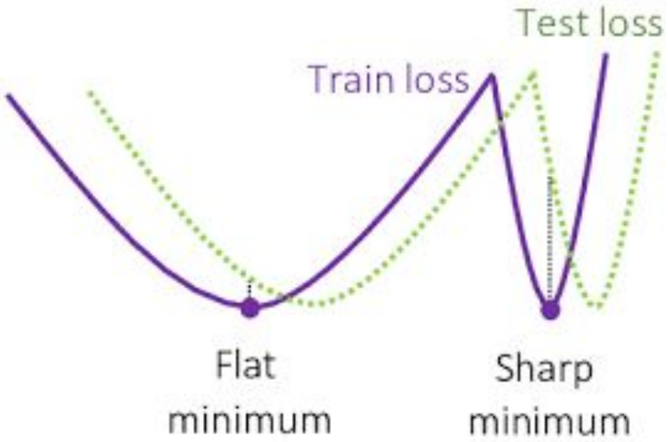


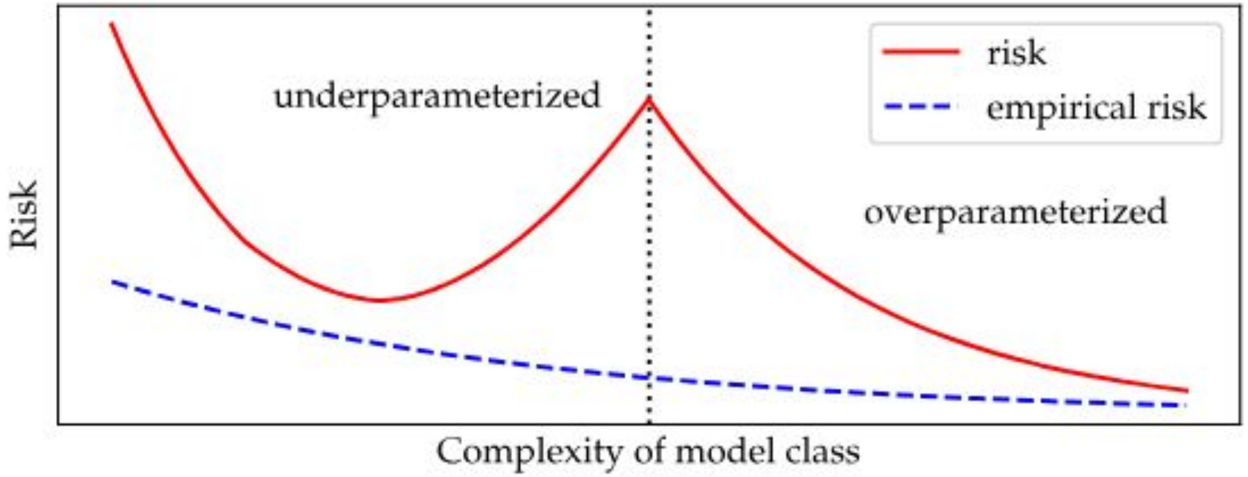
Benign Overfitting
under Label Noise

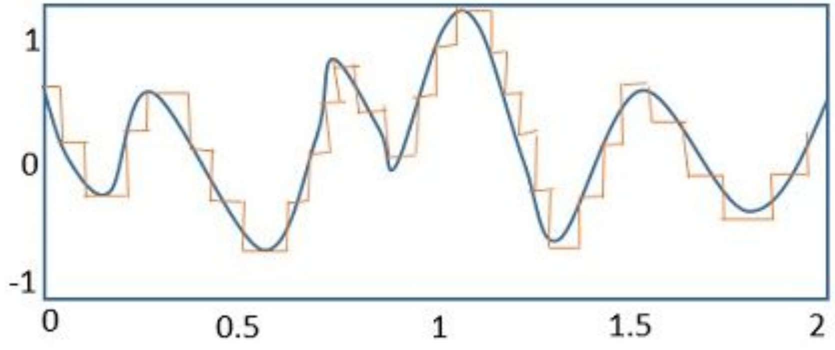


SGD Noise

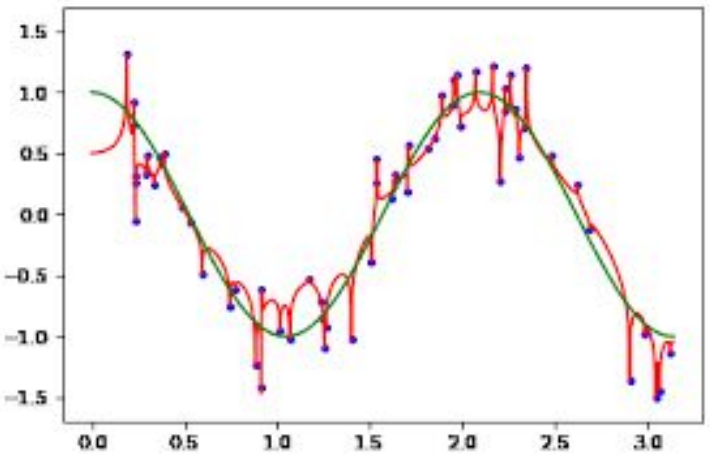
Scattered Views of Generalization & Robustness in Deep Learning



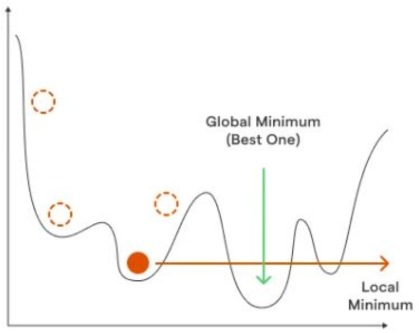




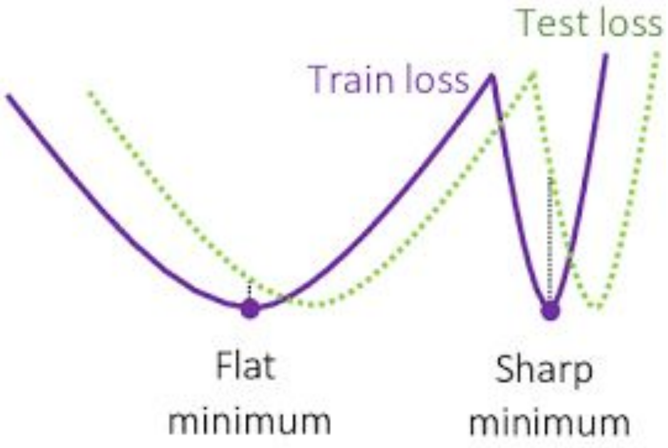
Quantization Noise



Benign Overfitting
under Label Noise

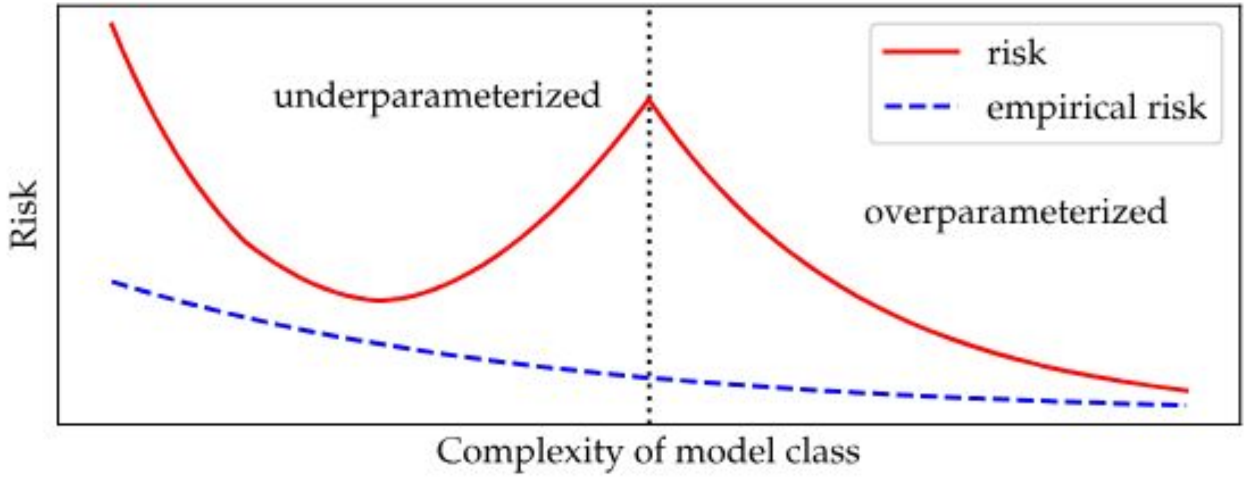


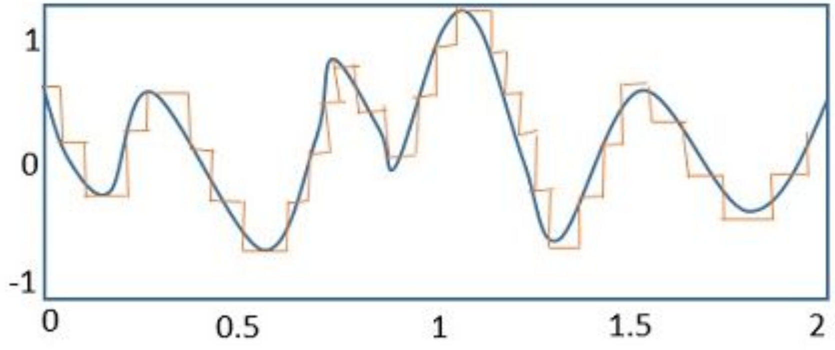
SGD Noise



Motivation

- Unified framework for studying robustness of a trained model
- Model selection using training-set

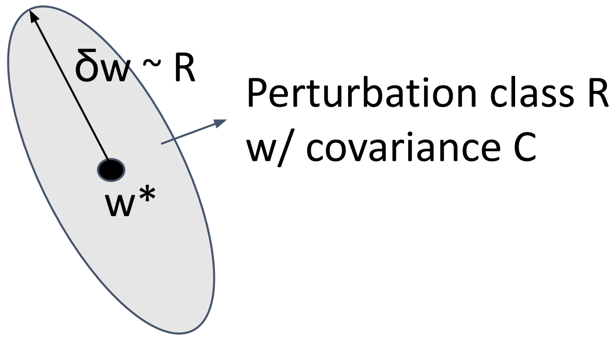




Quantization Noise

Test Prediction Variance (TPV)

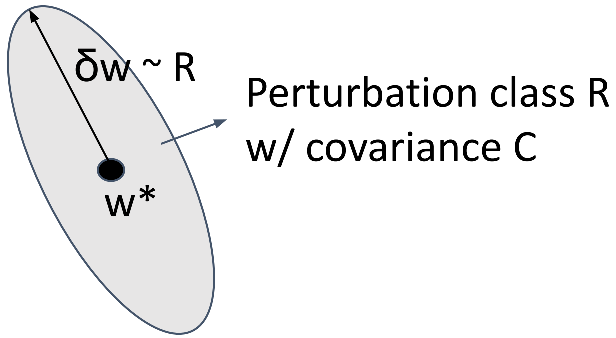
$$\mathrm{TPV} := \mathbb{E}_{x,\,\delta w}\left[\left\|f_{w^\star+\delta w}(x) - f_{w^\star}(x)\right\|^2\right]$$



TPV measures variance of predictions under perturbation class R

Test Prediction Variance (TPV)

$$\mathrm{TPV} := \mathbb{E}_{x,\,\delta w}\left[\left\|f_{w^\star+\delta w}(x) - f_{w^\star}(x)\right\|^2\right]$$



TPV measures variance of predictions under perturbation class R

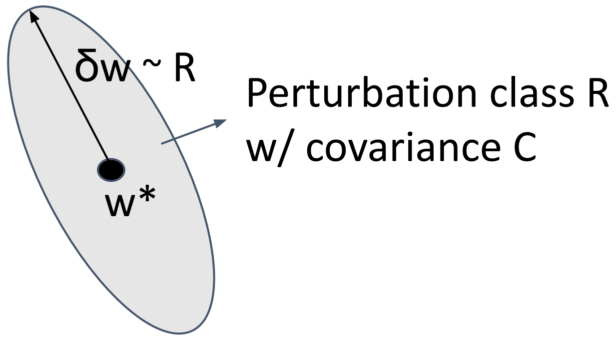
Many natural sources of randomness R in real-world:

- Fine-tuning on noisy label data
- Fine-tuning on data with covariate shift
- Quantization noise
- SGD noise at convergence
- Pruning noise

Each source induces parameter perturbations around w* directly or indirectly

Test Prediction Variance (TPV)

$$\mathrm{TPV} := \mathbb{E}_{x,\,\delta w}\left[\left\|f_{w^\star+\delta w}(x) - f_{w^\star}(x)\right\|^2\right]$$



Bias-Variance Decomposition

$$\mathcal{E}_{\text{test}} = \mathbb{E}_{x,y,R}\left[\left(f_{w^\star+\delta w}(x) - y\right)^2\right]$$

$$\mathcal{E}_{\text{test}} = \mathrm{TPV} + \mathbb{E}_x\big[\underbrace{\left(f_{w^\star}(x) - f^\star(x)\right)^2}_{\text{bias}^2}\big].$$

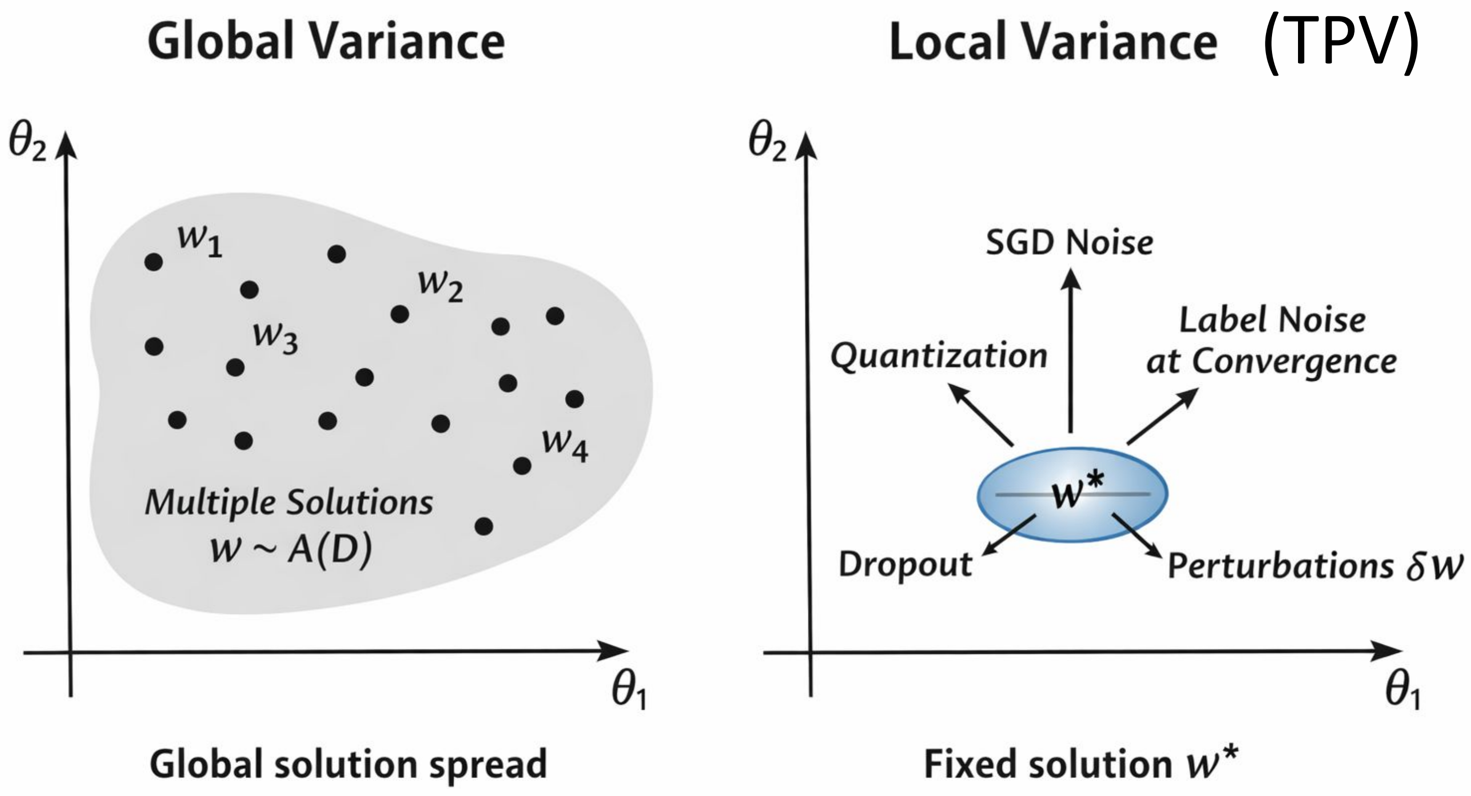
Local vs Global Bias-Variance Decomposition

Global Decomposition

- provides insight into properties of the learning algorithm

Local Decomposition

- provides insight into robustness of the trained model



TPV is Local Variance

Why is TPV Useful?

- Unifying view:
  - Overparameterization -> low TPV under label noise (benign overfitting)
  - Wide minima -> low TPV under SGD/quantization noise
  - Other noise sources can also be studied under TPV framework

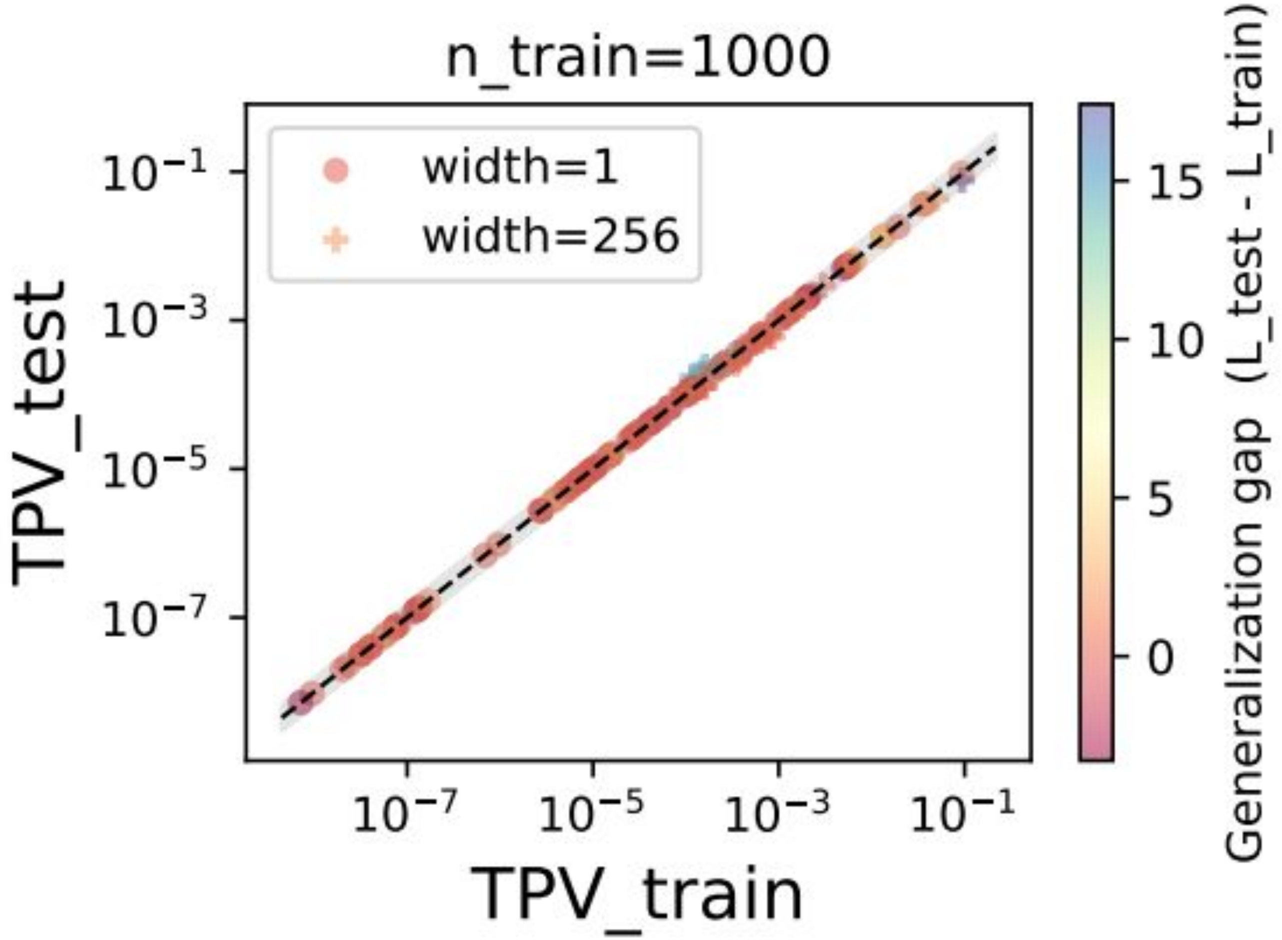
- Two properties of TPV:
  - TPV stability
    - $\mathrm{TPV}(w^*;X_{tr}) \approx \mathrm{TPV}(w^*;X_{te})$
  - TPV and Generalization
    - TPV correlated with test loss in low training loss regime

- Applications (due to above 2 properties):
  - Model/Training-recipe selection using training set
  - Label-free pruning algorithm

TPV Stability => TPV(w*;X_{tr}) ≈ TPV(w*;X_{te})

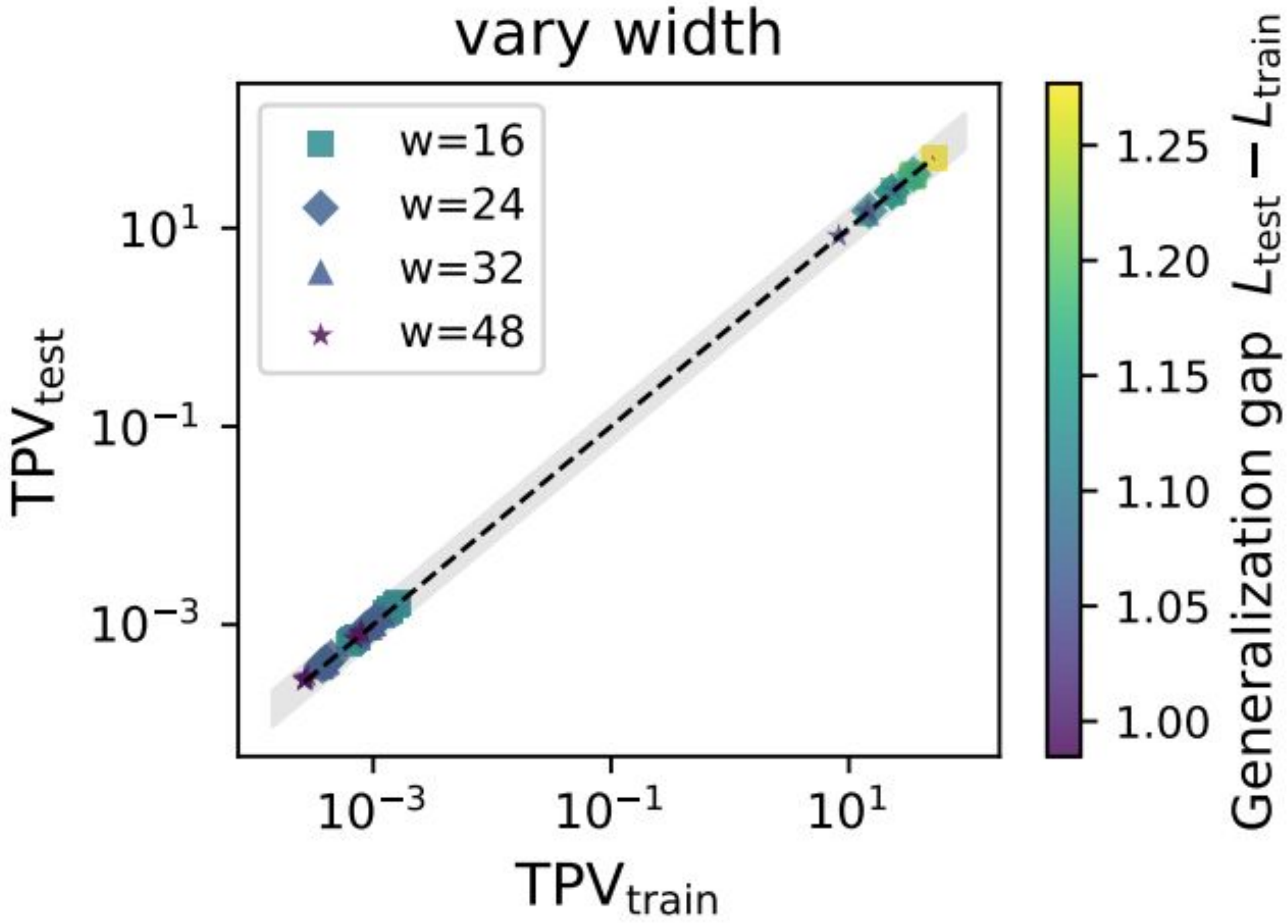
TPV stability (synthetic data)

- broad grid of:
  - architectures (depth/width),
  - data distributions/dimensions
  - perturbation sources
  - 324 configurations each w/ 50 runs



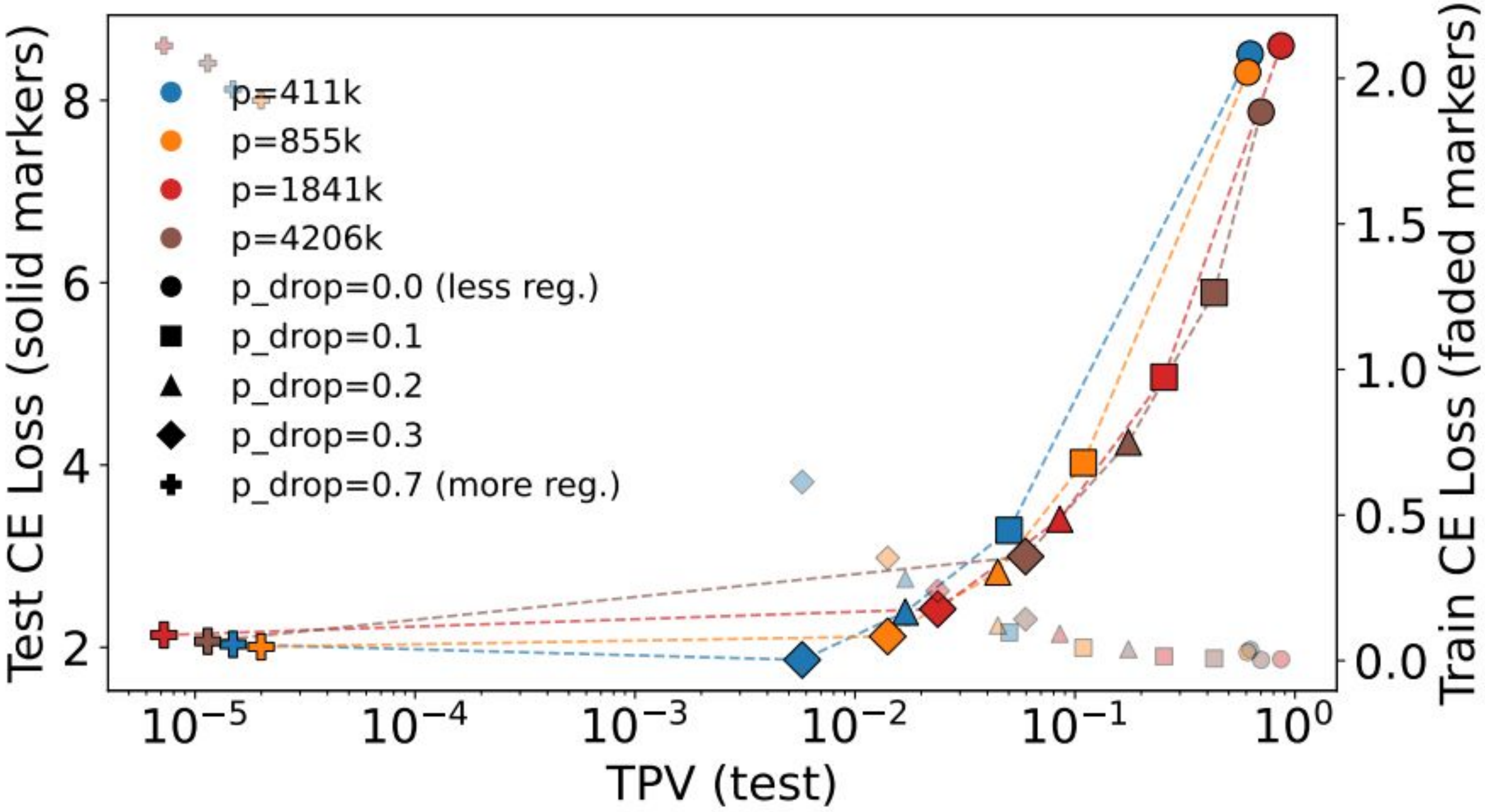
TPV Stability => TPV(w*;X_{tr}) ≈ TPV(w*;X_{te})

TPV stability (real data)



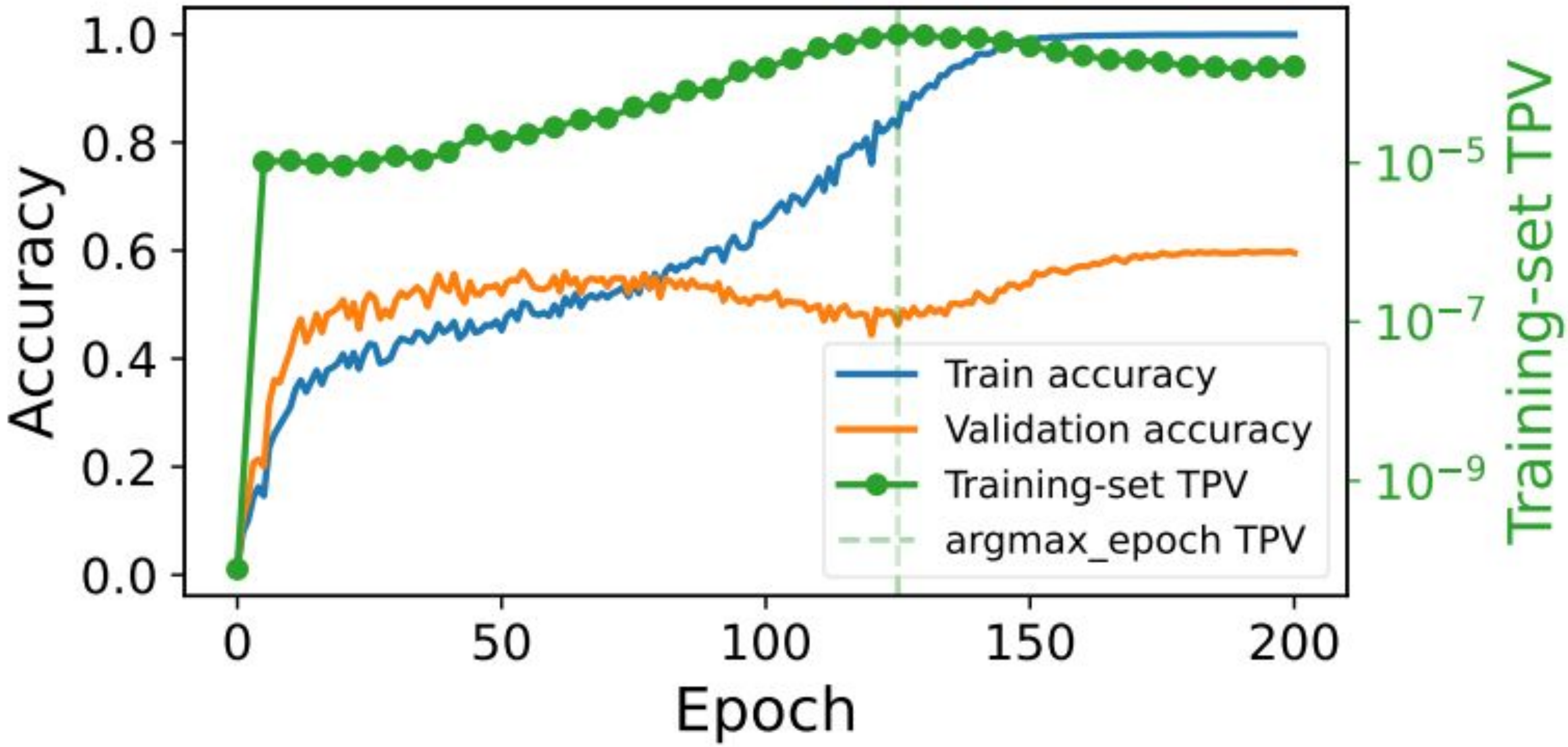
TPV & Generalization

Across Models



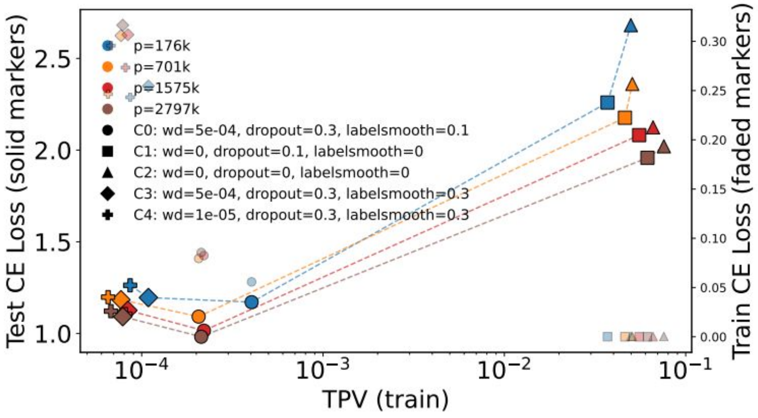
TPV & Generalization

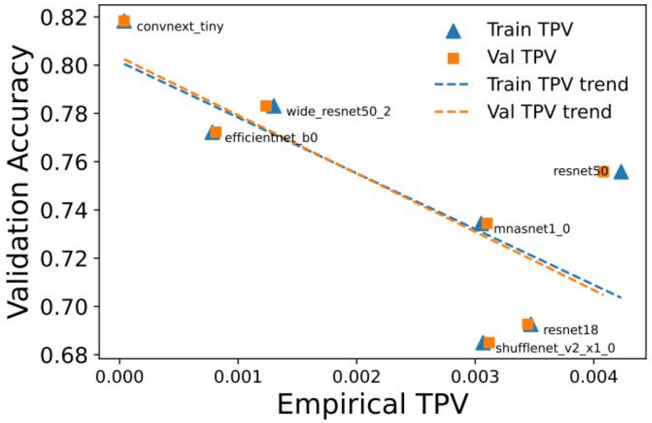
Across Epochs

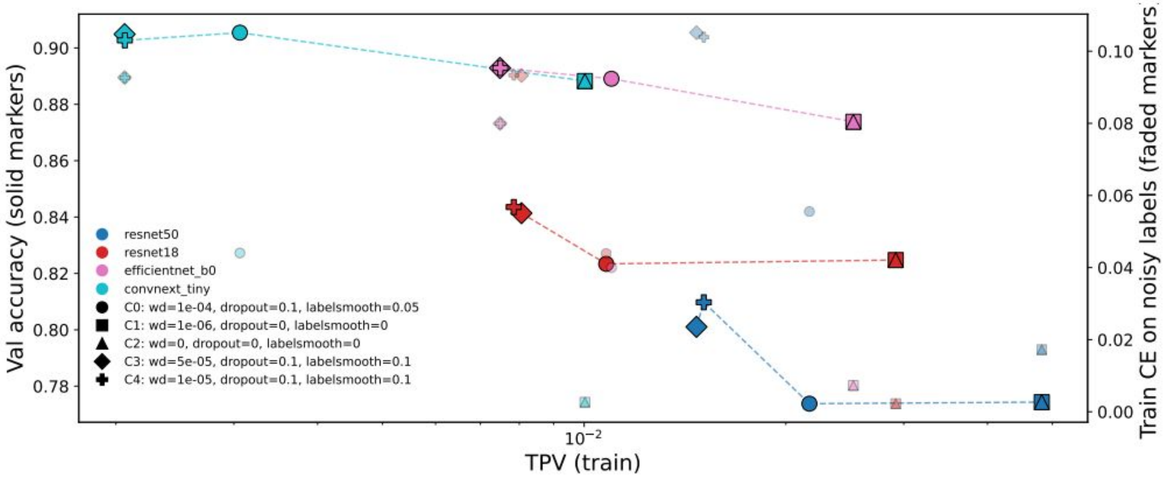


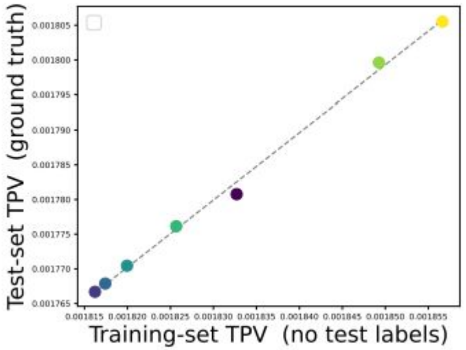
Applications of TPV

Model selection in in-distribution and transfer learning settings









Summary

- TPV provides a unifying theoretical framework for studying the robustness of trained models

- TPV stability and correlation with test loss enable training-set based model selection