Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

Lecture 6 - Covariate Shift

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Training ≠ Testing

- **Generalization performance**
  (the empirical distribution lies)
- **Covariate shift**
  (the covariate distribution lies)
- **Adversarial data**
  (the support of the distribution lies)
- **Label shift**
  (the label distribution lies)

$$p_{\mathrm{emp}}(x, y) \neq p(x, y)$$

$$p(x) \neq q(x)$$

$$\mathrm{supp}(p) \neq \mathrm{supp}(q)$$

$$p(y) \neq q(y)$$

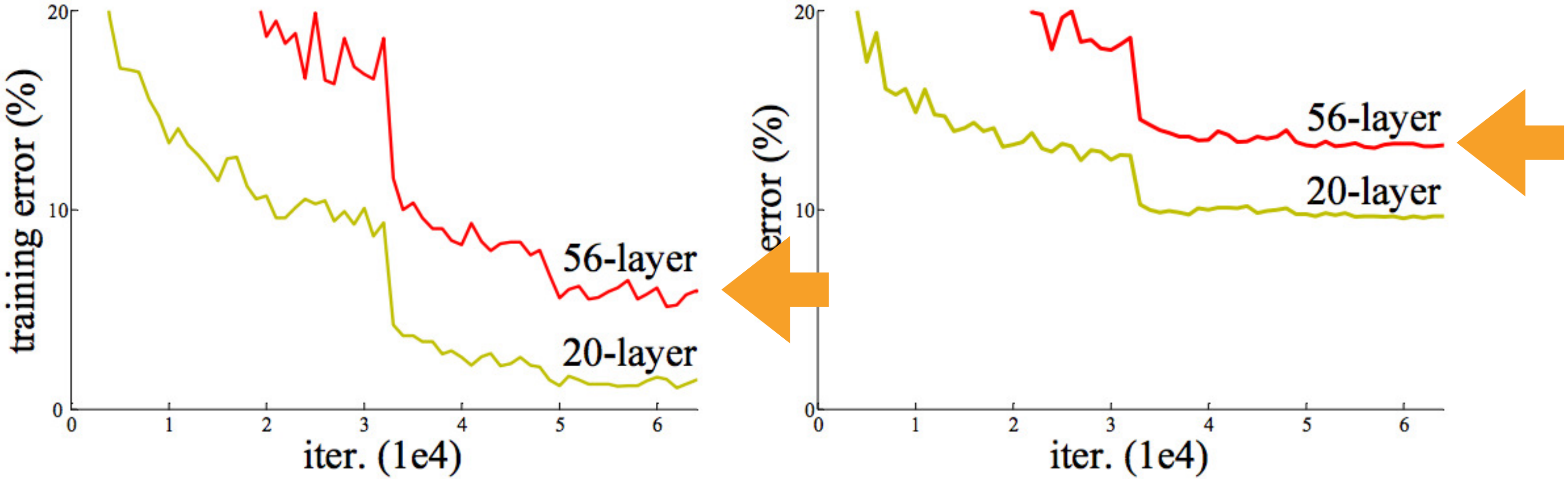
Recap - Generalization performance

Generalization performance

Generalization performance

Only cats and dogs?

- Images, too (e.g. He et al., 2015, ResNet paper)



- Alexa
  (‘Please turn off the coffee machine’ vs. ‘coffee machine off’)

Why?

- Data Distribution $p(x, y)$
- Dataset drawn from $p(x, y)$
- Training minimizes empirical risk (plus regularization)

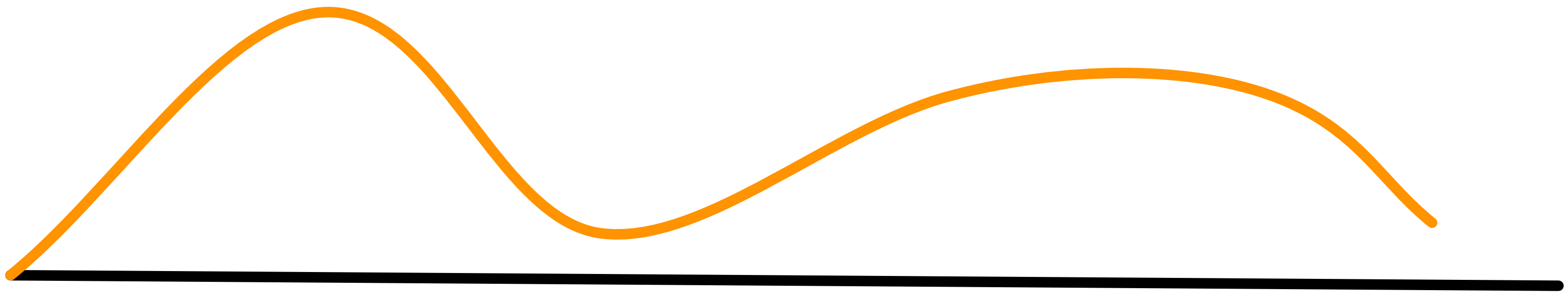
$$\underset{w}{\text{minimize}} \frac{1}{m} \sum_{I=1}^{m} l(f(x_i, w), y_i)$$

- At test time expected risk matters
  (all the other data we could have seen)

$$\mathbf{E}_{(x,y)\sim p} \left[l(f(x, w), y)\right]$$

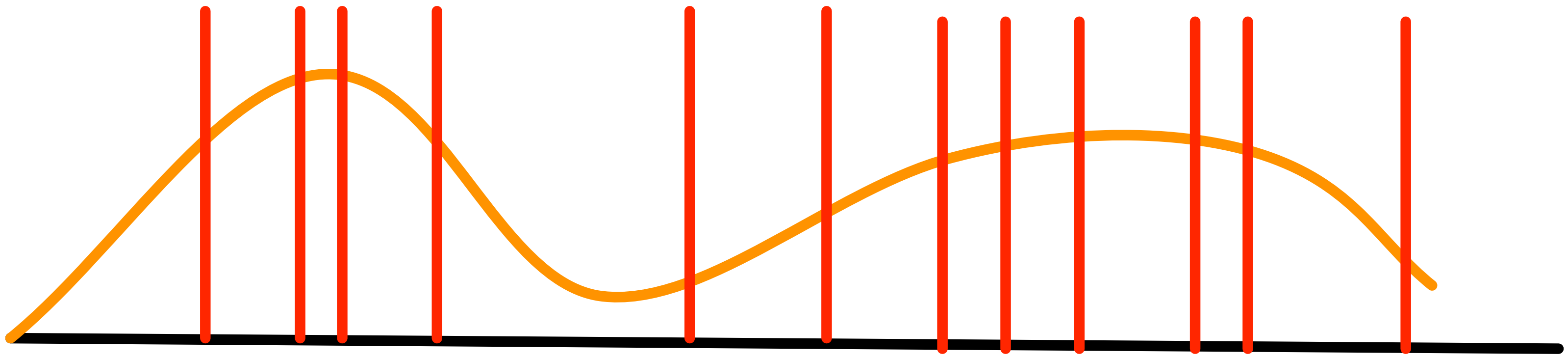
Why

Data Distribution



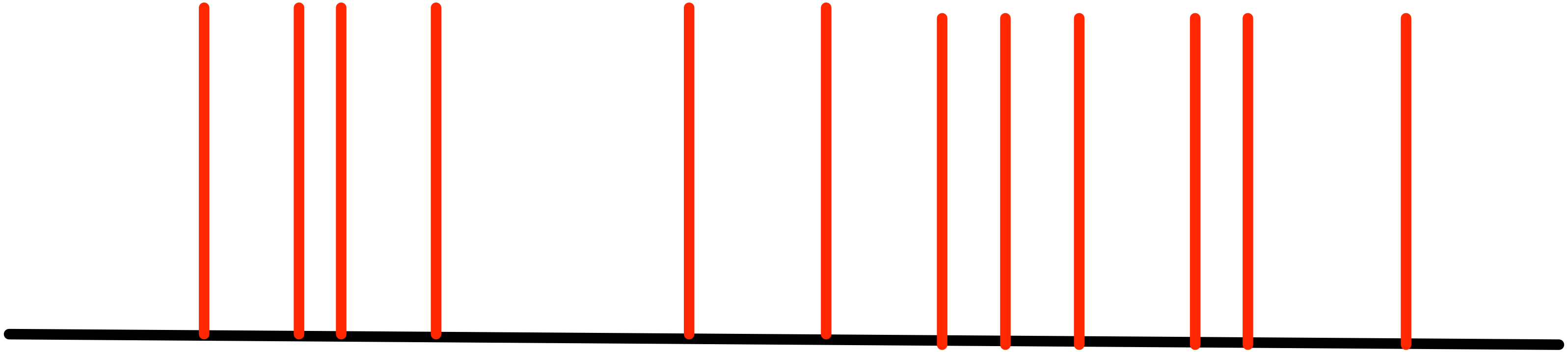
Why

Data Distribution with Empirical Sample



Why

Empirical Sample



Fixing it

- **Validation set**
  (hold out **separate** data that is not used for training)
- **Chernoff bound**

$$\Pr\left\{\frac{1}{m}\sum_{I=1}^{m} l(f(x_i),y_i)) - \mathbf{E}\left[l(f(x),y)\right] > \epsilon\right\} \leq \exp\left(-2m\epsilon^2\right)$$

- **Why does it work?**
  - Validation set was never used for training (often violated)
  - Loss bounded within [0,1] (otherwise rescale)

Fixing it

- **Validation set**
  (hold out **separate** data that is not used for training)
- **Chernoff bound**

$$\Pr\left\{\frac{1}{m}\sum_{I=1}^{m} l(f(x_i), y_i)) - \mathbf{E}\left[l(f(x), y)\right] > \epsilon\right\} \leq \exp\left(-2m\epsilon^2\right)$$

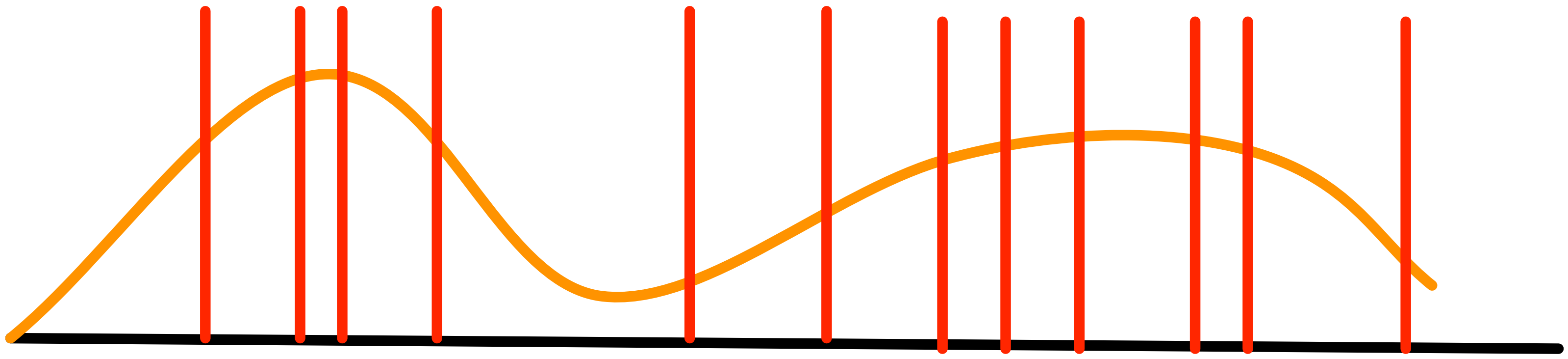
- Solving yields that with probability at least $1 - \delta$

$$R[f, p] \leq R_{\text{emp}}[f, X, Y] + \sqrt{-\frac{\log \delta}{2m}}$$

For $\delta = 0.05$ and $\epsilon = 0.01$ we have $m = 15{,}000$

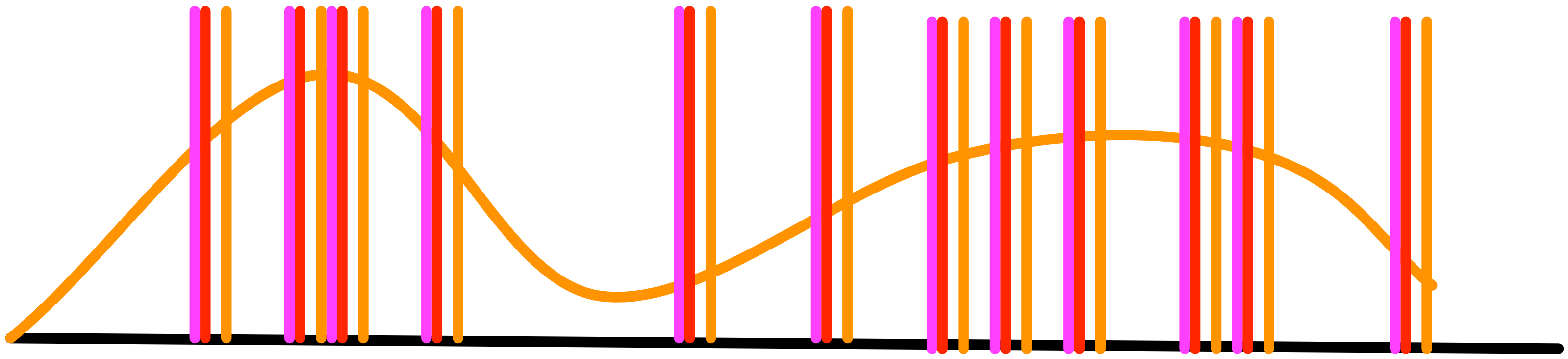
Fixing it

Data Distribution with Empirical Sample



Fixing it

- Input noise (more on this later)
- Dropout (noise within the layers)
- Smoothing f (weight decay or other regularization)



Key Takeaways

Training minimizes $R_{\text{emp}}[f, X, Y] := \frac{1}{m} \sum_{i=1}^{m} l(y_i, f(x_i))$

At test time we want to minimize

- Expected risk (data drawn from some distribution)
- Test error, if we have a specific set $\{x'_1, \ldots, x'_{m'}\}$

Good performance on training set doesn't guarantee good test performance, unless we regularize capacity or have independent validation set for calibration.

Training ≠ Testing

- **Generalization performance**
  (the empirical distribution lies) $$p_{\text{emp}}(x, y) \neq p(x, y)$$
- **Covariate shift**
  (the covariate distribution lies) $$p(x) \neq q(x)$$
- **Adversarial data**
  (the support of the distribution lies) $$\text{supp}(p) \neq \text{supp}(q)$$
- **Label shift**
  (the label distribution lies) $$p(y) \neq q(y)$$