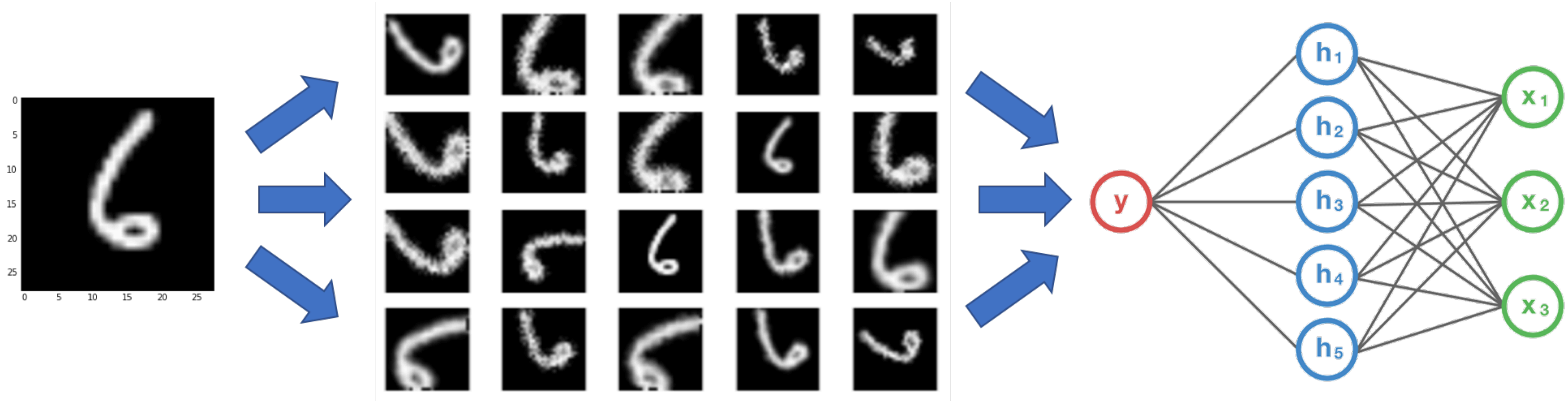
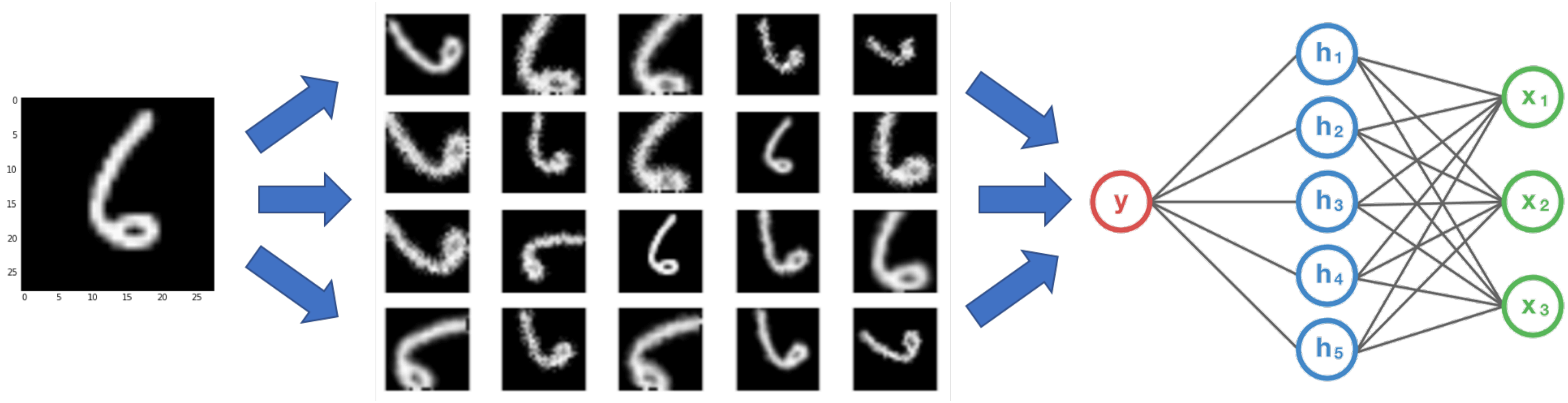
A Kernel Theory of Modern Data Augmentation

Tri Dao, Albert Gu, Alex Ratner, Virginia Smith, Chris De Sa, Chris Ré

Data augmentation is important to accuracy…

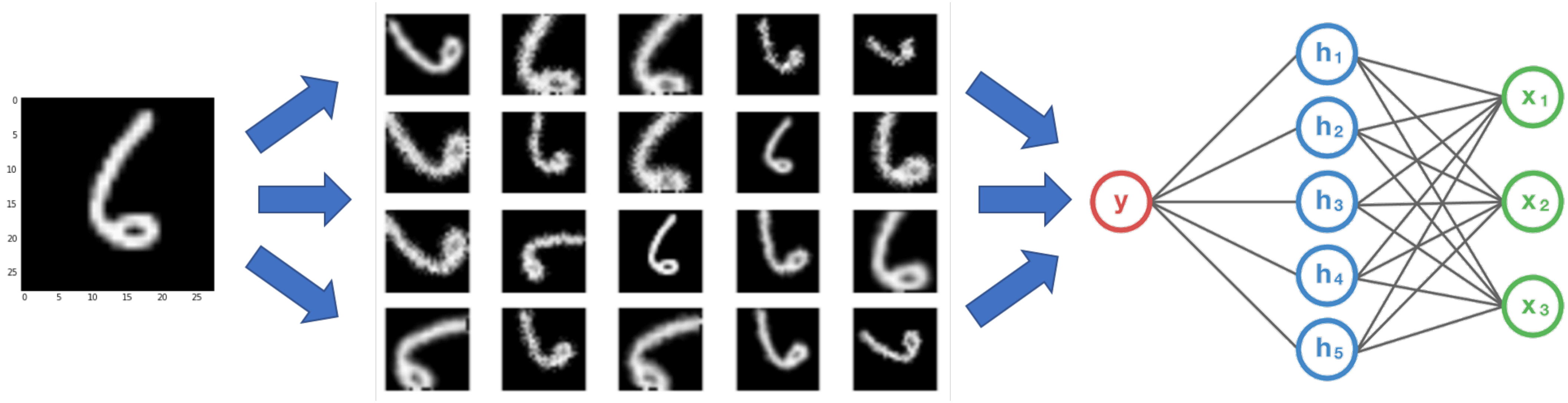


Data augmentation is important to accuracy…



3.7 pt. average gain across top ten CIFAR-10 models

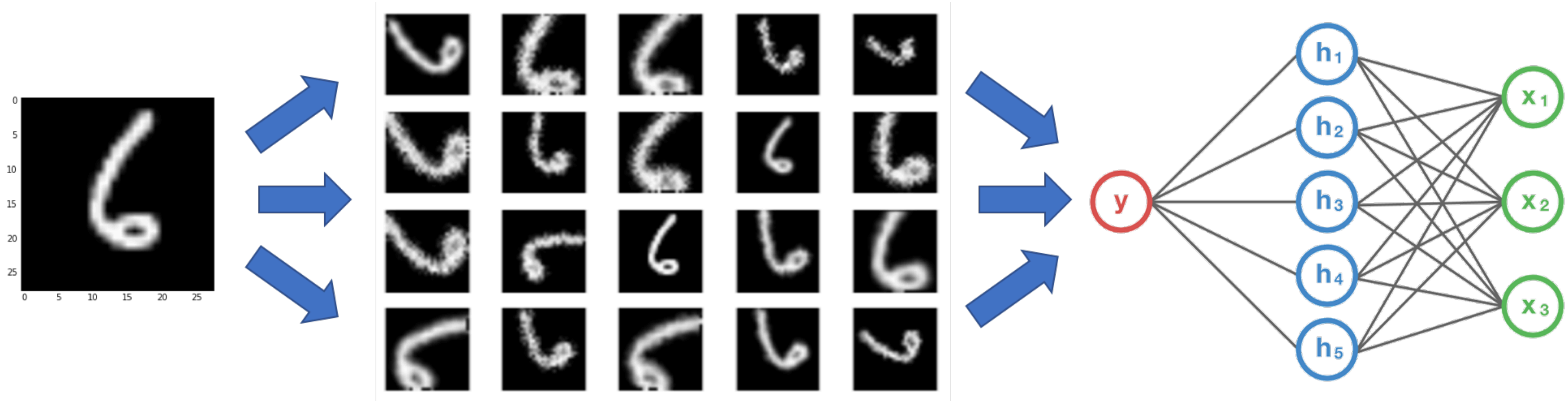
Data augmentation is important to accuracy…



3.7 pt. average gain across top ten CIFAR-10 models

13.9 pt. average gain for CIFAR-100

Data augmentation is important to accuracy…



3.7 pt. average gain across top ten CIFAR-10 models
13.9 pt. average gain for CIFAR-100

A form of weak supervision:
expresses domain knowledge (invariance)

… but is not well understood

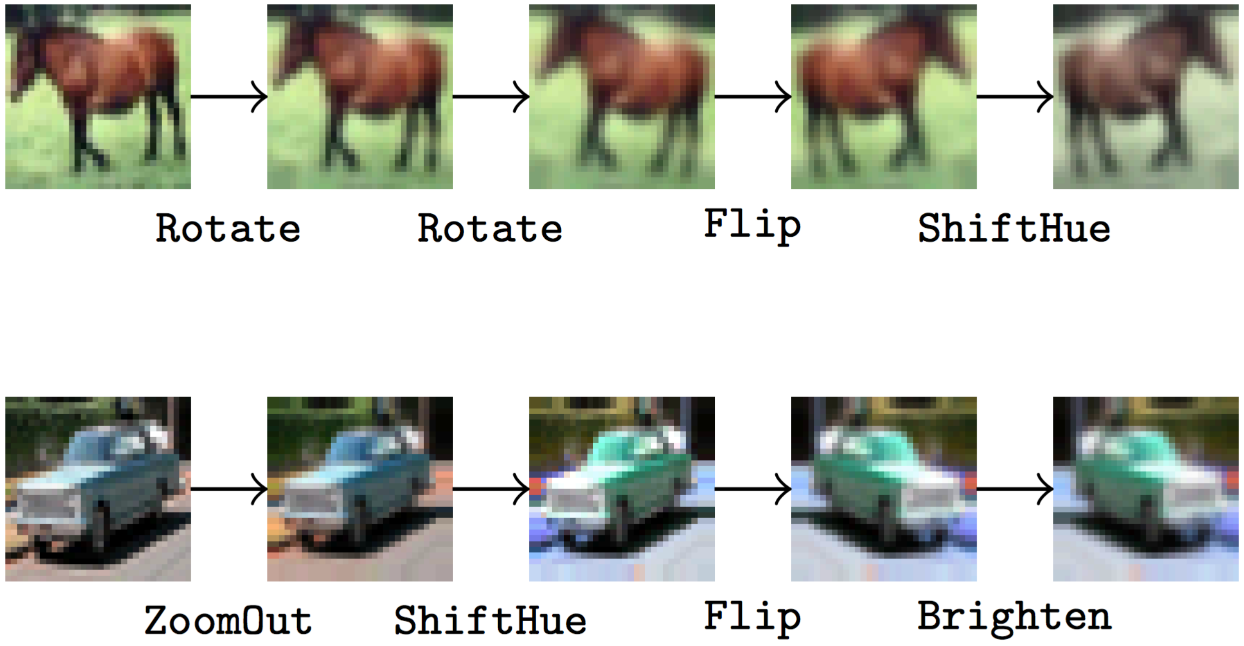
… but is not well understood

How does data augmentation affect the model?

- Learning process
- Parameters and decision surface

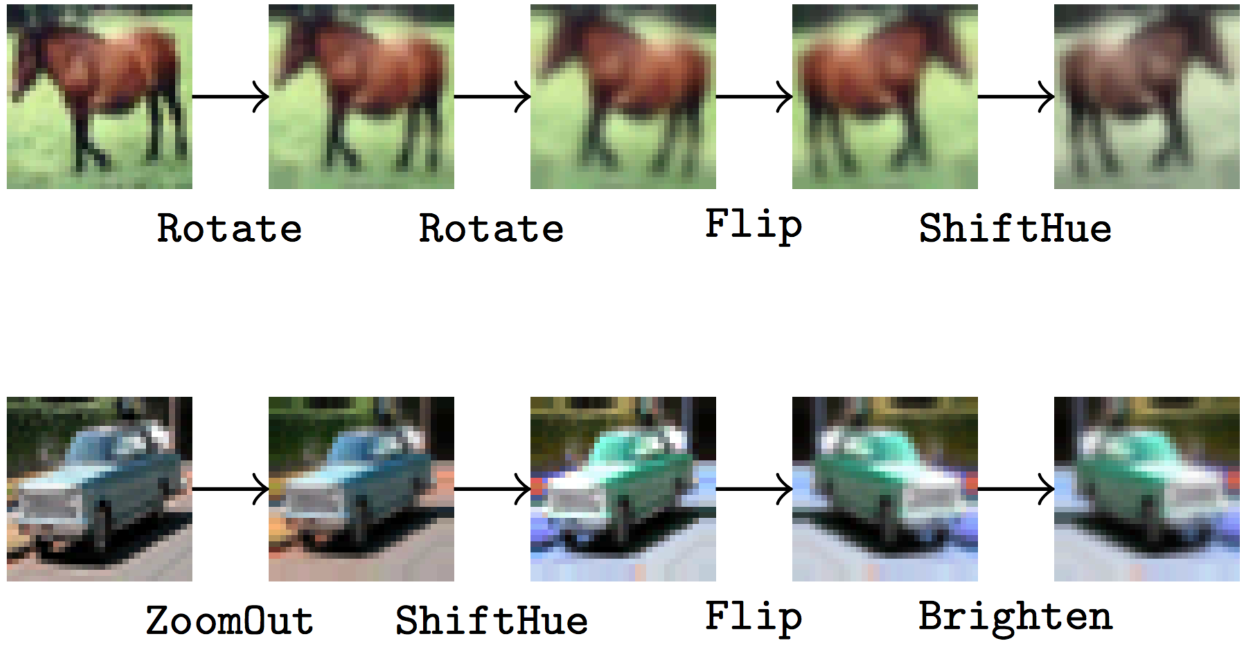
Augmentation as sequence modeling

- TANDA [Ratner et al., 2017]
- AutoAugment [Cubuk et al., 2018]



Augmentation as sequence modeling

- TANDA [Ratner et al., 2017]
- AutoAugment [Cubuk et al., 2018]



Model augmentation as a Markov chain

Augmentation as kernels

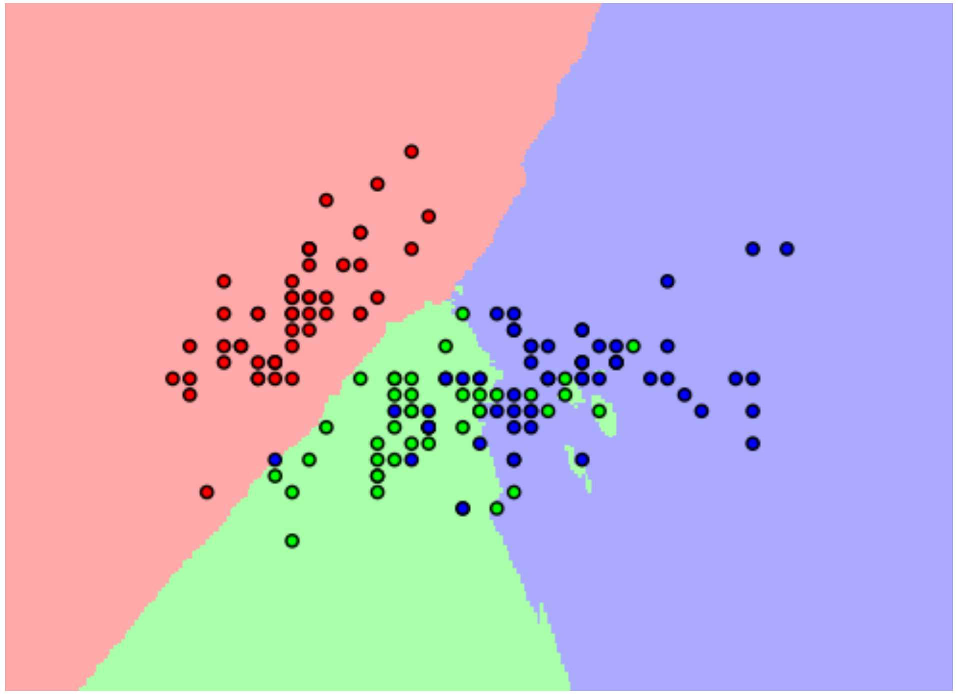
Base classifier: k-nearest neighbors

+

Data augmentation

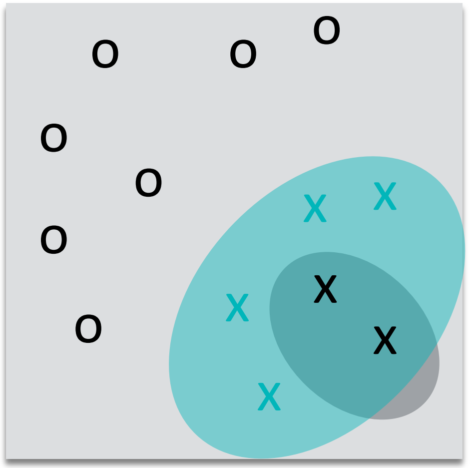
=

Asymptotic kernel classifier



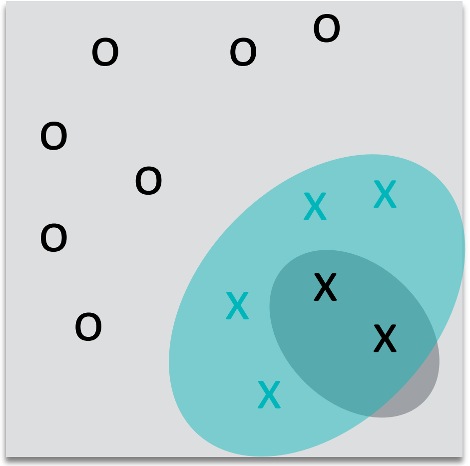
Effects of data augmentation on kernel classifiers

Effects of data augmentation on kernel classifiers

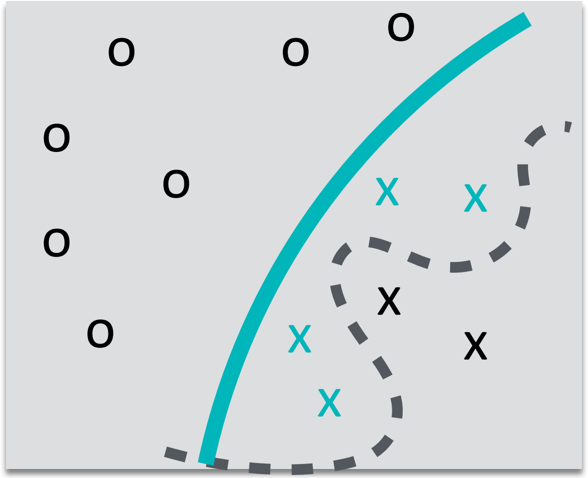


Invariance

Effects of data augmentation on kernel classifiers

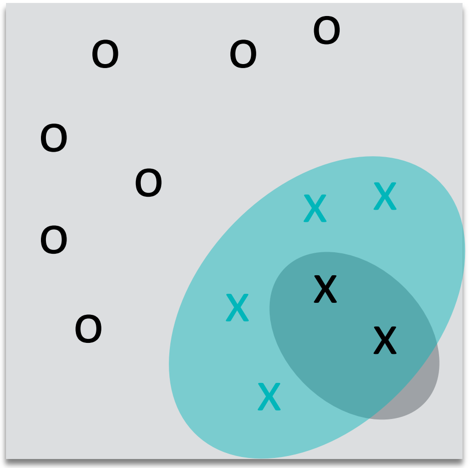


Invariance

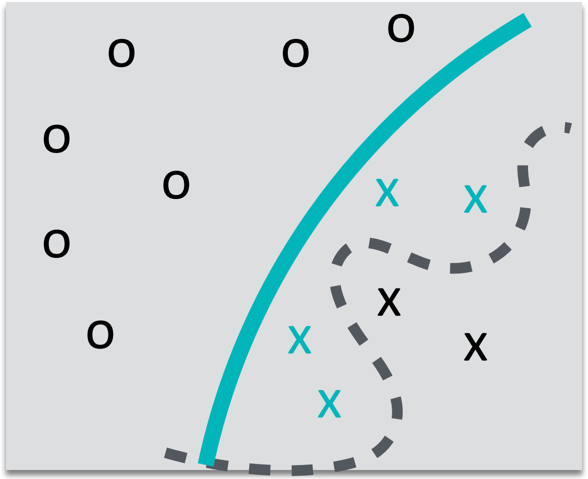


Regularization

Effects of data augmentation on kernel classifiers



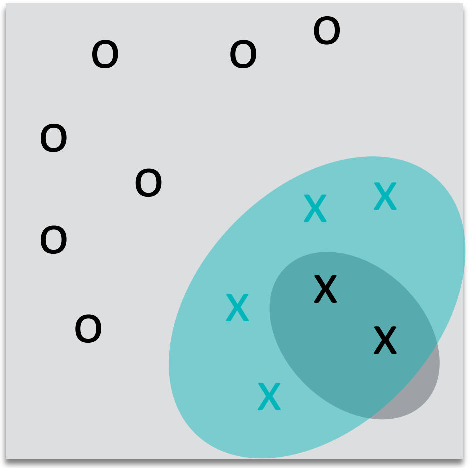
Invariance



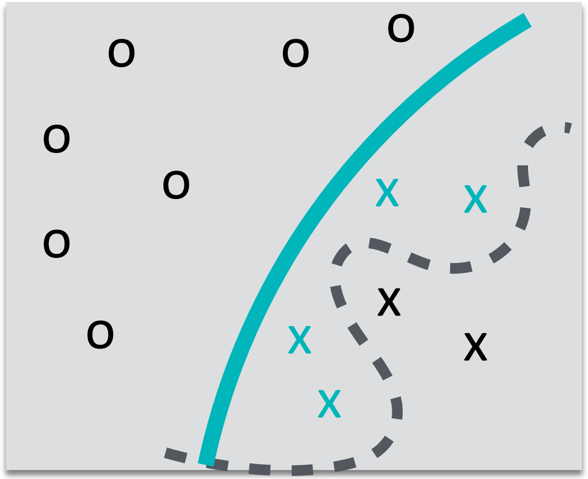
Regularization

Practical utility

Effects of data augmentation on kernel classifiers



Invariance

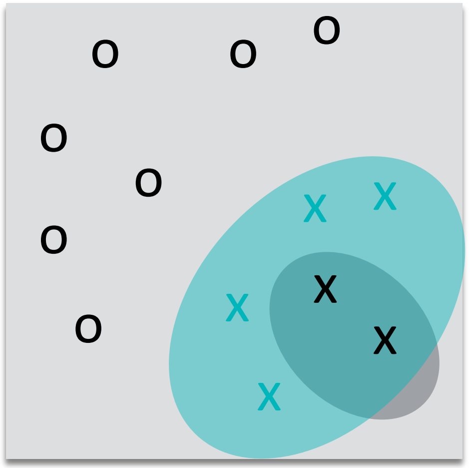


Regularization

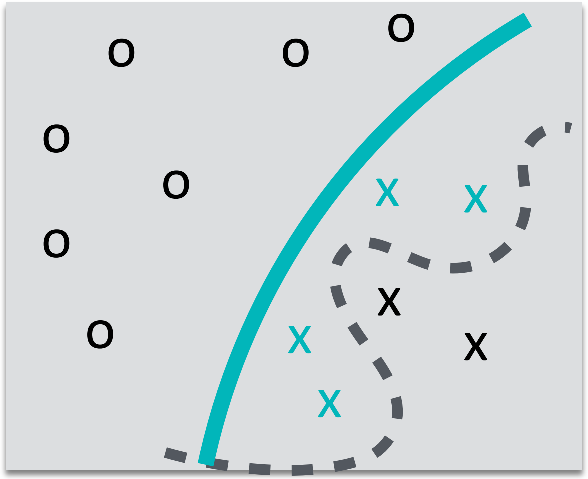
Practical utility

speeding up training

Effects of data augmentation on kernel classifiers



Invariance



Regularization

Practical utility

speeding up training

as a diagnostic

Model of data augmentation: kernel classifier

Non-augmented: $$\min_w \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top \phi(x_i))$$

Loss function

Feature map

Model of data augmentation: kernel classifier

Non-augmented: $$\min_w \frac{1}{n}\sum_{i=1}^{n} \ell(w^\top \phi(x_i))$$

Loss function

Feature map

Augmented: $$\min_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{z_i \sim T(x_i)} \ell(w^\top \phi(z_i))$$

Transformed versions of data point

Data augmentation effects

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i\sim T(x_i)}\ell(w^\top\phi(z_i)) \approx \frac{1}{n}\sum_{i=1}^{n}\ell(w^\top\mathbb{E}_{z_i\sim T(x_i)}\phi(z_i))$$

Average of augmented features
(i.e. kernel mean embedding)

Data augmentation effects

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i\sim T(x_i)}\ell(w^\top\phi(z_i)) \approx \frac{1}{n}\sum_{i=1}^{n}\ell(w^\top\mathbb{E}_{z_i\sim T(x_i)}\phi(z_i))$$

Average of augmented features
(i.e. kernel mean embedding)

1st order effect:
induces invariance
by feature
averaging

Data augmentation effects

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z_i\sim T(x_i)}\ell(w^\top\phi(z_i)) \approx \frac{1}{n}\sum_{i=1}^{n}\ell(w^\top\mathbb{E}_{z_i\sim T(x_i)}\phi(z_i))$$

Average of augmented features
(i.e. kernel mean embedding)

1st order effect: induces invariance by feature averaging

2nd order effect: reduces model complexity via a data-dependent regularization

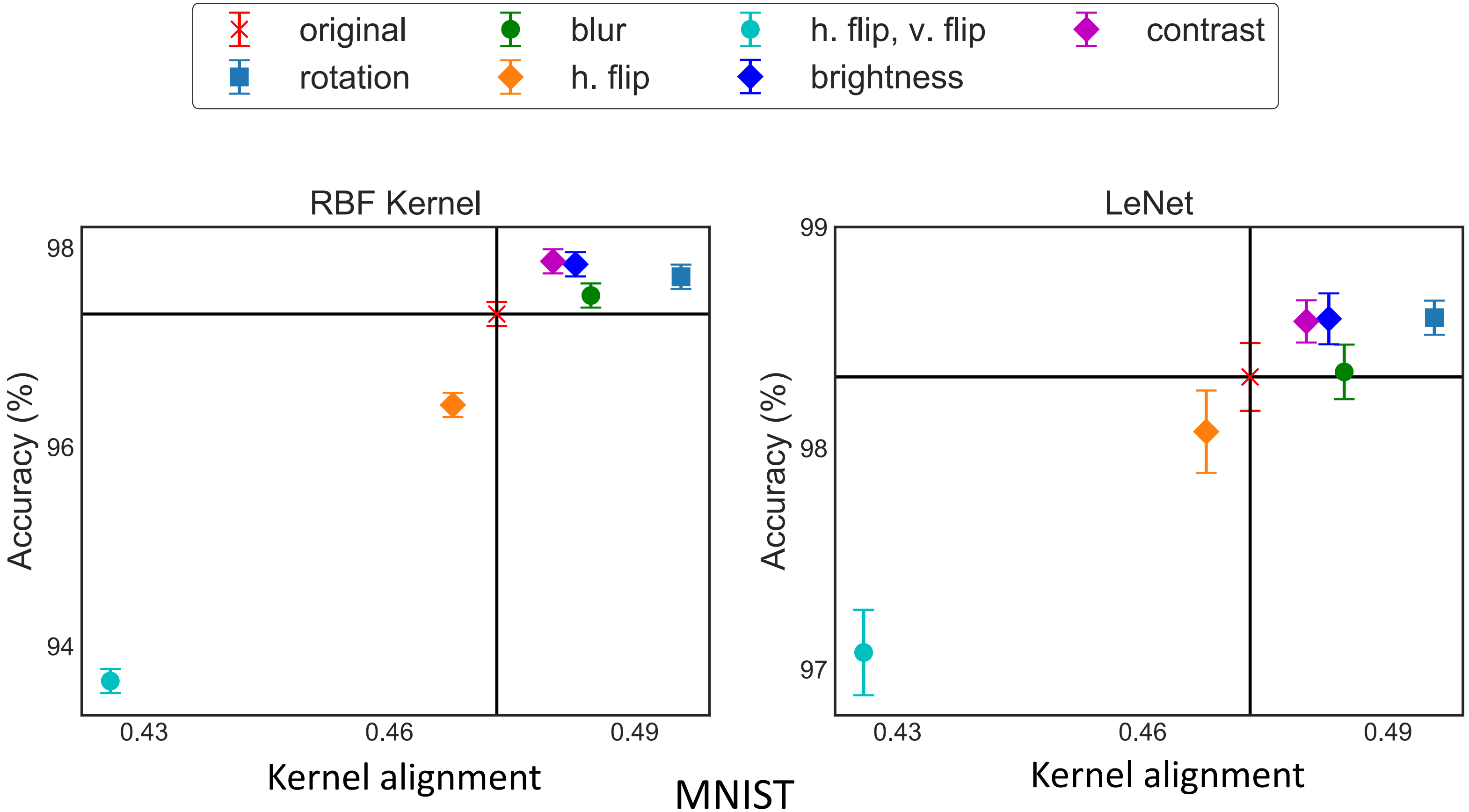
A diagnostic: kernel alignment metric

Averaged features: $\psi(x) = \mathbb{E}_{z \sim T(x)} \phi(z)$

Kernel target alignment [Cristianini et al., 2002]:

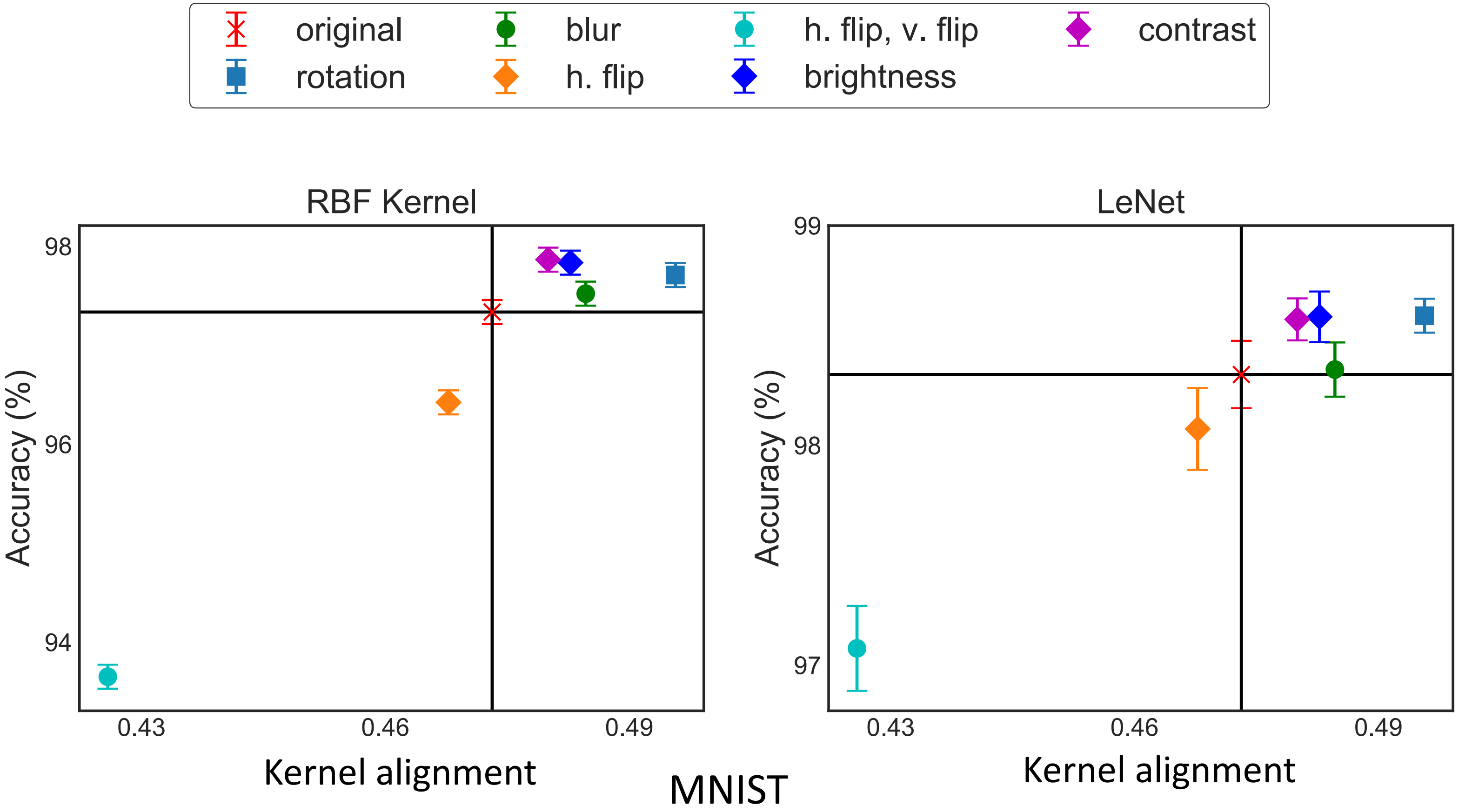
how well separated are features from different classes

A diagnostic: kernel alignment metric



MNIST

A diagnostic: kernel alignment metric



Kernel alignment correlates with accuracy.

Summary

- Data augmentation + k-NN = asymptotic kernel classifier.
- Data augmentation induces invariance and regularizes.
- Application in speeding up training and diagnostics.

Tri Dao
trid@stanford.edu

Poster #227 on Tuesday Jun 11th at 6:30pm