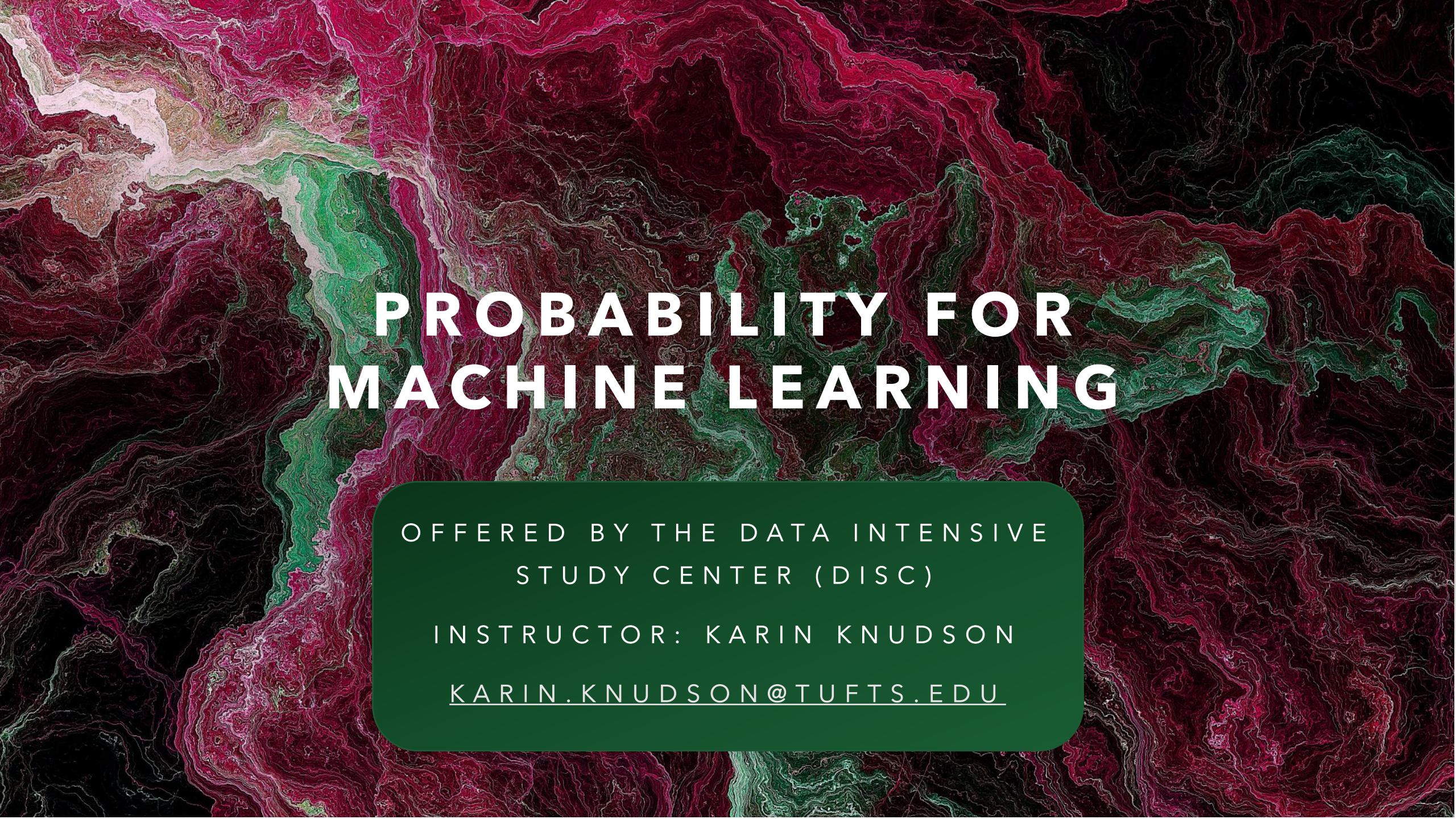


PROBABILITY FOR MACHINE LEARNING

OFFERED BY THE DATA INTENSIVE STUDY CENTER (DISC)

INSTRUCTOR: KARIN KNUDSON

KARIN.KNUDSON@TUFTS.EDU

WHAT WE'VE COVERED SO FAR

- Interpretations of probability
- Random variables (discrete and continuous)
- Probability mass functions, probability density functions
- Expectation, variance, covariance
- Joint, conditional, marginal probabilities
- Independence
- Bayes theorem
- Applications of Bayes theorem for statistical learning – beta/binomial coin-flipping example, ridge reg. and lasso
- Priors, likelihoods, posteriors
- Maximum likelihood, maximum a posteriori (MAP estimates)

WHAT WE'VE COVERED SO FAR

- Linear regression: maximum likelihood approach
- Linear regression with regularization: ridge regression and lasso
- Loss/error functions
- Loss/error functions as derived from likelihood or posterior distributions
- Logistic regression for classification (model, error function, options for regularization)
- Neural network: model, error function
- Gradient descent and stochastic gradient descent

POSTERIOR DISTRIBUTIONS REVISTED: BEYOND THE MAXIMUM

- In our coin flipping example, we mostly looked at the maximum a posteriori (MAP) estimate of the bias of the coin
- BUT, we actually knew the full posterior distribution of the bias of the coin.

POSTERIOR DISTRIBUTIONS REVISTED: BEYOND THE MAXIMUM

- In our coin flipping example, we mostly looked at the maximum a posteriori (MAP) estimate of the bias of the coin
- BUT, we actually knew the full posterior distribution of the bias of the coin.
- What else might we be interested in here?

POSTERIOR DISTRIBUTIONS REVISTED: BEYOND THE MAXIMUM

- In our coin flipping example, we mostly looked at the maximum a posteriori (MAP) estimate of the bias of the coin
- BUT, we actually knew the full posterior distribution of the bias of the coin.
- What else might we be interested in here?
  - Mean?
  - Variance?
  - Some function of the coin-bias?

POSTERIOR DISTRIBUTIONS REVISTED: BEYOND THE MAXIMUM

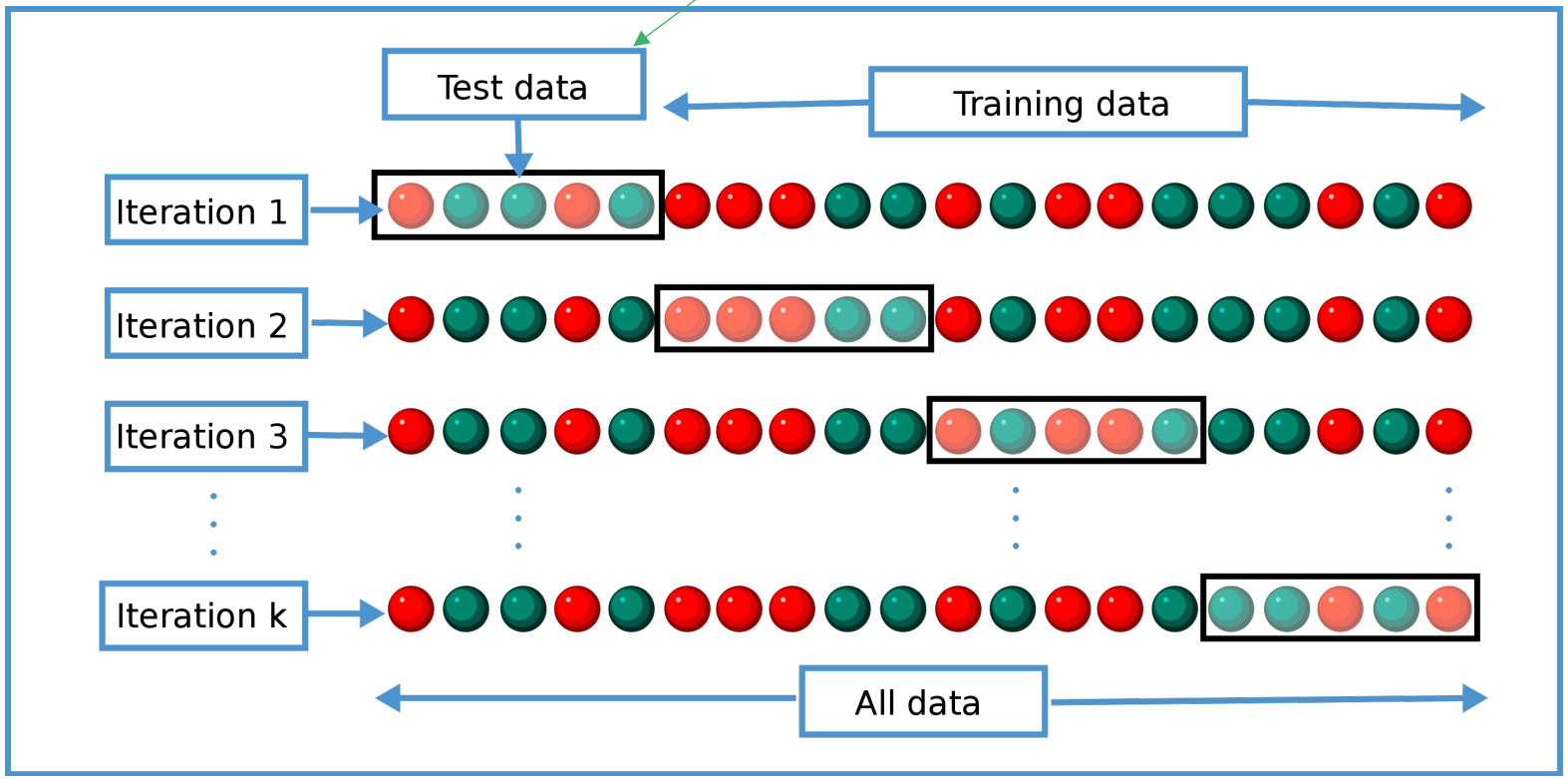
The posterior distribution gives us lots of information beyond the maximum that we can use – e.g. posterior means, quantifications of uncertainty

If we can sample from the posterior distribution, we can compute lots of quantities of interest

Sometimes we have a posterior distribution that might not be nice at all, but from which we can still sample (e.g. using MCMC) – that's not a bad place to be!

Or, validation set

CROSS VALIDATION



Can use to tune model parameters

Want to hold out a separate set of test data to evaluate final model

BIAS-VARIANCE DECOMPOSITION

$$\text{expected loss} = \text{bias}^2 + \text{variance} + \text{noise}$$

bias – how does our prediction differ from an optimal prediction?
(high bias might mean a poor choice of model)

variance – how sensitive is our choice of y to a specific dataset?
(high variance means our model might be overly flexible, sensitive to particular dataset)

bias – how does our prediction differ from an optimal prediction?
variance – how sensitive is our choice of y to a specific dataset?

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] \text{ optimal prediction (for squared loss function)}$$

$$y(\mathbf{x}) \text{ model prediction}$$

$$\mathbb{E}_{data}[(y(\mathbf{x})-h(\mathbf{x}))^2] = [\mathbb{E}_{data}(y(\mathbf{x})-h(\mathbf{x}))]^2+\mathbb{E}_{data}[(y(\mathbf{x})-\mathbb{E}_{data}(y(\mathbf{x}))^2]$$

$$\begin{aligned}
\text{expected loss} &= \iint (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t)\, d\mathbf{x}dt \\
&= \int (y(\mathbf{x}) - h(\mathbf{x}))^2\ p(\mathbf{x})\, d\mathbf{x} + \\
&\iint (h(\mathbf{x}) - t)^2 p(\mathbf{x}, t) d\mathbf{x}dt \\
&= \int [\mathbb{E}_{data}(y(\mathbf{x}) - h(\mathbf{x}))]^2 p(\mathbf{x})\, d\mathbf{x} + \int \mathbb{E}_{data}[(y(\mathbf{x}) - \mathbb{E}_{data}(y(\mathbf{x}))^2] p(\mathbf{x})\, d\mathbf{x} + \\
&\iint (h(\mathbf{x}) - t)^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x}dt
\end{aligned}$$

bias^2

variance

noise

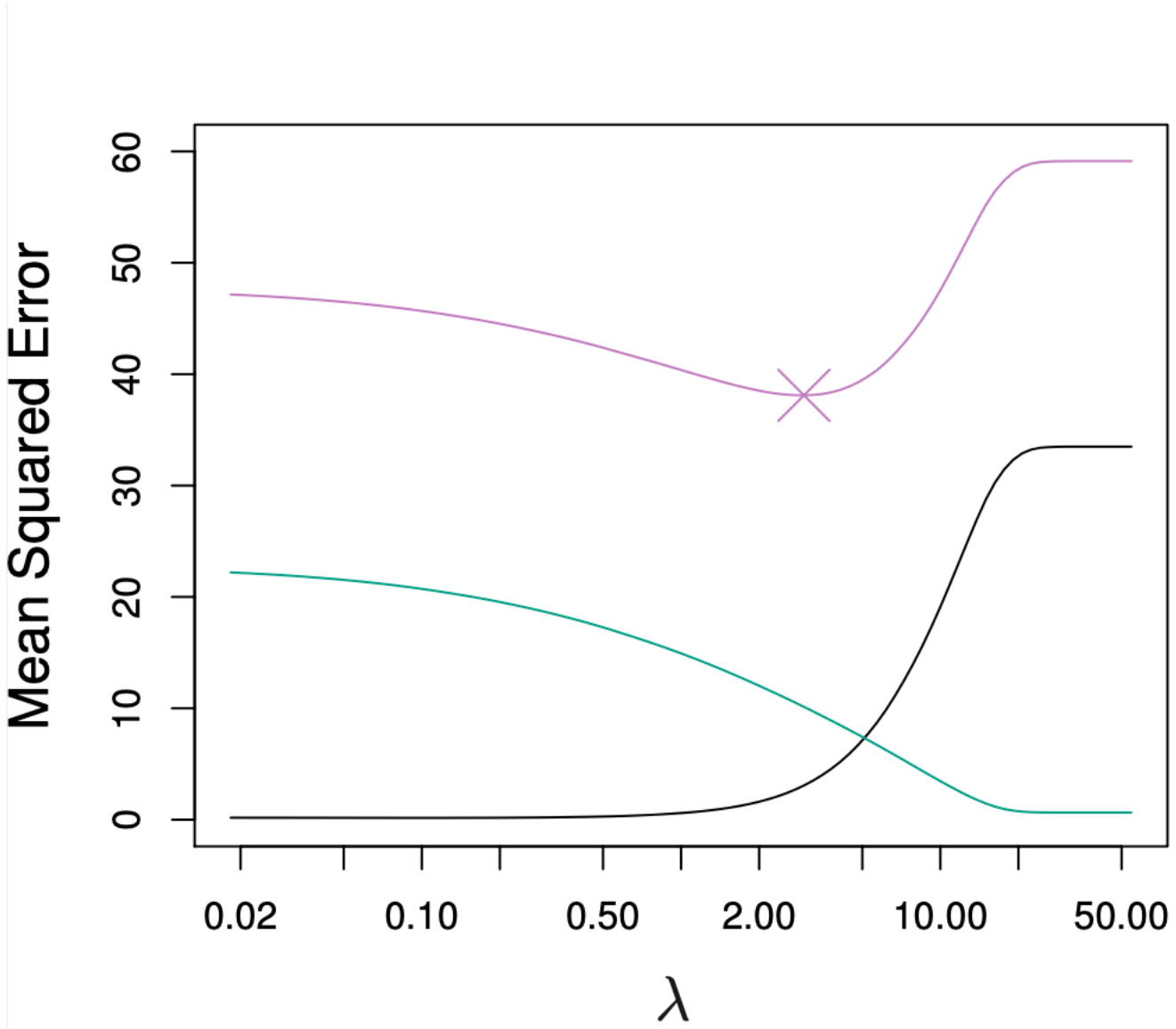
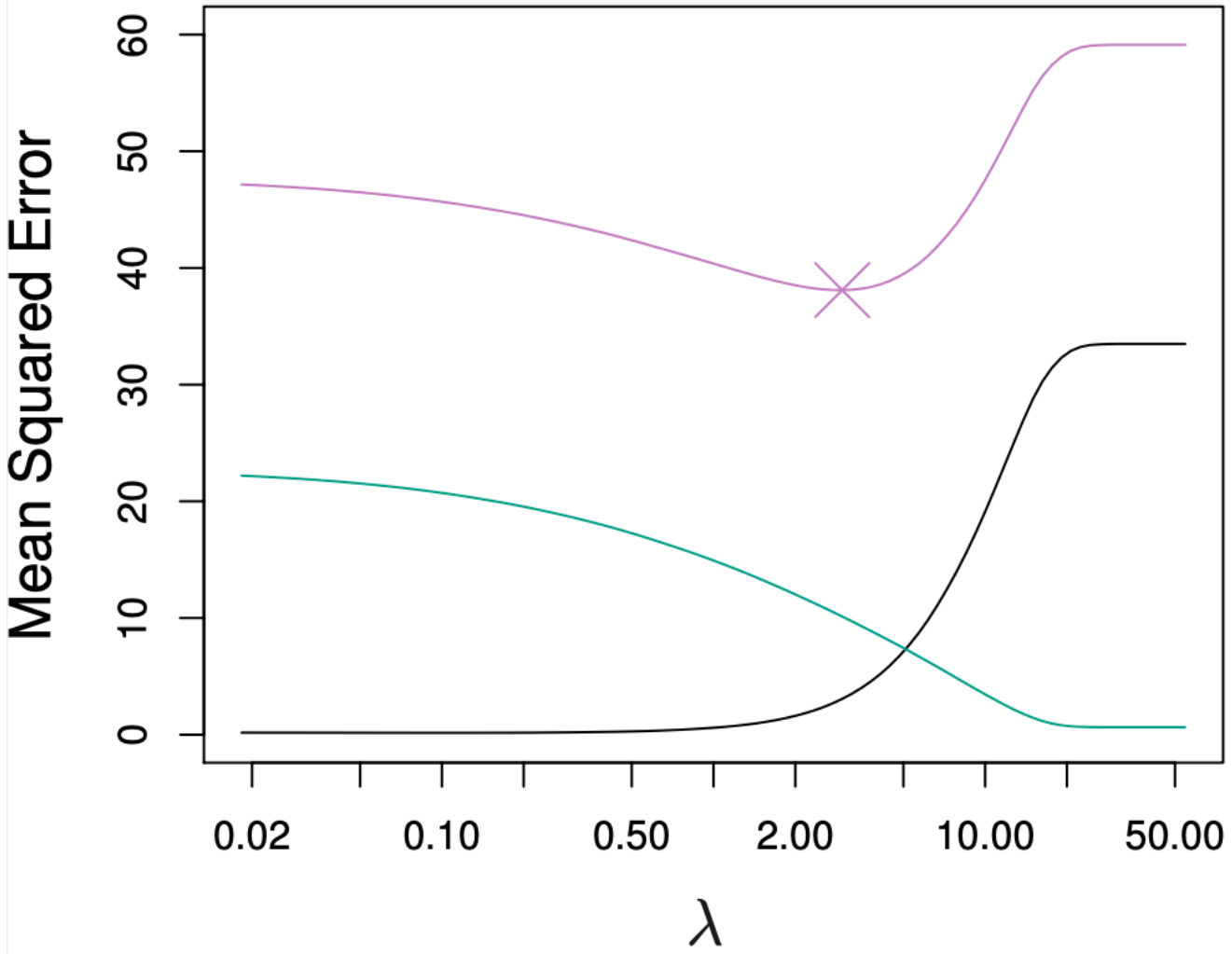


Figure from An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "



Test mean squared error

Squared bias

variance

Figure from An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

REFERENCE

Pattern Recognition and Machine Learning by Christopher Bishop, esp. Ch3-5:

https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani