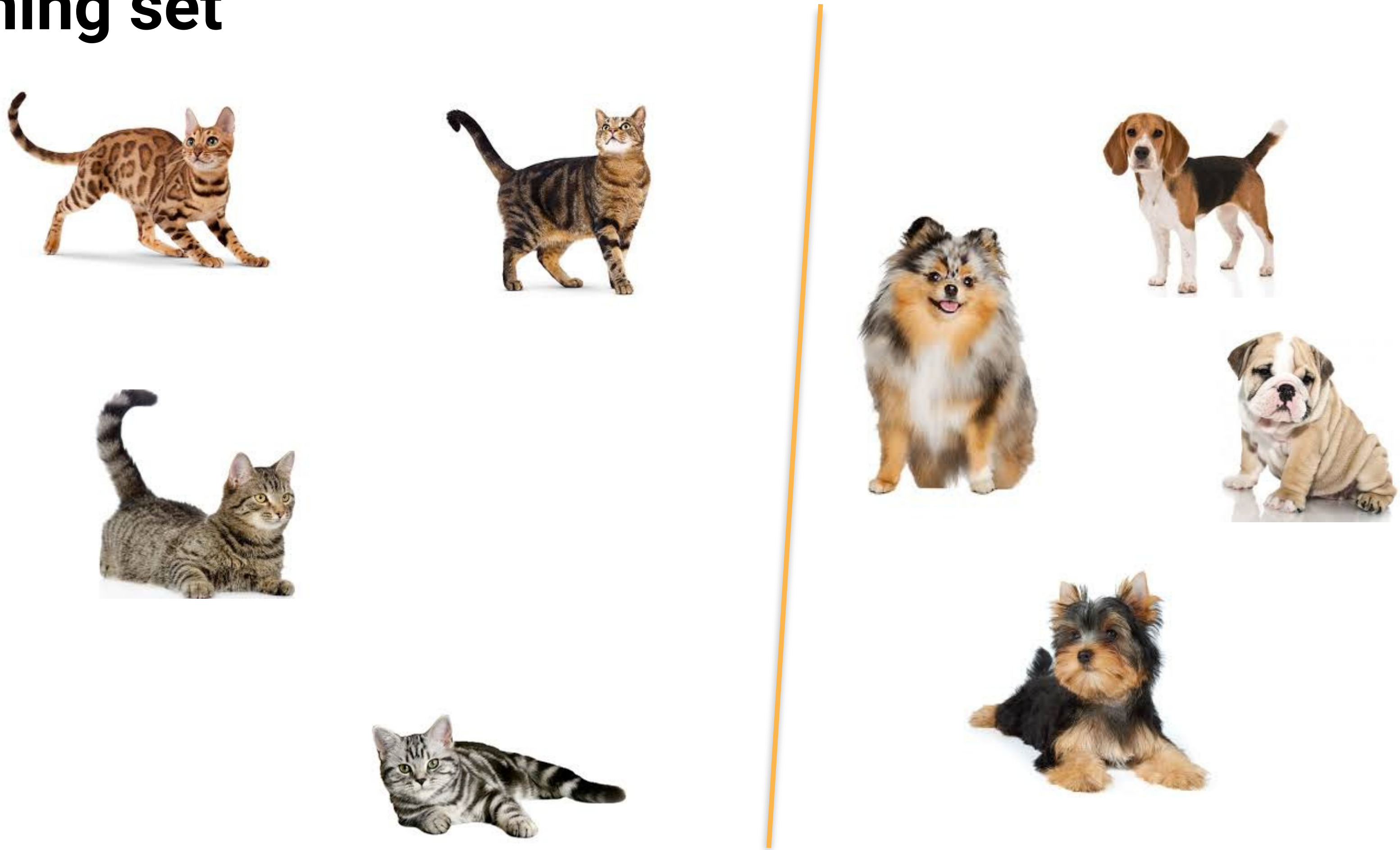
Label shift

Training set



Test set

Why would anyone be so stupid?

Label Shift $q(x,y) = q(y)p(x|y)$

- **Medical diagnosis**
  - Train on data with few sick patients (in CA)
  - Test in South Dakota after Sturgis Biker Rally when $q(\mathrm{C19}) > p(\mathrm{C19})$ while COVID19 symptoms $p(\mathrm{symptoms} \mid \mathrm{C19})$ are still the same.
- **Speech recognition**
  - Train on newscast data before election
  - Test on newscast after election (new topics, names, discussions, but still same language)

Label Shift Correction when we know $p(y)$ and $q(y)$

- Given trained model with $p(y \mid x)$
- We want $q(y \mid x)$

$$q(y \mid x) = \underbrace{\frac{q(x \mid y)q(y)}{q(x)}}_{\text{Bayes Rule}} = \frac{\overbrace{p(x \mid y)}^{= q(x \mid y)}p(y)}{p(x)} \cdot \frac{q(y)}{p(y)} \cdot \underbrace{\frac{p(x)}{q(x)}}_{\text{Drop this}} \propto p(y \mid x)\frac{q(y)}{p(y)}$$

- TL;DR - When you have $q(y)$, fixing models is easy.

Label Shift Correction when we know $p(y \mid x)$ and $q(y)$

- Given trained model with $p(y \mid x)$
- We want $q(y \mid x)$

$$q(y \mid x) = \frac{q(x \mid y)q(y)}{q(x)} = \frac{p(x \mid y)p(y)}{p(x)} \cdot \frac{q(y)}{p(y)} \cdot \frac{p(x)}{q(x)} \propto p(y \mid x)\frac{q(y)}{p(y)}$$

- Train original model on p(x,y)
- Reweight estimates via $q(y \mid x) \propto p(y \mid x)\frac{q(y)}{p(y)}$
- Renormalize

Label Shift

$$q(x, y) = q(y)p(x|y)$$

- Data generating process p(x|y) is unchanged
- Labels change since the underlying cause changed
- Need to reweight according to $\beta(y) = \dfrac{q(y)}{p(y)}$ to get

$$\int dq(x, y) l(f(x), y) =$$

$$\int dq(y) \int dp(x \mid y) l(f(x), y) =$$

$$\int dp(y) \frac{q(y)}{p(y)} \int dp(x \mid y) l(f(x), y) = \int dp(x, y) \frac{q(y)}{p(y)} l(f(x), y)$$

We don’t have samples from q(y)!

Label Shift

$$q(x, y) = q(y)p(x|y)$$

- Data generating process p(x|y) is unchanged
- Labels change since the underlying cause changed
- Need to reweight according to $\beta(y) = \frac{q(y)}{p(y)}$ to get

$$\int dq(x, y) l(f(x), y) =$$

$$\int dq(y) \int dp(x \mid y) l(f(x), y) =$$

$$\int dp(y) \frac{q(y)}{p(y)} \int dp(x \mid y) l(f(x), y) = \int dp(x, y) \frac{q(y)}{p(y)} l(f(x), y)$$

We don't have samples from q(y)!

Label Shift

$$q(x, y) = q(y)p(x|y)$$

- **Key Idea - measure the estimates on test set**
  - p(x|y) is the same for training and test
  - Use distribution of predictions per label (error confusion matrix) is

$$p(\hat{y}, y) = \int \hat{p}(\hat{y} \mid x)p(x \mid y)p(y)dx$$

  - Match distribution of predictions on training and test set.
- **Spectral algorithm** (linear equation, Lipton et al. 2018)

$$q(\hat{y}) = \int \hat{p}(\hat{y} \mid x)q(x)dx = \sum_y p(\hat{y}, y)\beta_y$$

Simple Algorithm

$C = 0$ and $q = 0$

for $i = 1$ to m do (training set)

$\quad C[:, y[i]] += p(: | x[i])$

for $i = 1$ to m' do (test set)

$\quad q += p(y | x'[i])$

$b = C^{-1} q$

Better solution

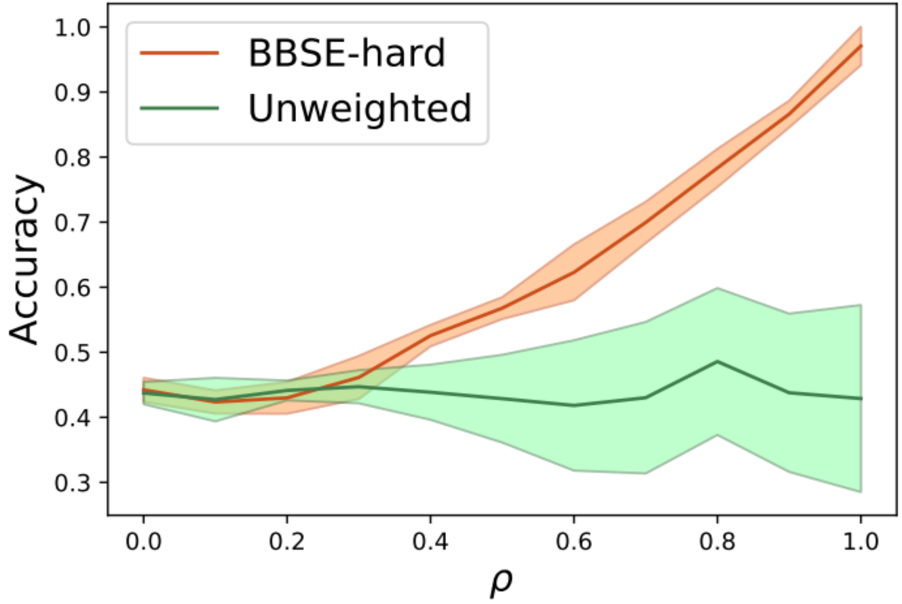
$$\underset{b}{\text{minimize}} \quad \|q - Cb\|^2$$

$$\text{subject to} \quad b[y] \geq 0 \quad \text{and} \quad \sum_y b[y]p[y] = 1$$

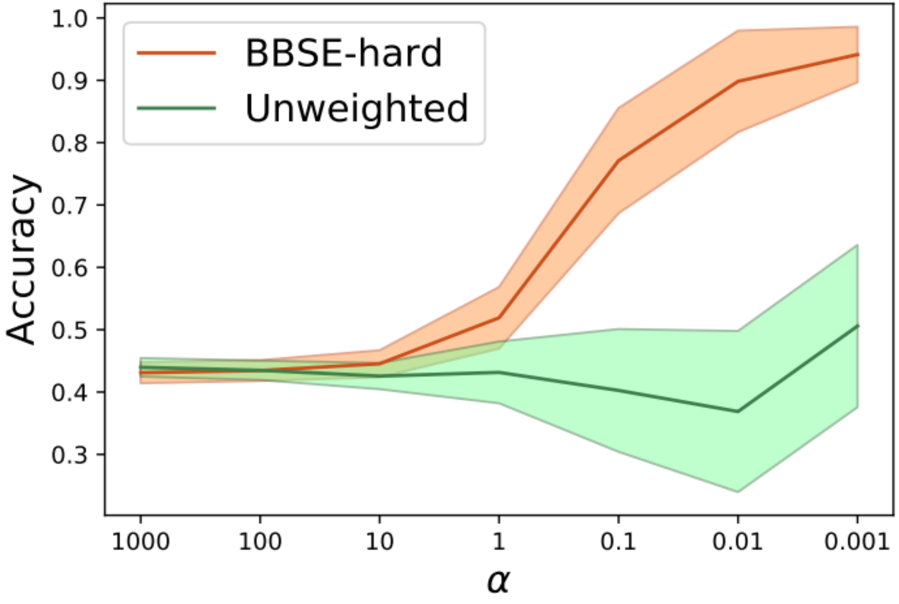
Guarantees

- **Robust under misspecification**
  - Even if the estimates $\hat{y}(x)$ are wrong, calibration is OK: (same errors on hold-out and test set)
  - Confusion matrix and label vector are concentrated: (use matrix Bernstein inequality)
- **Simple algorithm**
  - Cubic in number of classes, linear in sample size.

Black Box Shift Correction on CIFAR10



Tweaking one class probability



Dirichlet prior over shifts

Extensions

- **Streaming data**
  Estimate weights while observing data
  (e.g. via SGD on moment matching)
- **Large label sets**
  - Feature moment matching (via MMD)
  - GAN moment matching
  - Classifier of scores between training and test set
- **Better objective**
  Use KL-divergence to calibrate $q(y)$ against $\beta(y)p(y)$

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)

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

- **Two-Sample Tests**
  (distributions don't match)

$$p \neq q$$

- **Label shift**
  (the label distribution lies)

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

Things we didn't cover

- **Covariate Drift**
  - Things change slowly over time, e.g. language, user preferences, disease symptoms
  - Geographic preferences (Canada vs. USA search behavior, demographics)
  - Strategy: estimate $\frac{p(x,t)}{p(x,t')} \propto \exp(g(x,t) - g(x,t'))$ as a time-varying function.

Things we didn't cover

- **Concept Drift**
  - Dependency $p(y \mid x)$ changes slowly over time.
  - Train classifier $p(y \mid x, t)$ instead.
- **Concept Shift**
  Much bigger problem if concept shifts between training and test set. No real guarantees possible.