6.7900: Machine Learning
Lecture 7

Lecture start: Tues/Thurs 2:35pm

Who's speaking today? Prof. Tamara Broderick

Course website: gradml.mit.edu

Questions? Ask here or on piazza.com/mit/fall2024/67900/

Materials: Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

Last Time

I. Visualizing regression
II. Uncertainty
III. Ridge regression
IV. More flexible/complex features

Today

I. Evaluation for supervised learning
II. Choosing hyperparams
III. Validation data and empirical risk

More complex features

More complex features

- Is it always better to use more complex/flexible feature sets?

More complex features

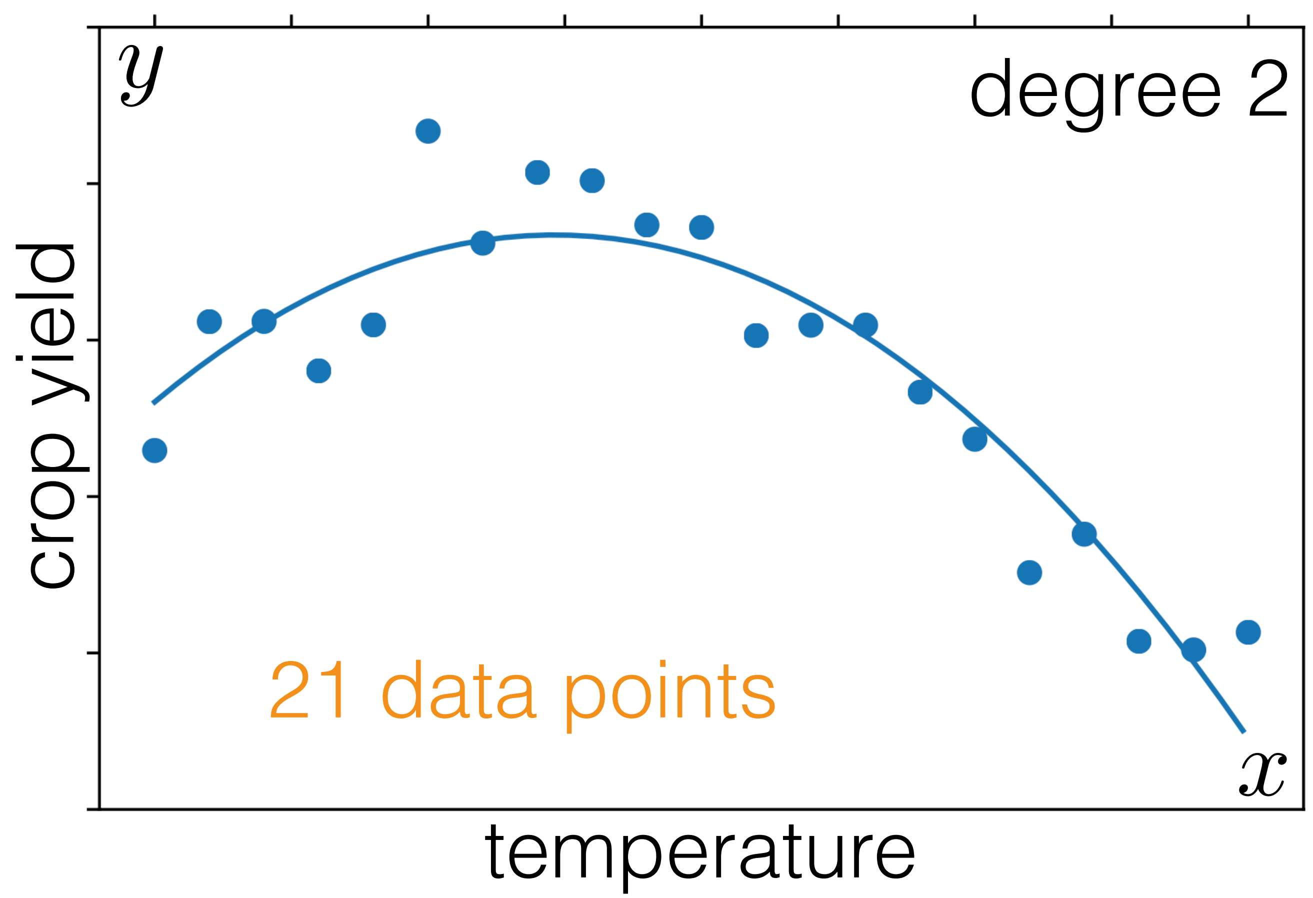
- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree

More complex features

- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$

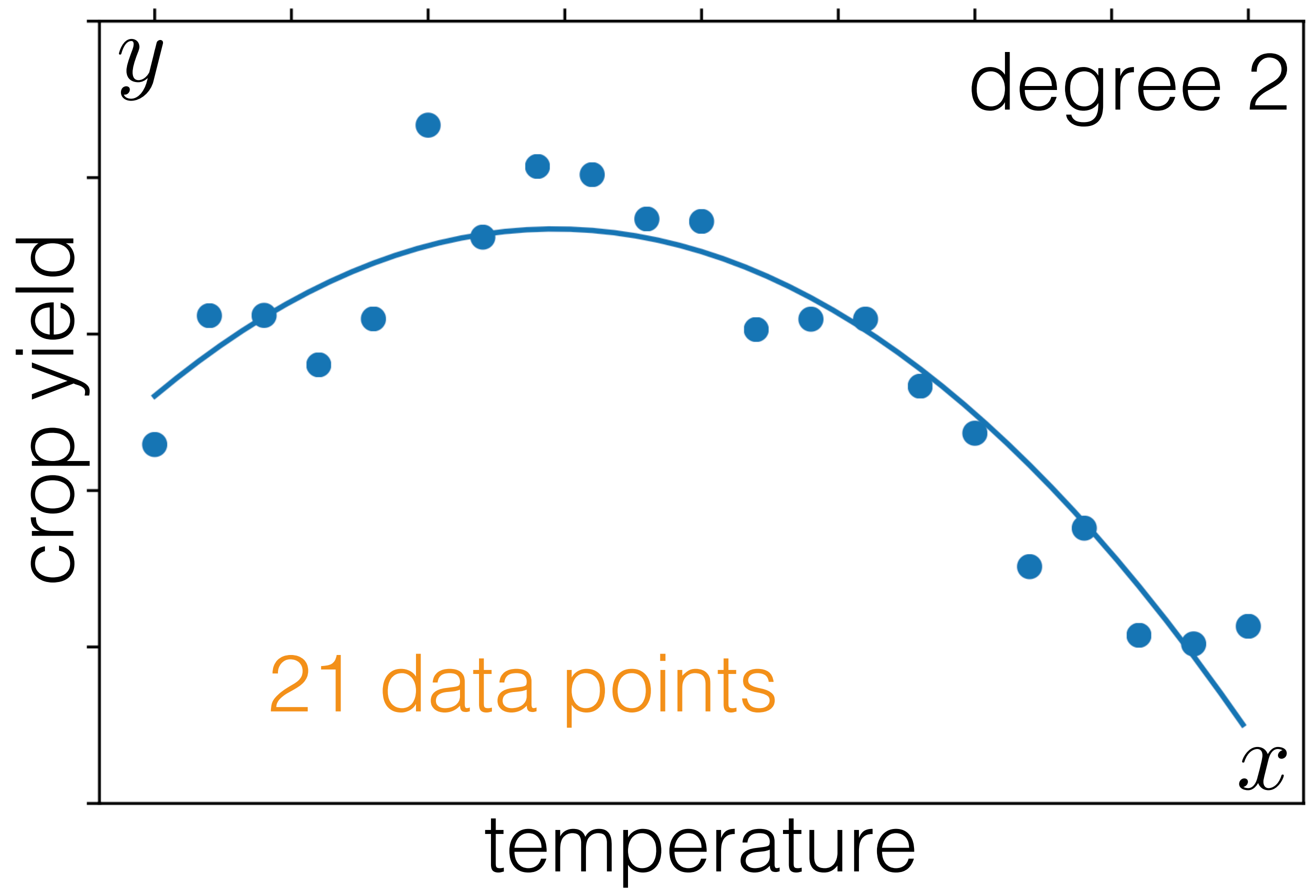
More complex features

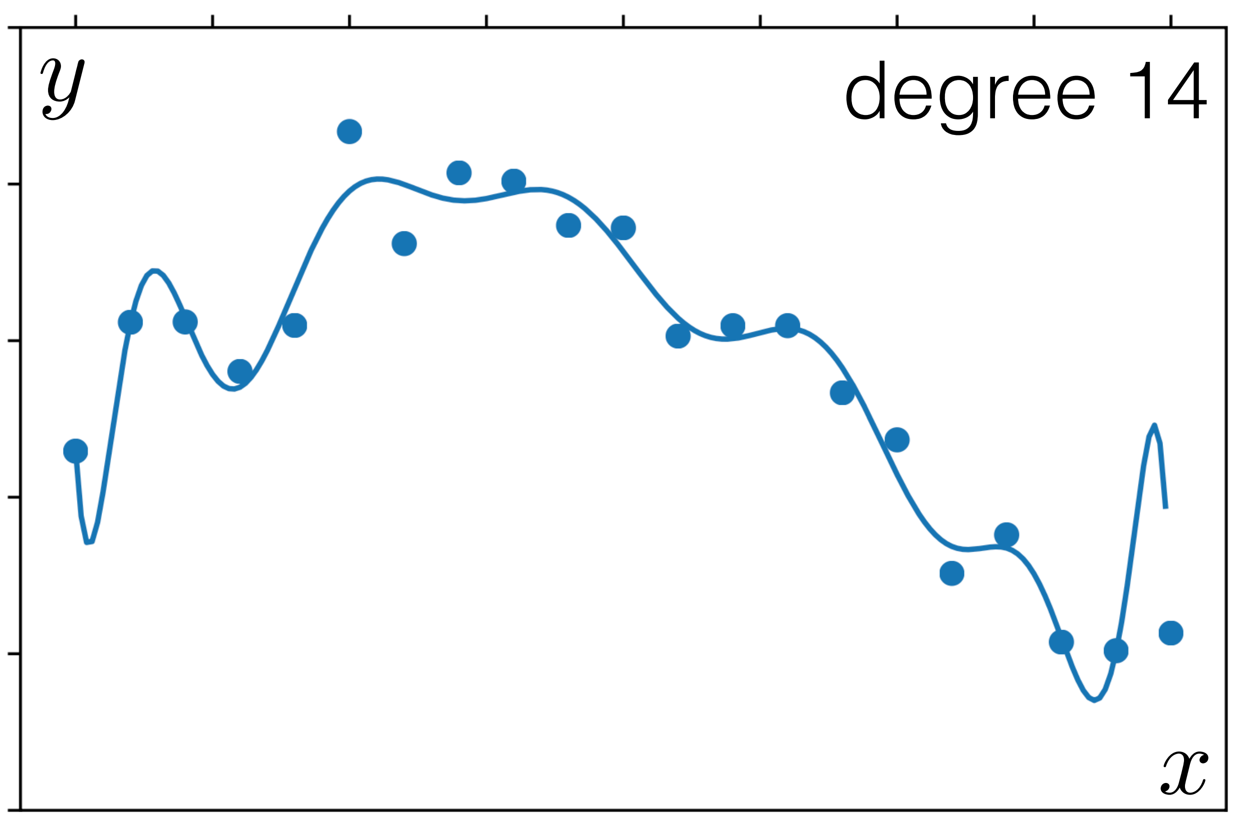
- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$



More complex features

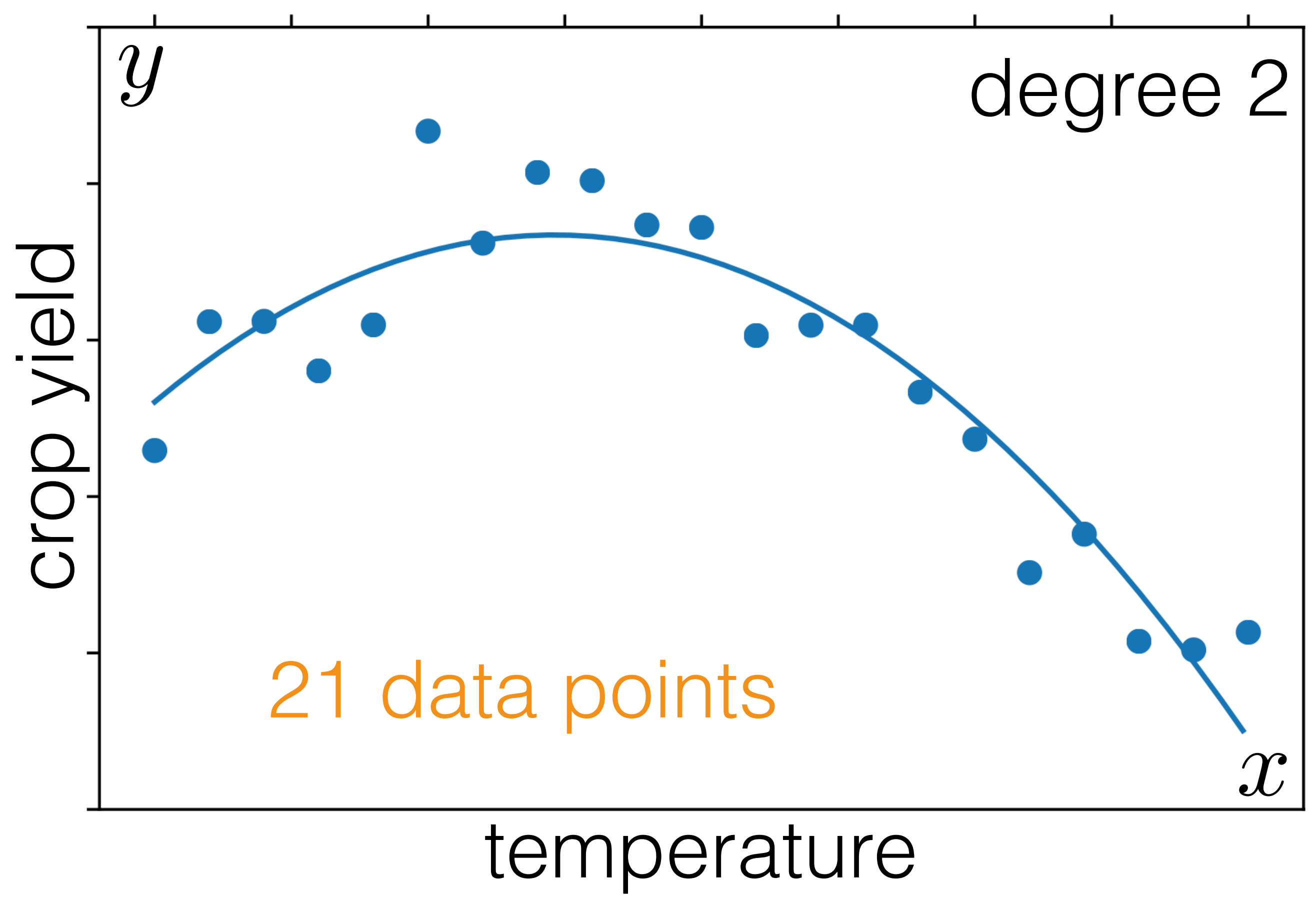
- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$

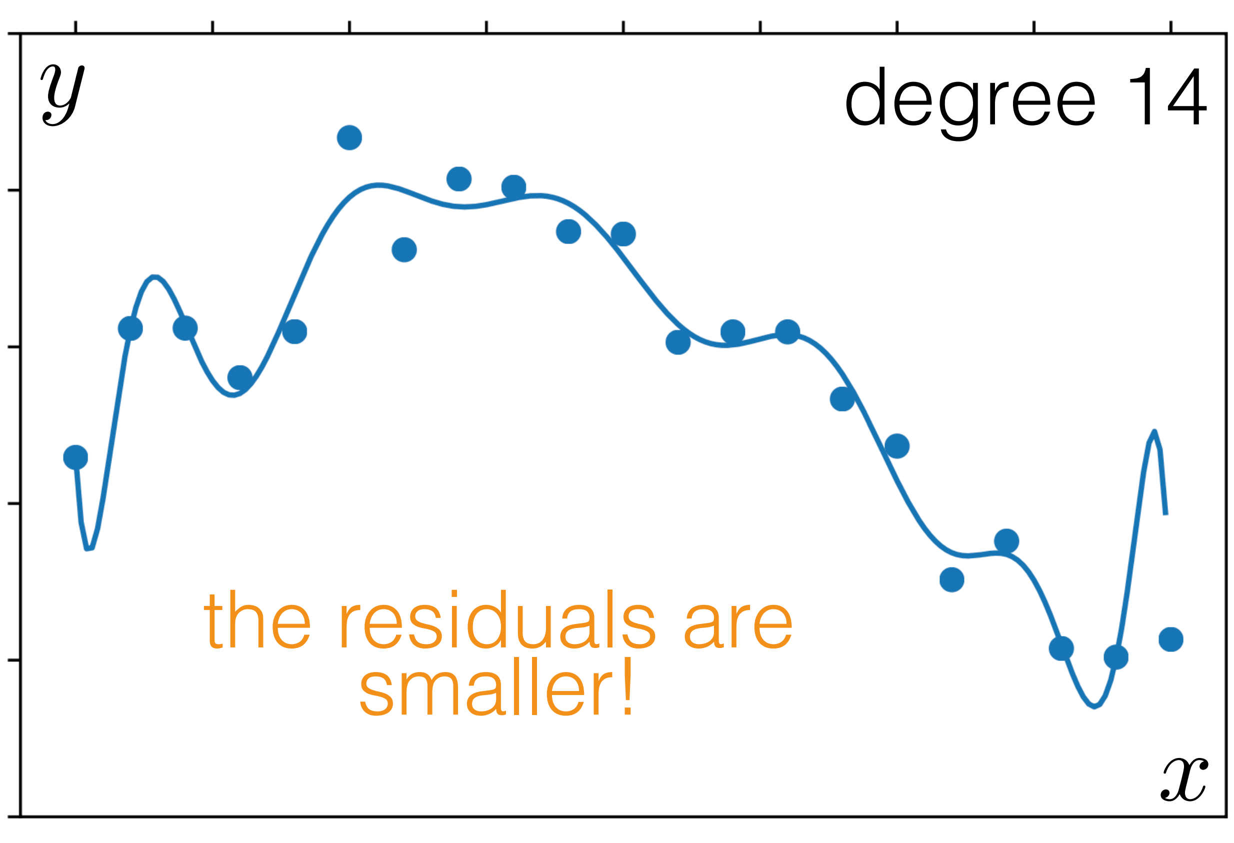




More complex features

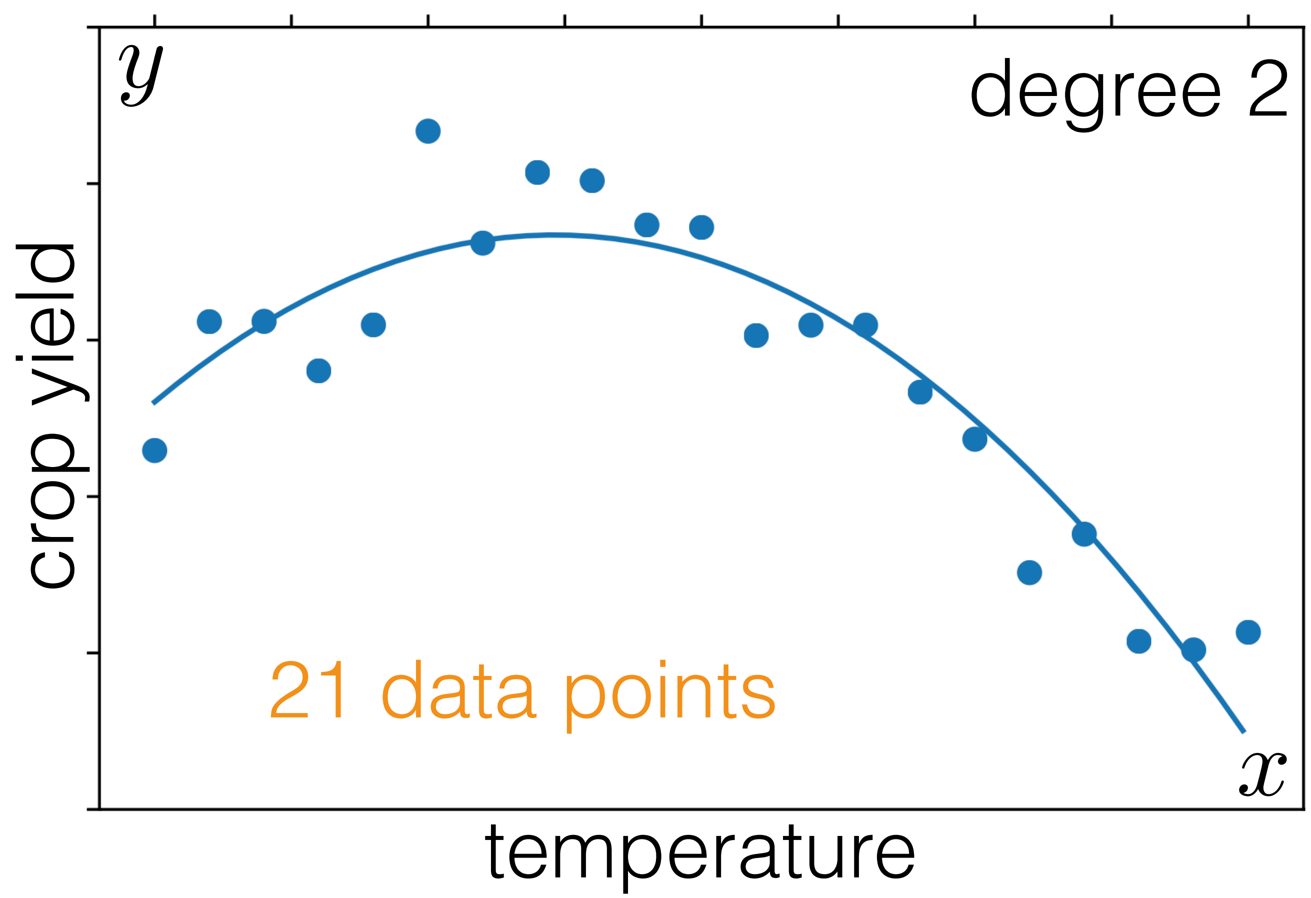
- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$

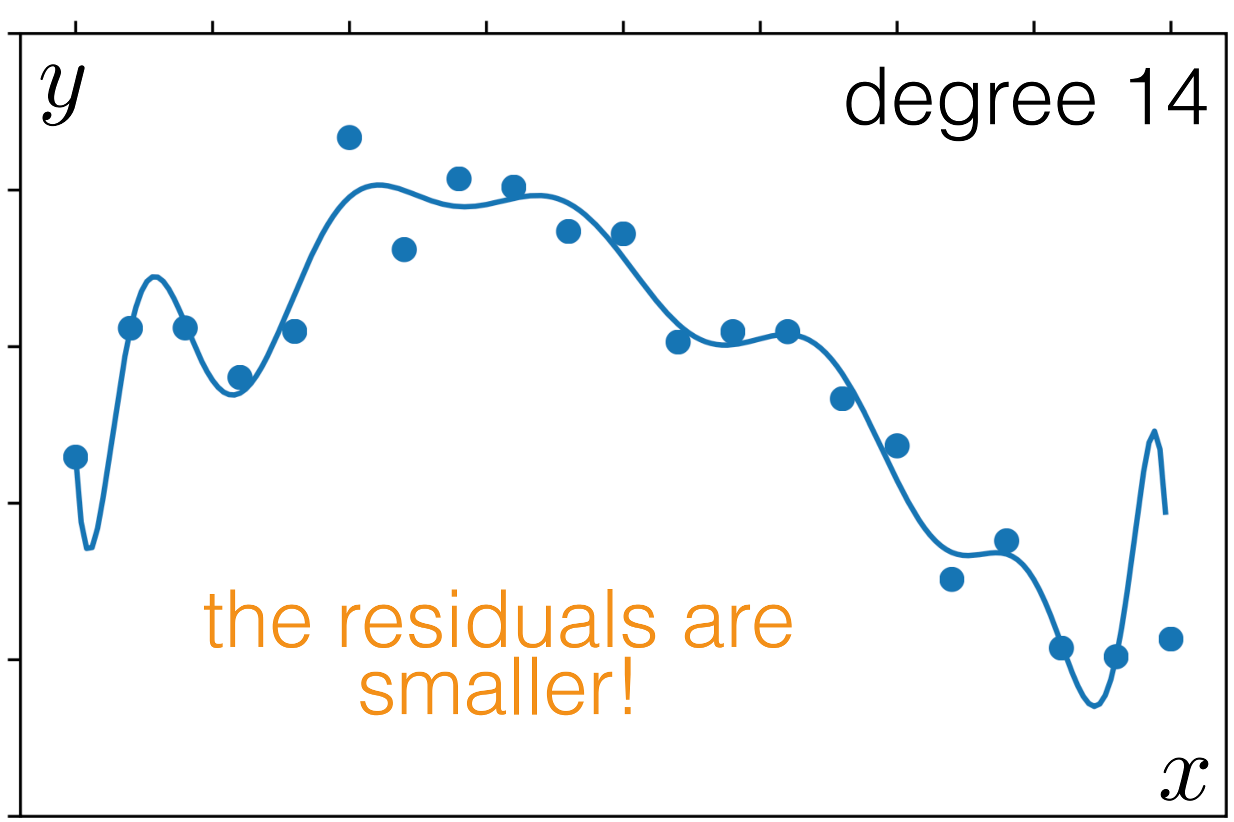


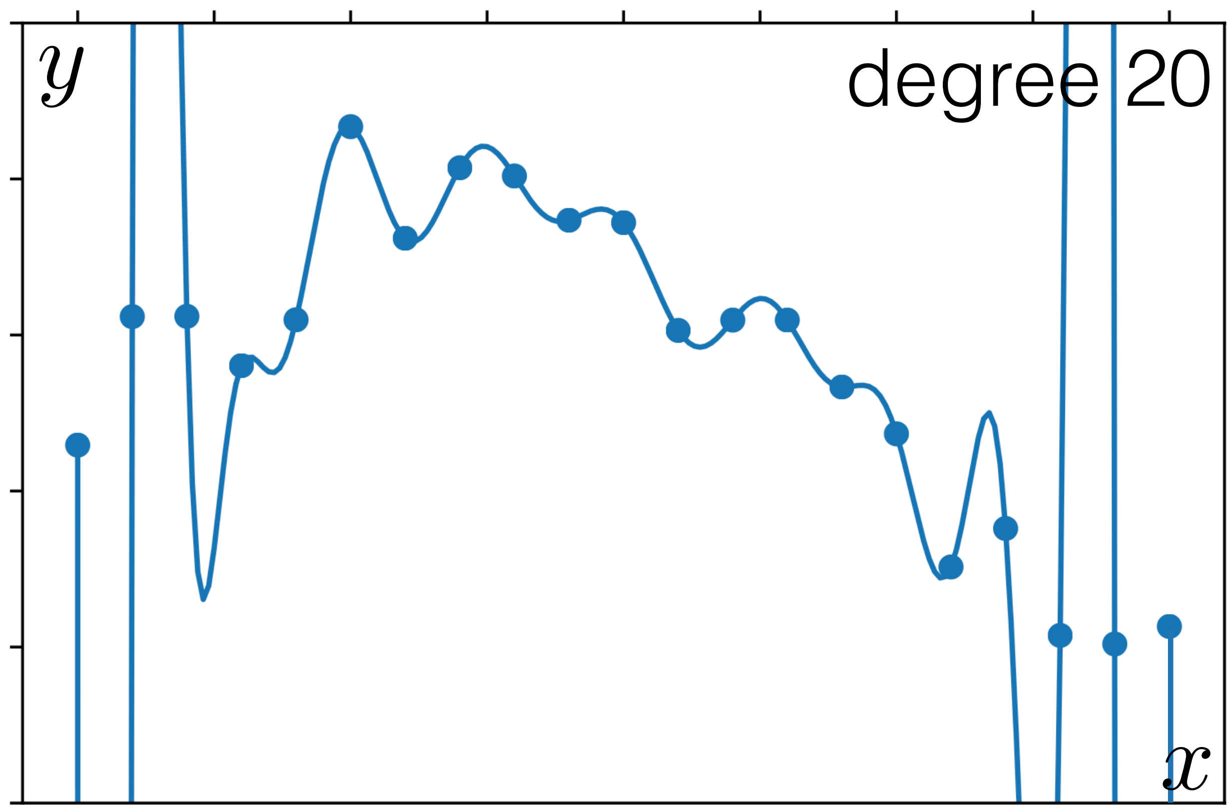


More complex features

- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$

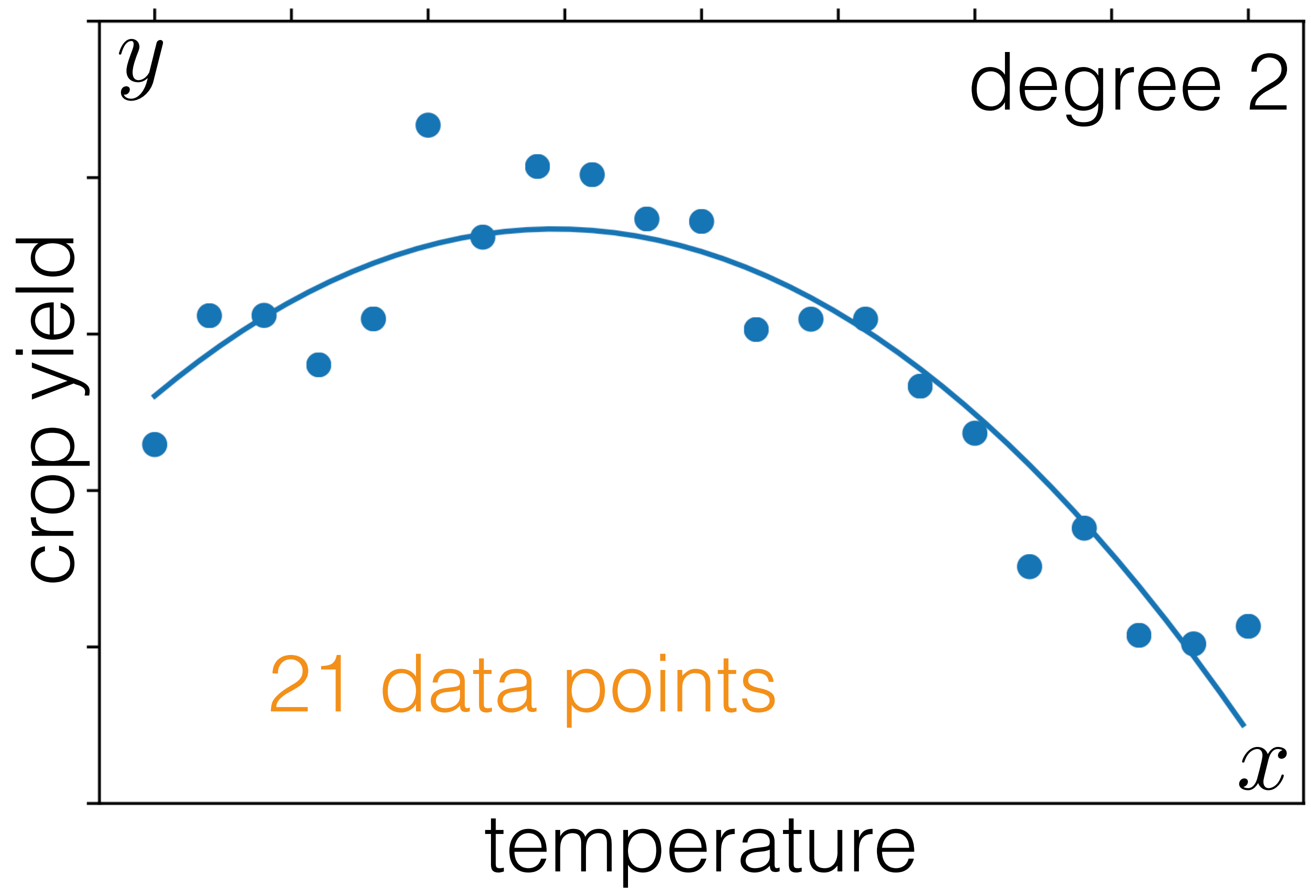


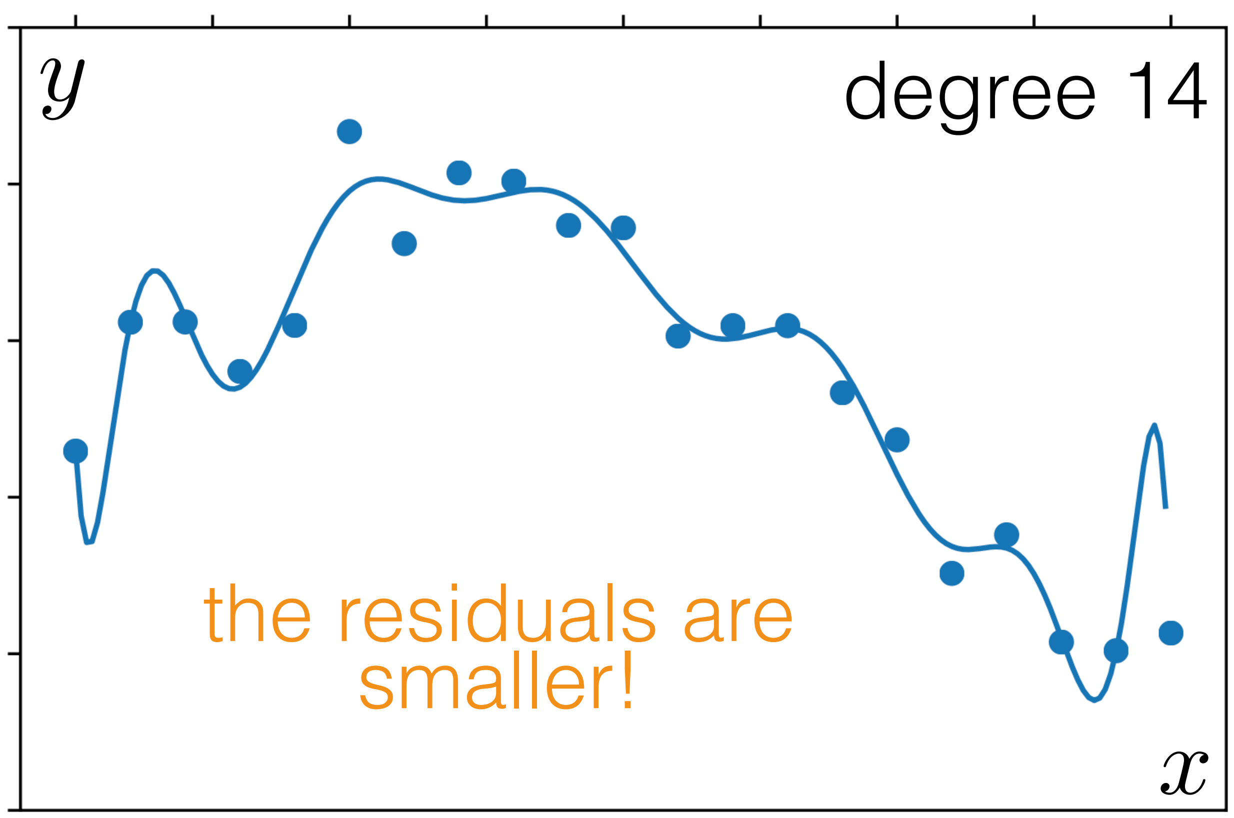


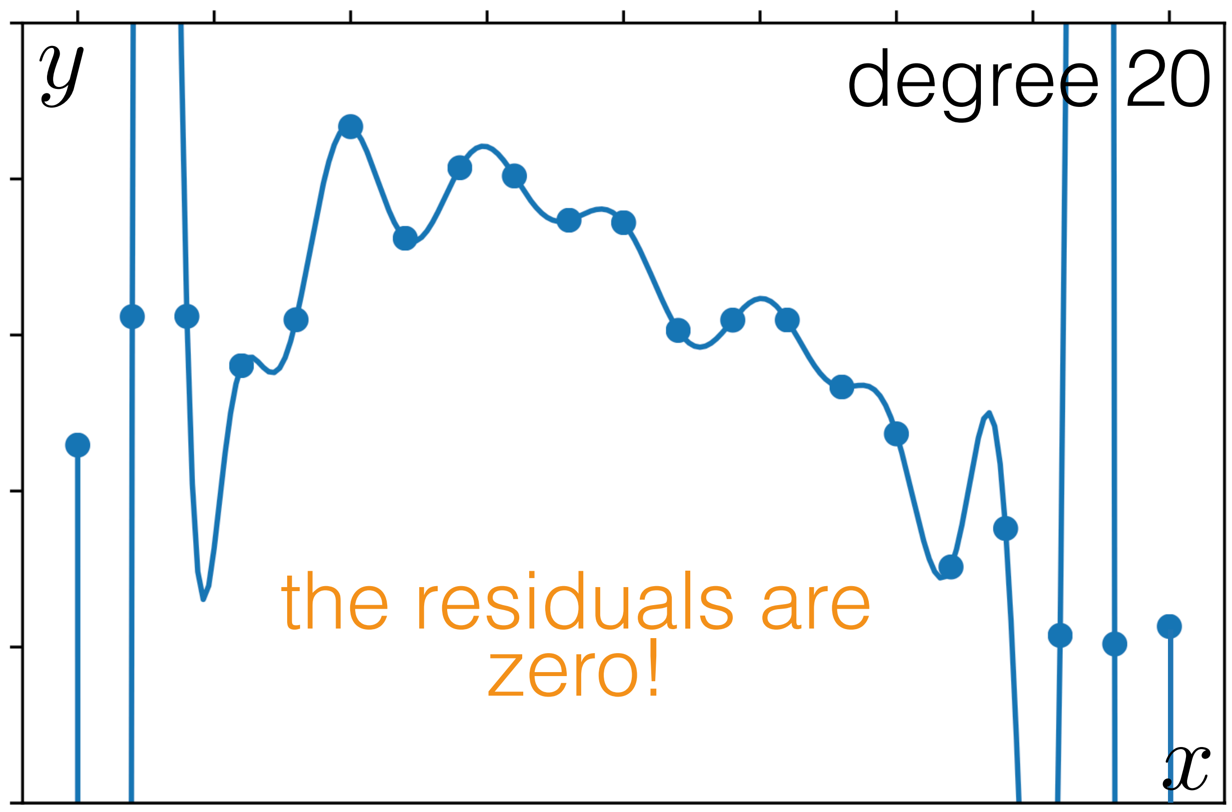


More complex features

- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$

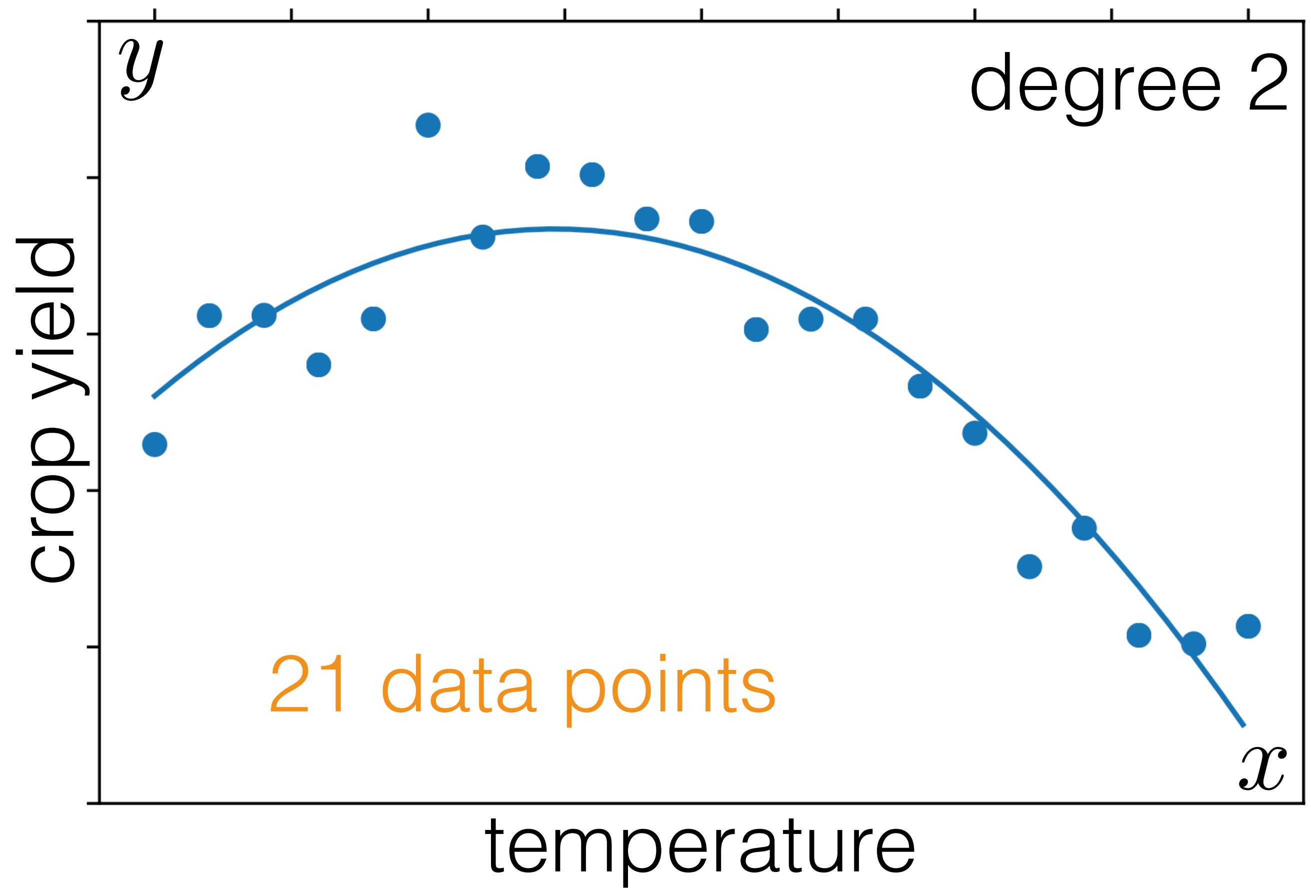


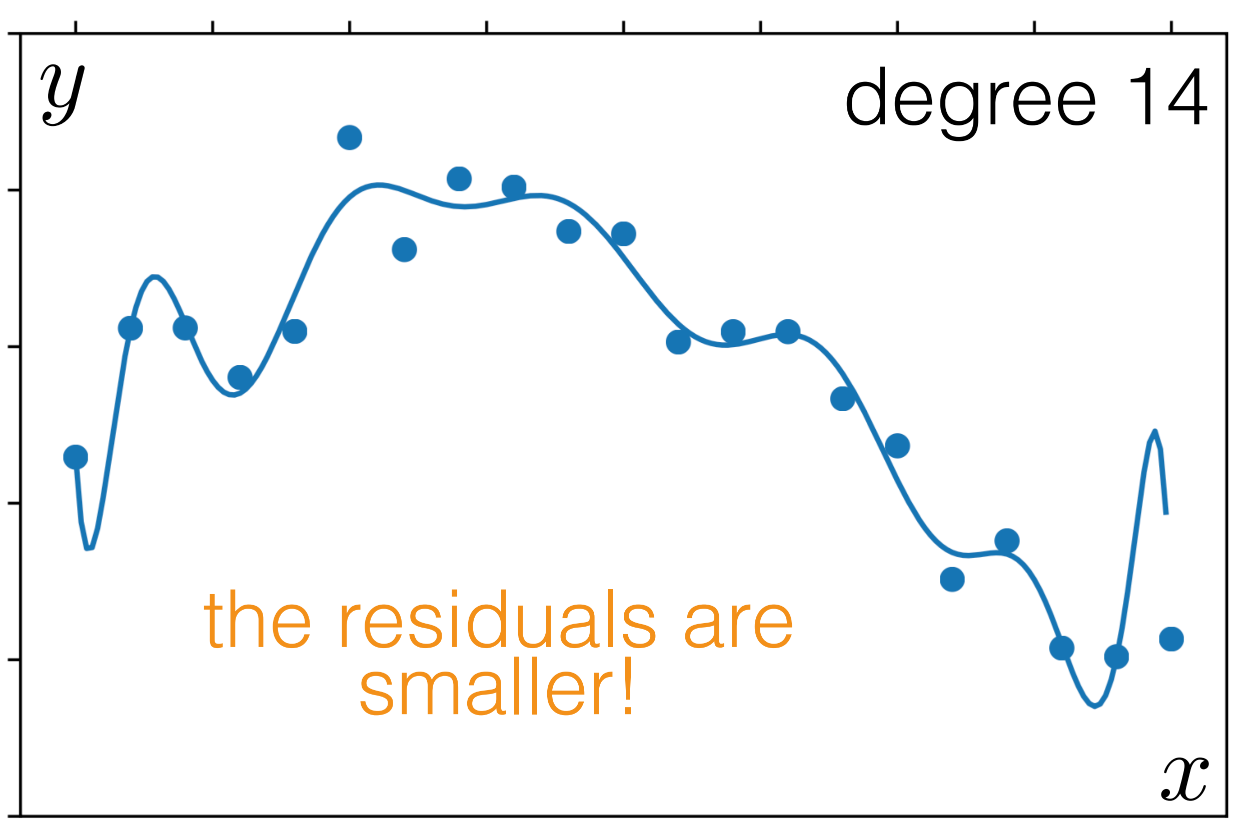


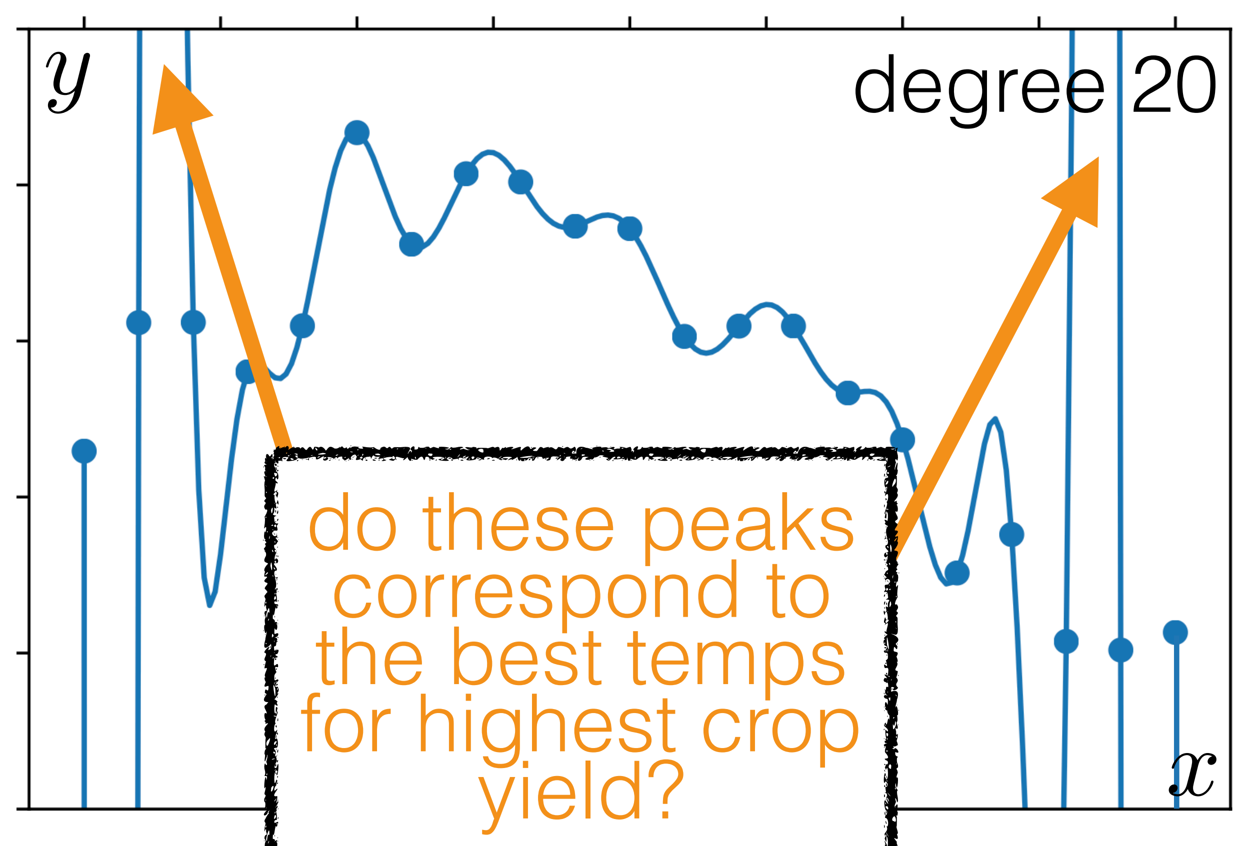


More complex features

- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$

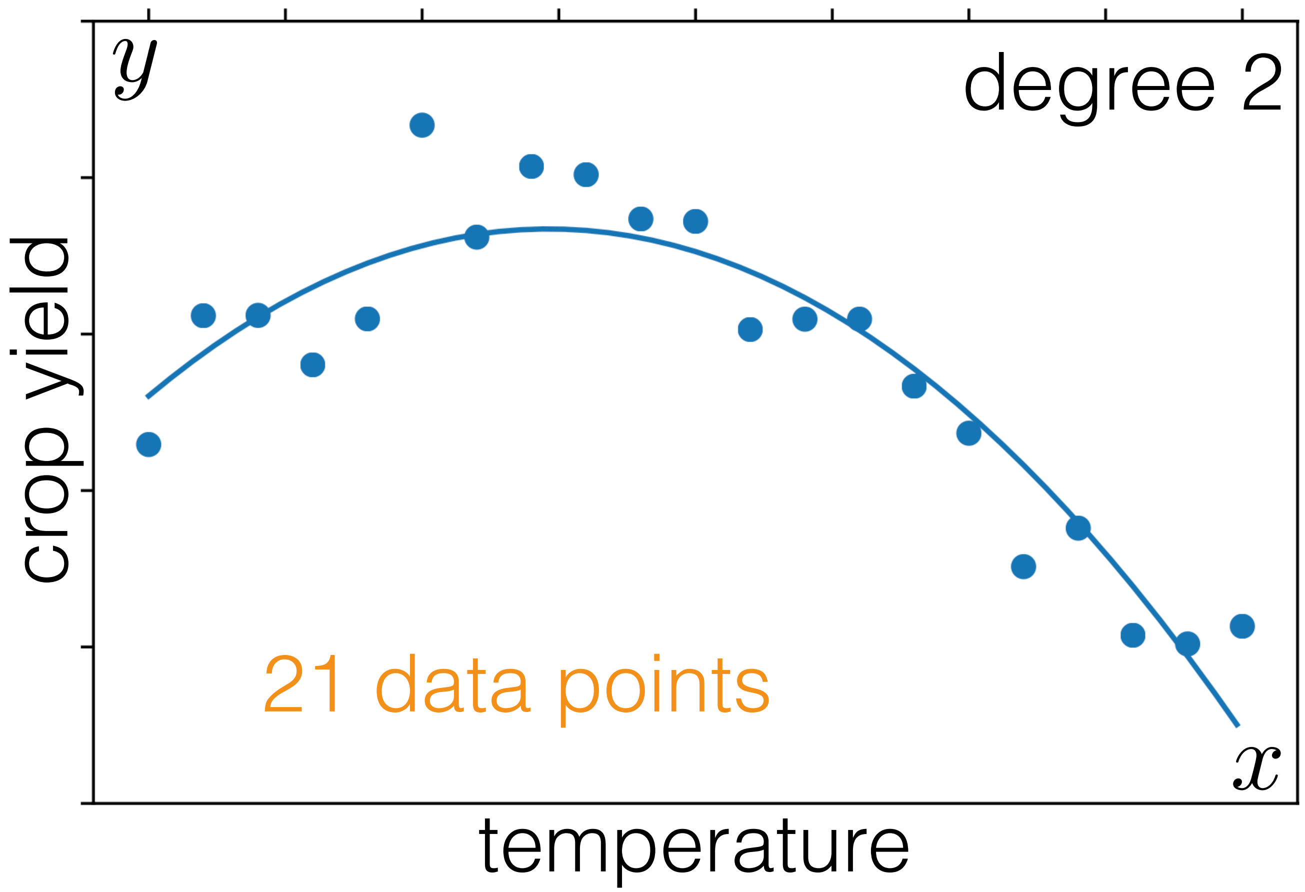


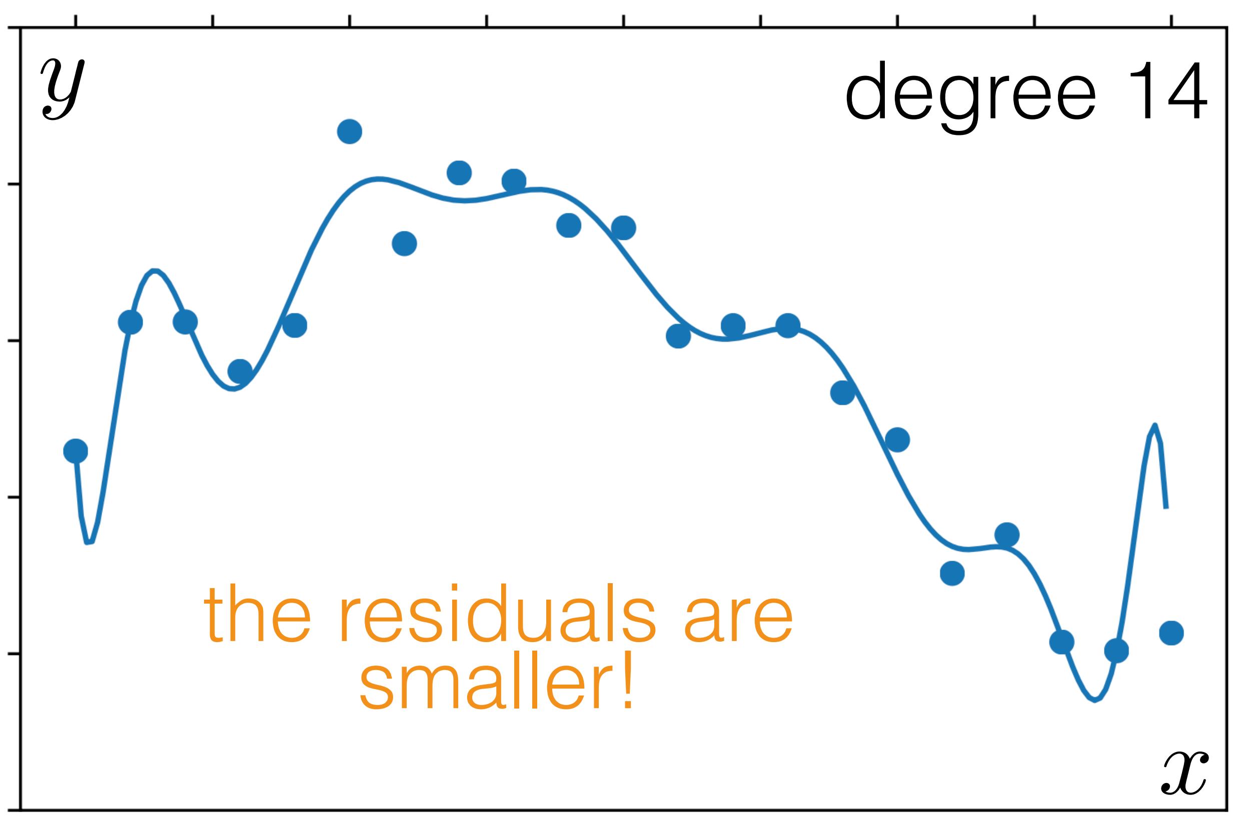


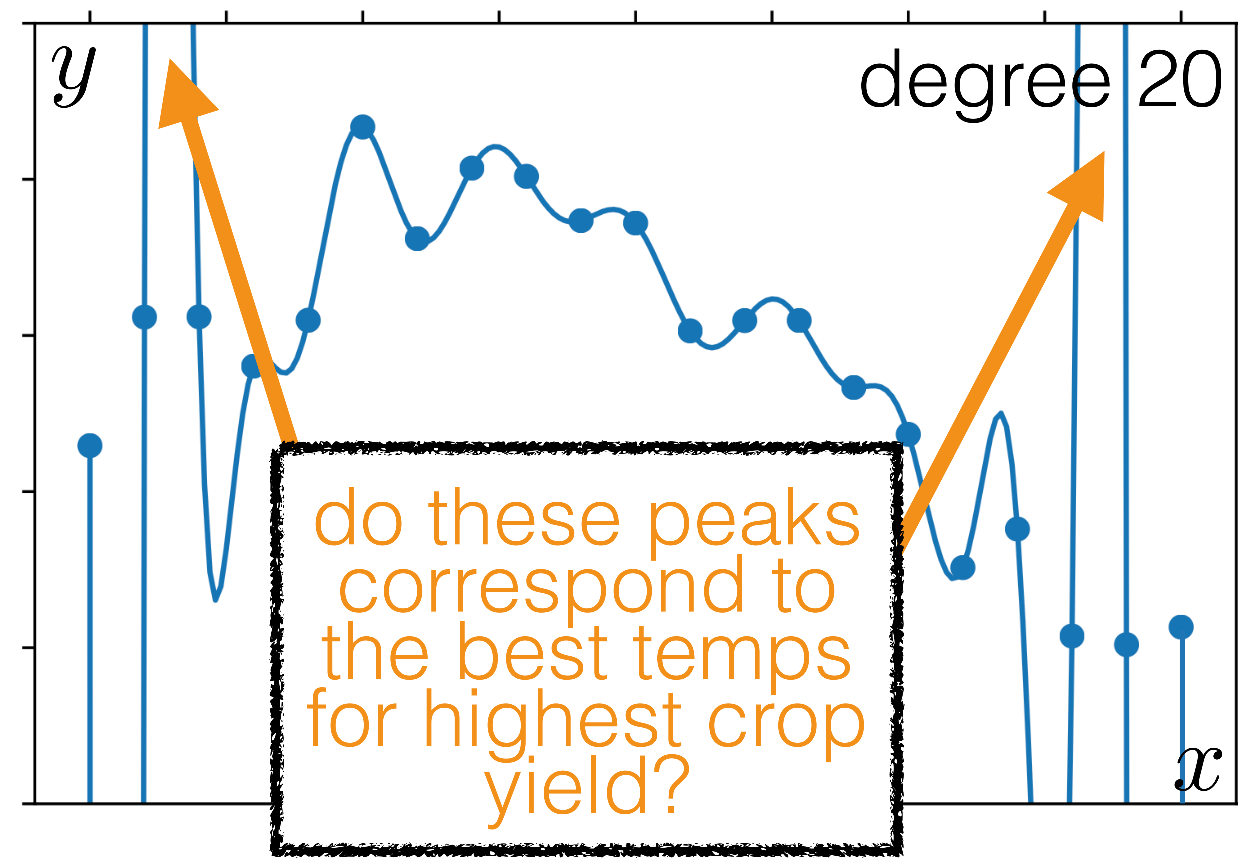


More complex features

- Is it always better to use more complex/flexible feature sets? OLS with polynomial features up to specified degree
- E.g. deg 2: $\phi(x) = [1, x, x^2]^\top$
- Always plot your data!
- Harder in higher dimensions







Common questions in machine learning

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
    - E.g. choose the max degree of polynomial features
    - E.g. choose the ridge regression hyperparameter λ

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X,Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
  - If we knew this value, we could just report it for question #1 and compare the values for different $h_{\mathcal{D}}$ for question #2

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
  - If we knew this value, we could just report it for question #1 and compare the values for different $h_{\mathcal{D}}$ for question #2 plot

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
  - If we knew this value, we could just report it for question #1 and compare the values for different $h_{\mathcal{D}}$ for question #2 plot
  - The whole problem is that we don't know it (déjà vu)

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
  - If we knew this value, we could just report it for question #1 and compare the values for different $h_{\mathcal{D}}$ for question #2 plot
  - The whole problem is that we don't know it (déjà vu)
- Like before, we've got to estimate it. A key difference is that now we assume $h_{\mathcal{D}}$ was already fit using training data

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
  - If we knew this value, we could just report it for question #1 and compare the values for different $h_{\mathcal{D}}$ for question #2 plot
  - The whole problem is that we don't know it (déjà vu)
- Like before, we've got to estimate it. A key difference is that now we assume $h_{\mathcal{D}}$ was already fit using training data
- What tools do we have to potentially help us?

Common questions in machine learning

1. How well will my decision rule $h_{\mathcal{D}}$ perform on new data (not the training data $\mathcal{D}$)? E.g. new planting, new (spam?) email
2. How can I choose among a set of possible decision rules $h_{\mathcal{D}}$ that I've already fit to training data?
   - E.g. choose the max degree of polynomial features
   - E.g. choose the ridge regression hyperparameter λ
     - From data rather than "prior" information
     - "Empirical Bayes"

- We've already specified performance of a decision rule on new data: the risk of a new point (X, Y) is $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
  - If we knew this value, we could just report it for question #1 and compare the values for different $h_{\mathcal{D}}$ for question #2 plot
  - The whole problem is that we don't know it (déjà vu)
- Like before, we've got to estimate it. A key difference is that now we assume $h_{\mathcal{D}}$ was already fit using training data
- What tools do we have to potentially help us? board

Empirical risk over the training data

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data:
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

Empirical risk over the training data

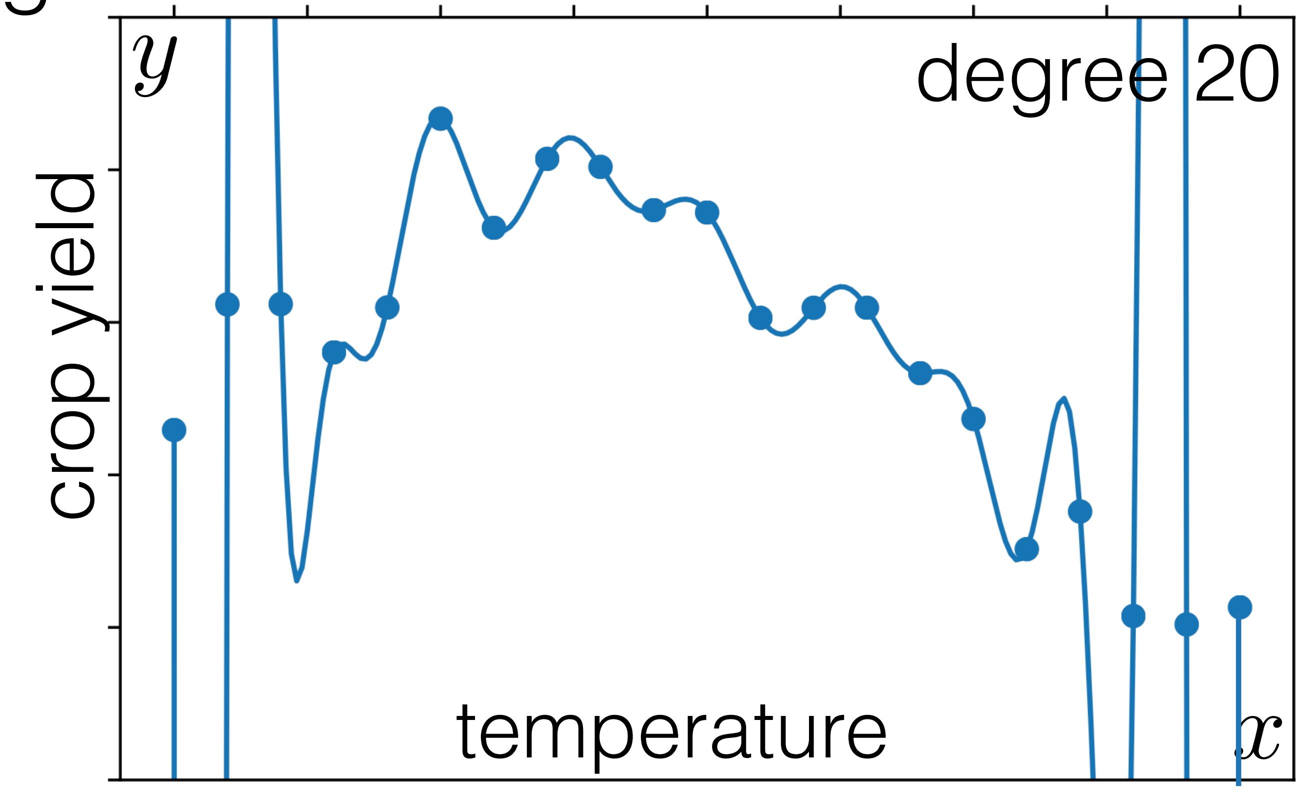
- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?

  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

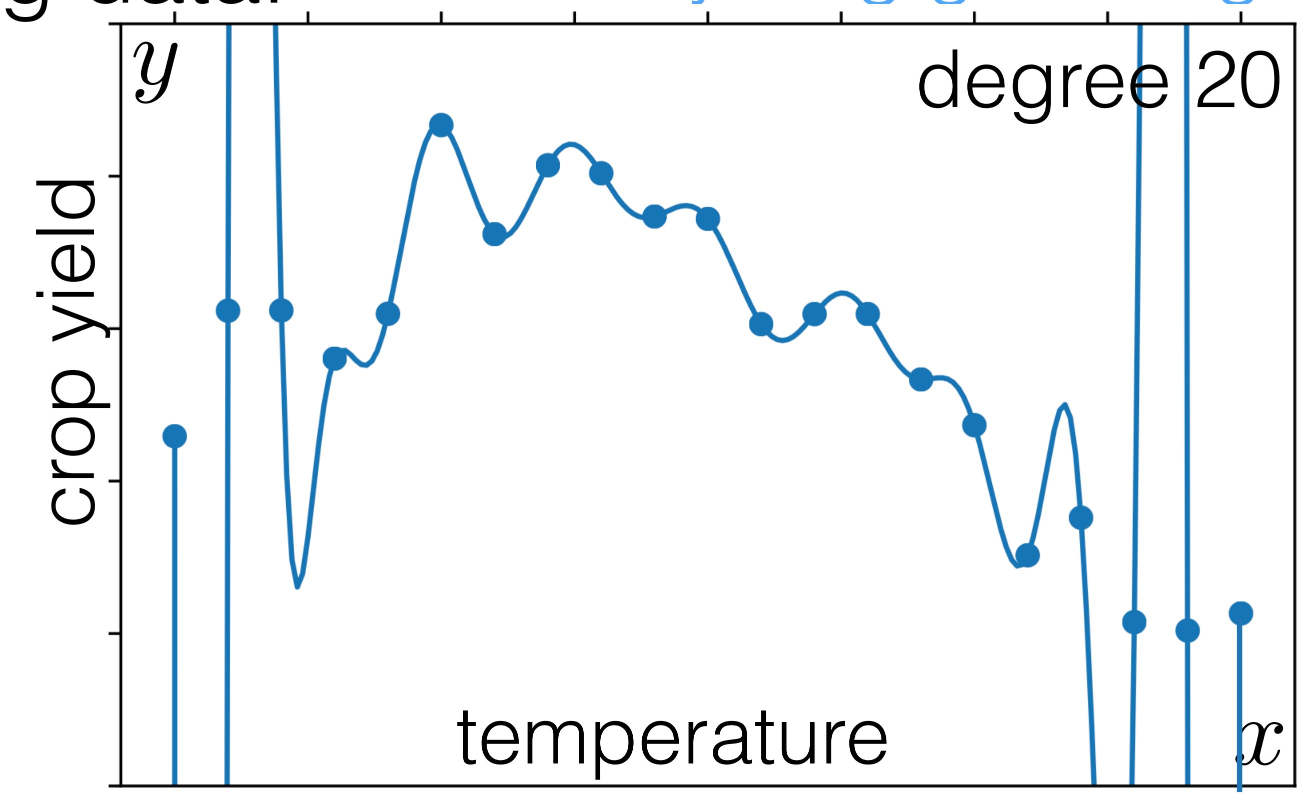
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero



Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data:
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

could anything go wrong?

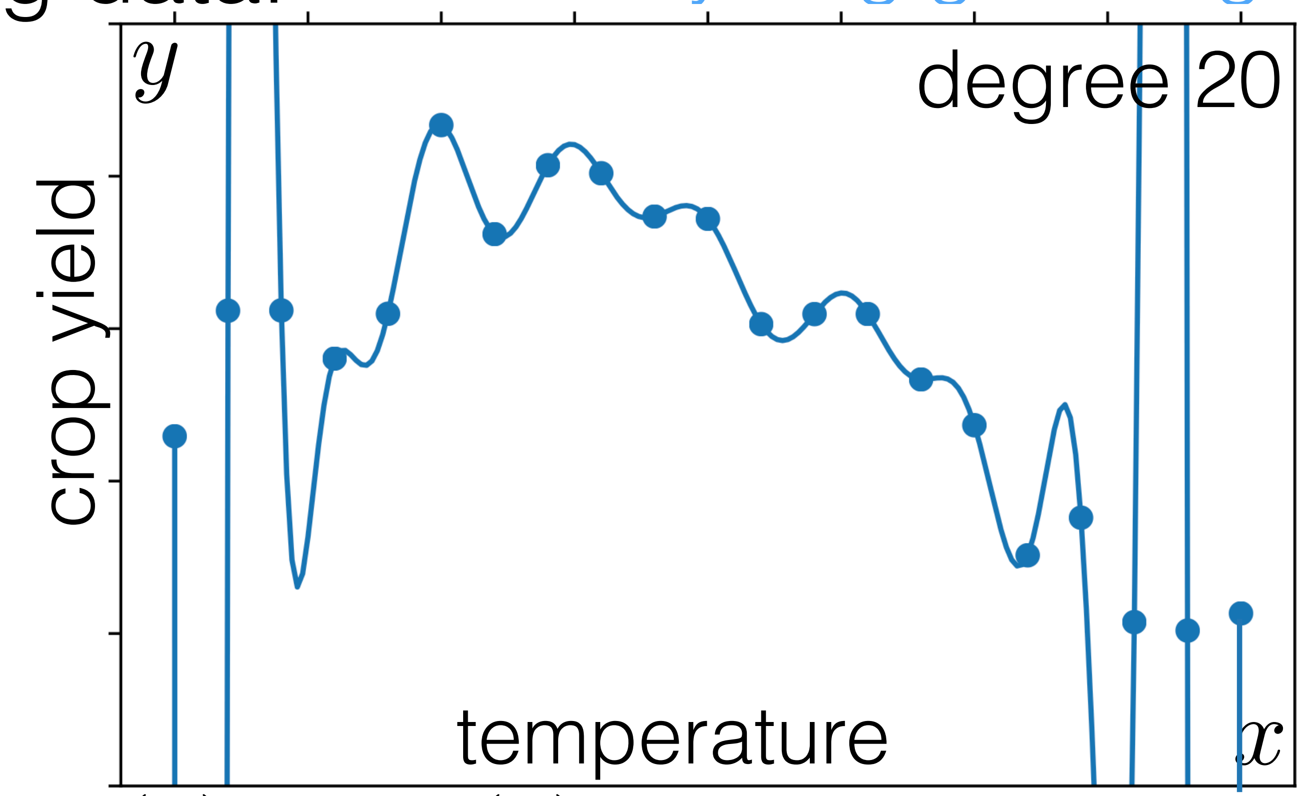


- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero
- Is the problem that we don't have enough training data?

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data:
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

could anything go wrong?

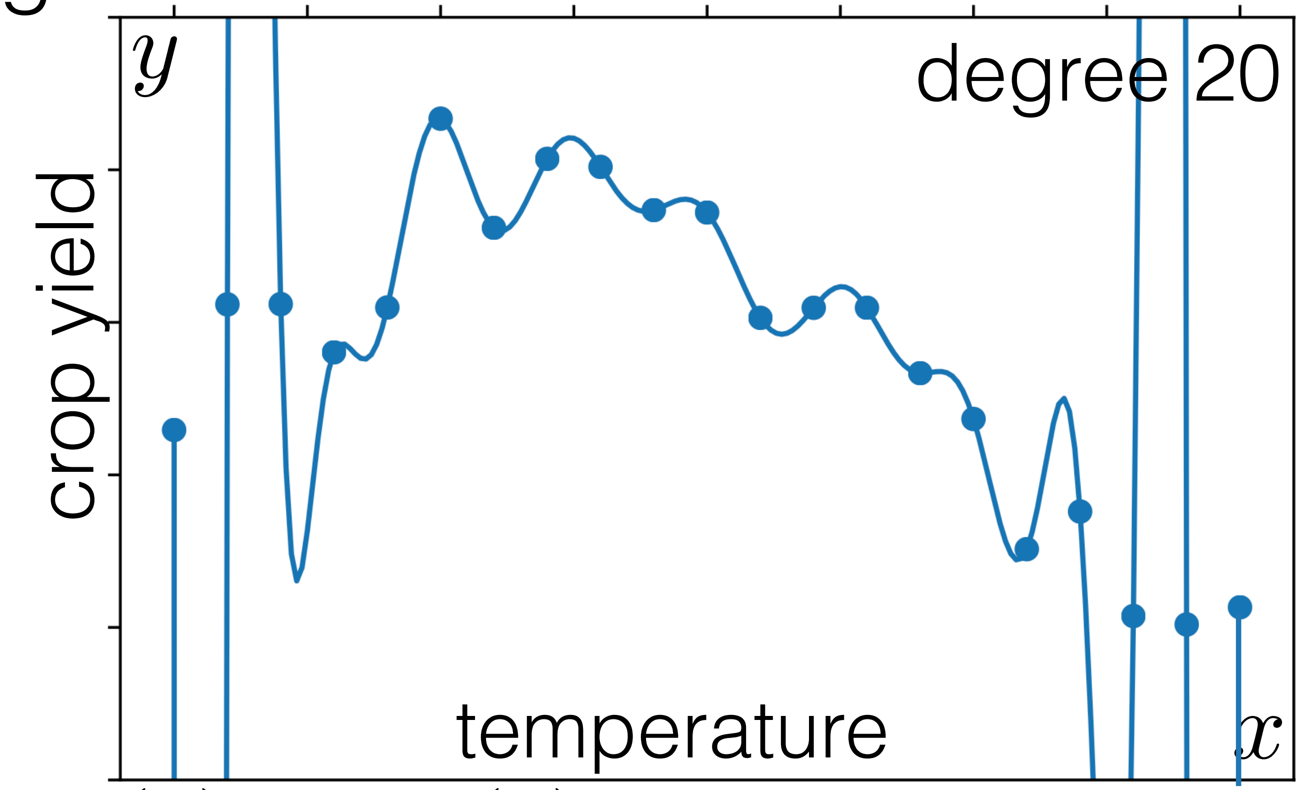


- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero
- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

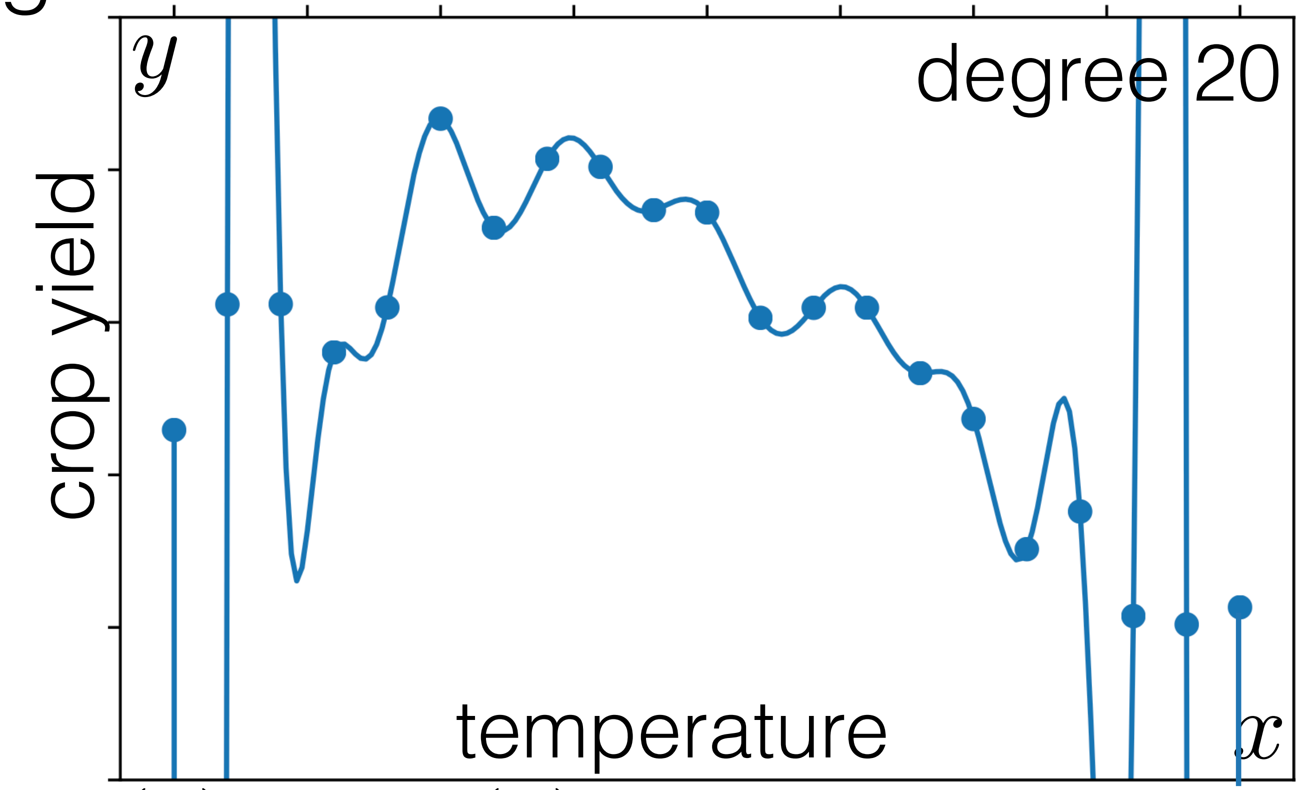
could anything go wrong?



- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero
- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N*

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero

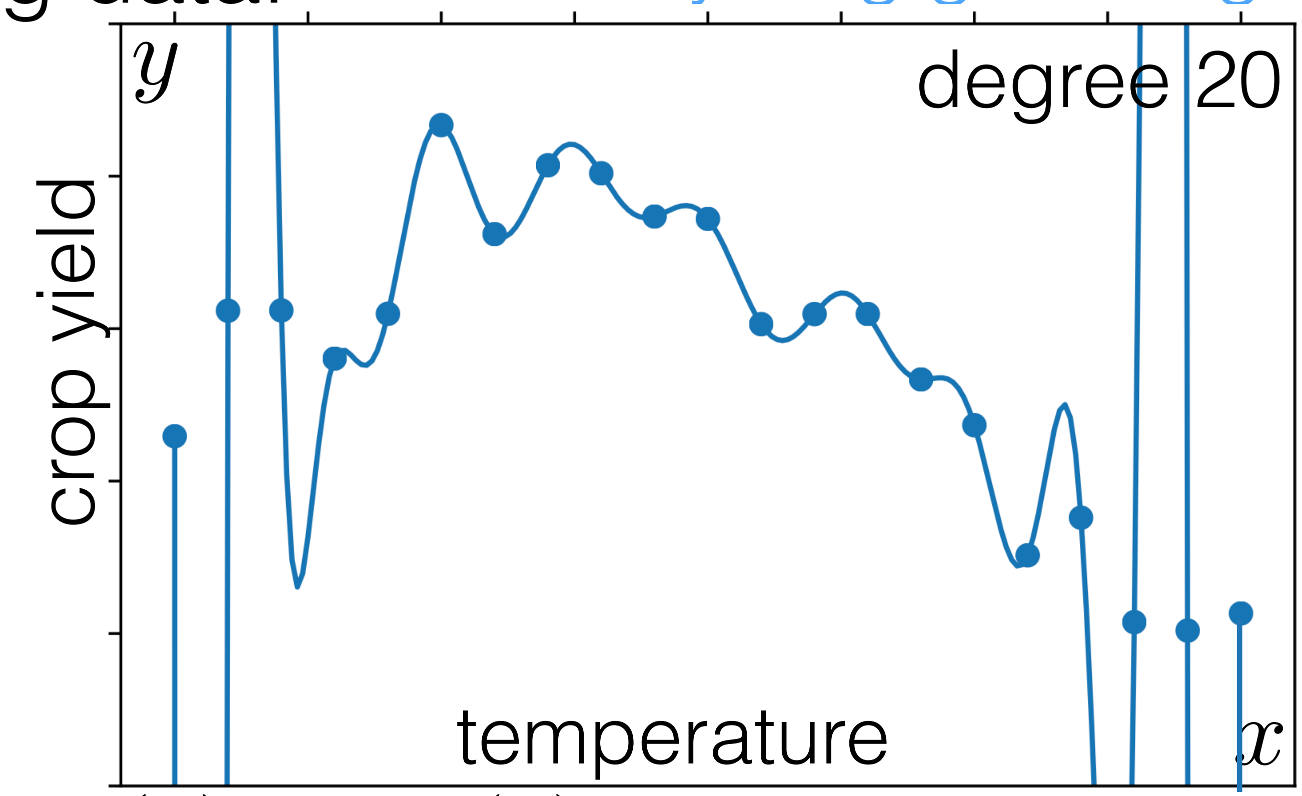


- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N*
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \nrightarrow \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$

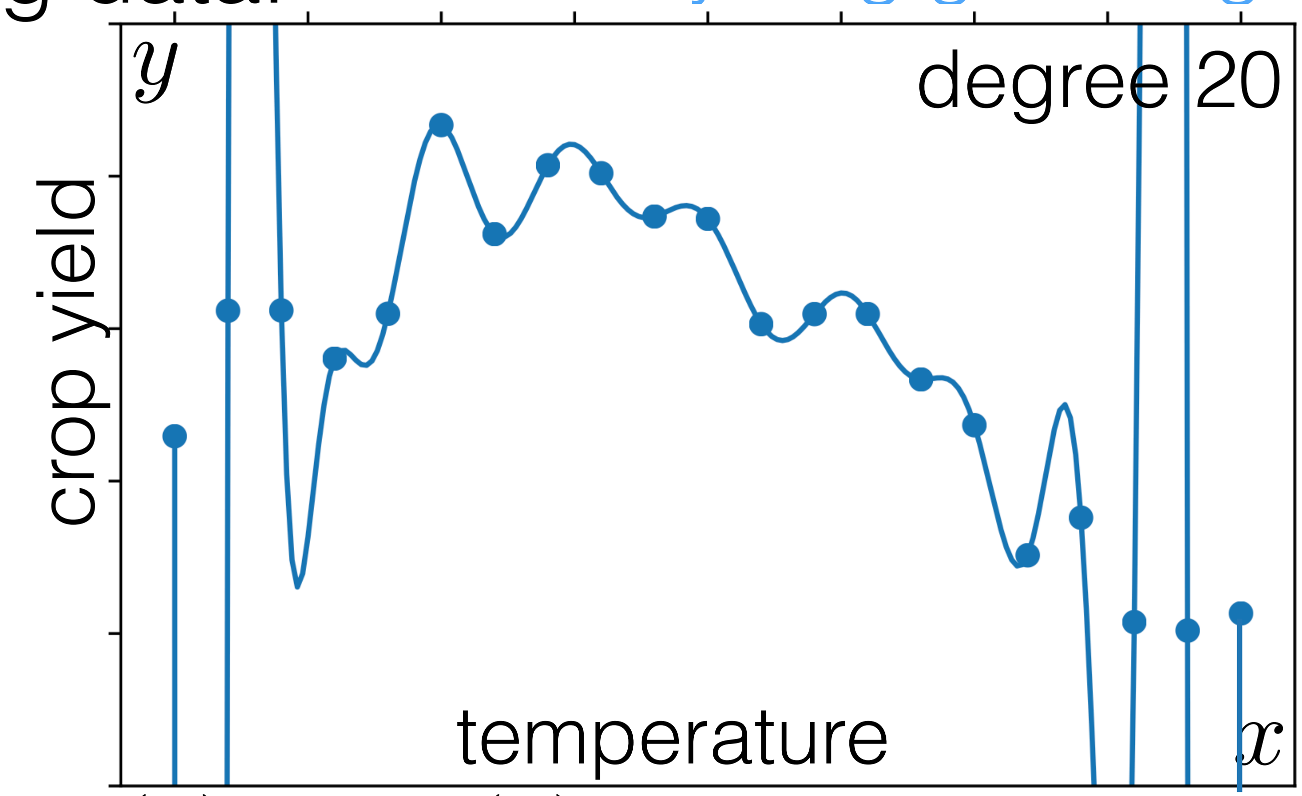
could anything go wrong?



- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero
- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N*
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \not\to \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$

Empirical risk over the training data

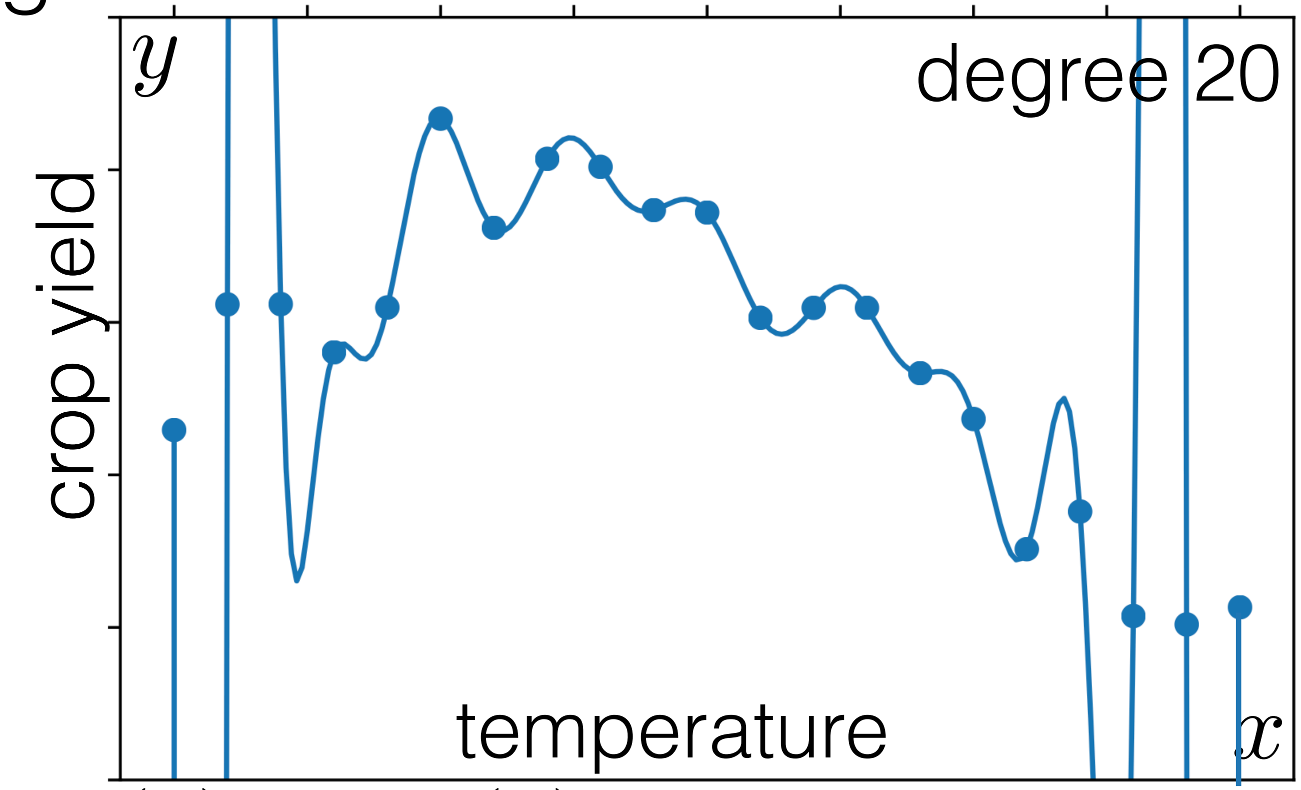
- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero



- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N*
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \not\to \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
- Which assumption isn't true?

Empirical risk over the training data

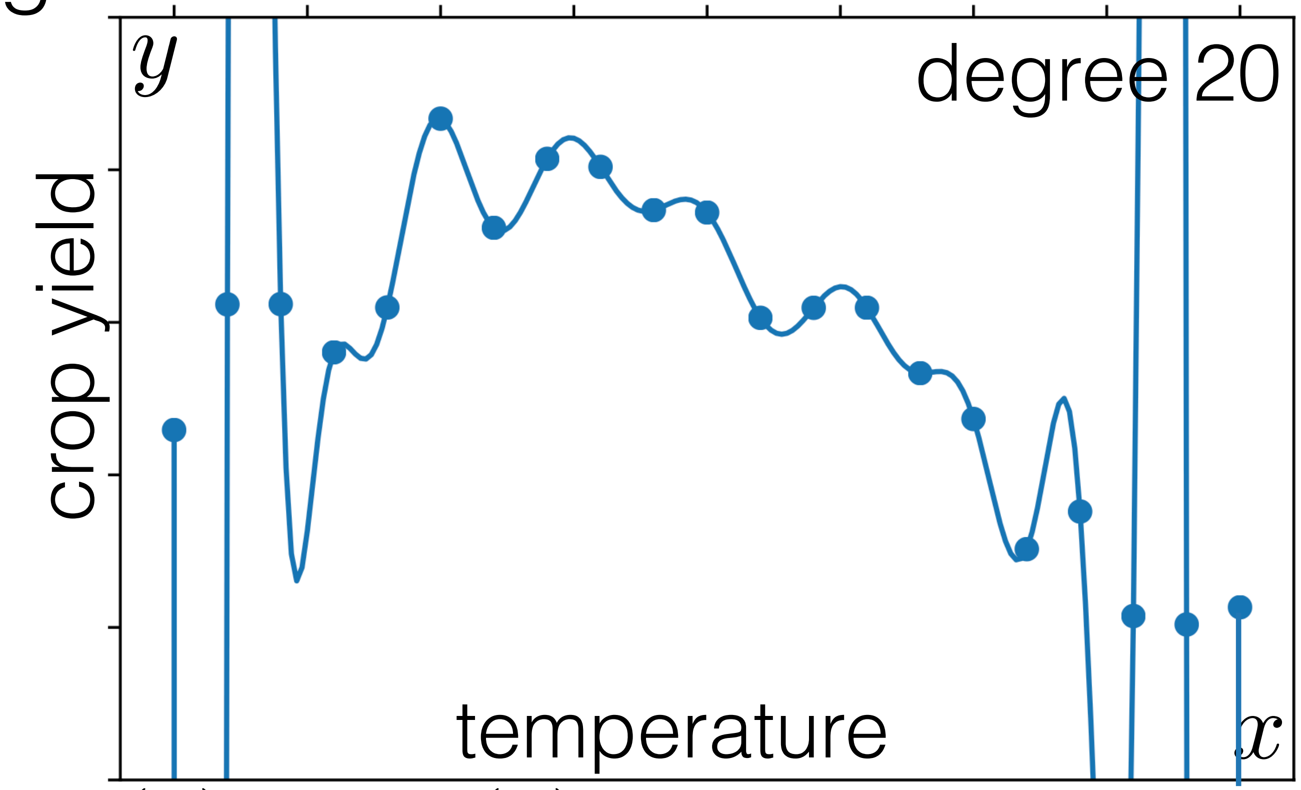
- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero



- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any N notation can be misleading!
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \not\to \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
- Which assumption isn't true?

Empirical risk over the training data

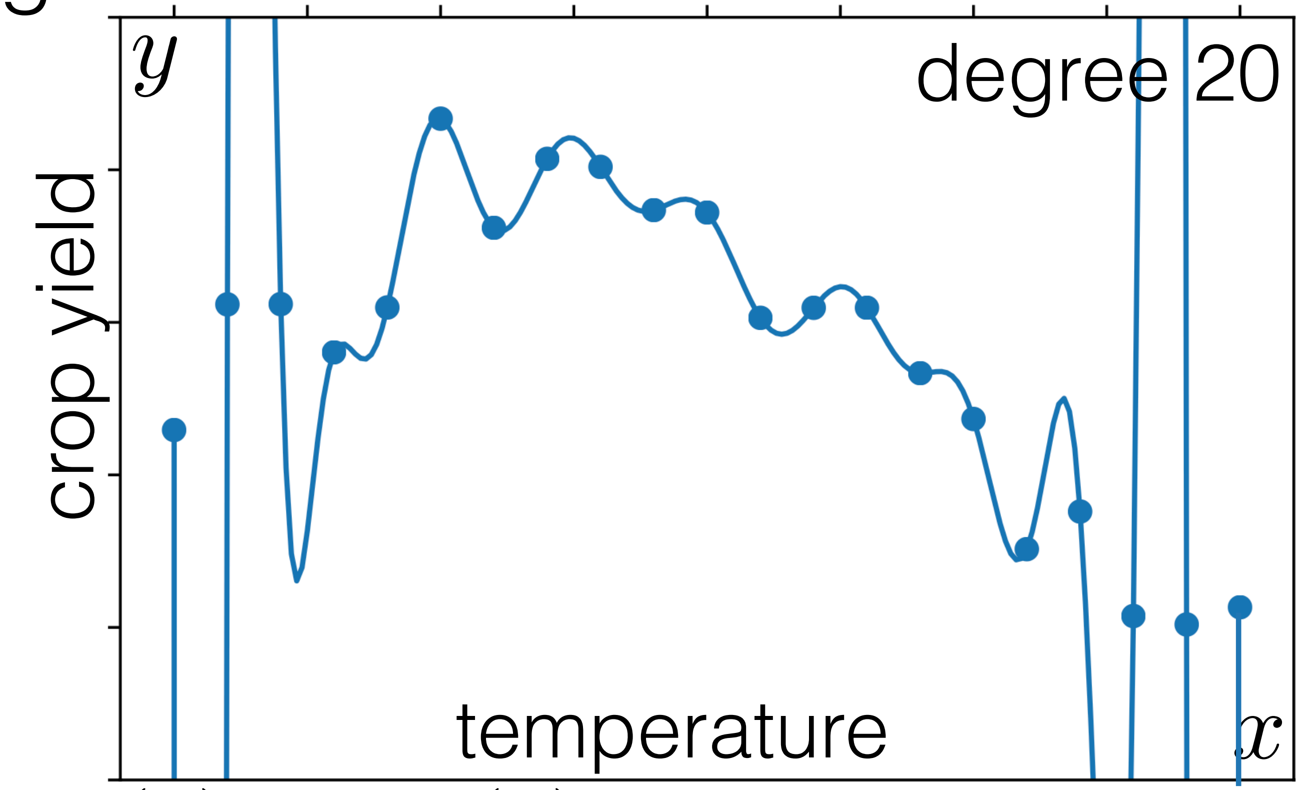
- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero



- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N* notation can be misleading!
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \not\to \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
- Which assumption isn't true? The terms in the sum aren't iid

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero

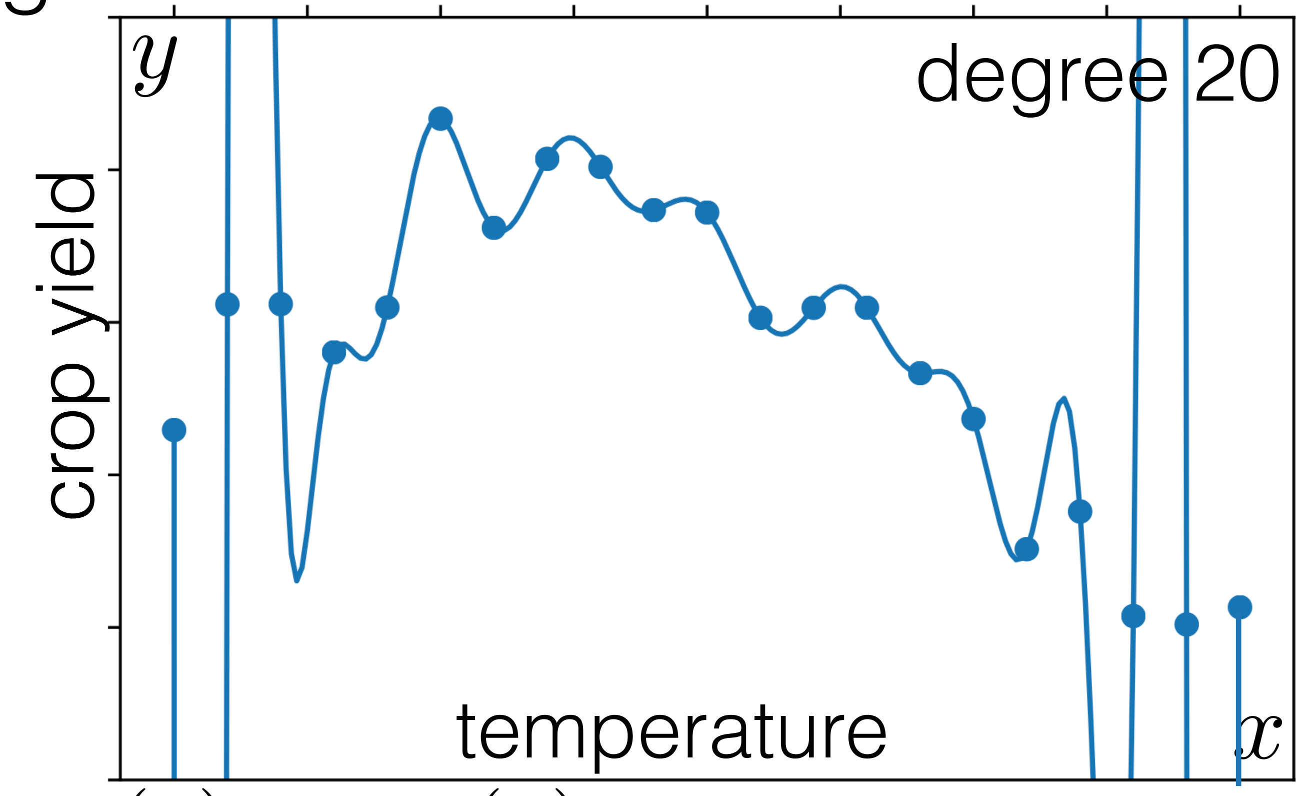


- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N* notation can be misleading!
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \not\to \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
- Which assumption isn't true? The terms in the sum aren't iid

$$L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) = f(X^{(n)}, Y^{(n)}, \mathcal{D} = \{(X^{(n')}, Y^{(n')})\}_{n'=1}^{N})$$

Empirical risk over the training data

- Proposal: Let's estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the training data: could anything go wrong?
  $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)}))$
- An example from polynomial regression where the estimate is 0, but the actual risk is (very) non-zero



- Is the problem that we don't have enough training data?
  - For any training point, let $h_{\mathcal{D}}(X^{(n)}) = Y^{(n)}$; else $h_{\mathcal{D}}(x) = 0$
  - Risk estimate is 0 for any *N* notation can be misleading!
  - So $\frac{1}{N}\sum_{n=1}^{N} L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) \nrightarrow \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$
- Which assumption isn't true? The terms in the sum aren't iid

$$L(Y^{(n)}, h_{\mathcal{D}}(X^{(n)})) = f(X^{(n)}, Y^{(n)}, \mathcal{D} = \{(X^{(n')}, Y^{(n')})\}_{n'=1}^{N})$$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(X^{(n)}, Y^{(n)}, \mathcal{D} = \{(X^{(n')}, Y^{(n')})\}_{n'=1}^{N})$$

=? (the Law of Large Numbers doesn't tell us)

Empirical risk over validation data

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$
- New proposal: estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the validation data (*not* the training data):

$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$
- New proposal: estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the validation data (*not* the training data):

$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$

- We can condition on the training data and conclude (if all of our other assumptions hold) that:

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)})) = \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$$

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$
- New proposal: estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the validation data (*not* the training data):
$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$
- We can condition on the training data and conclude (if all of our other assumptions hold) that:
$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)})) = \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$$
- Some options:

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$
- New proposal: estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the validation data (*not* the training data):
$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$
- We can condition on the training data and conclude (if all of our other assumptions hold) that:
$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)})) = \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$$
- Some options:
  A. Collect validation data separately from training data.

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$
- New proposal: estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the validation data (*not* the training data):
$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$
- We can condition on the training data and conclude (if all of our other assumptions hold) that:
$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)})) = \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$$
- Some options:
  A. Collect validation data separately from training data.
  B. Uniformly at random partition available data into training and validation data.

Empirical risk over validation data

- Idea: Let's get *M* **validation data** points that are totally separate from the training data: $\mathcal{D}' = \{(X^{(m)}, Y^{(m)})\}_{m=N+1}^{N+M}$
- New proposal: estimate $\mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$ with the empirical average of loss over the validation data (*not* the training data):
$$\frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)}))$$
- We can condition on the training data and conclude (if all of our other assumptions hold) that:
$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=N+1}^{N+M} L(Y^{(m)}, h_{\mathcal{D}}(X^{(m)})) = \mathbb{E}[L(Y, h_{\mathcal{D}}(X))]$$
- Some options:
  A. Collect validation data separately from training data.
  B. Uniformly at random partition available data into training and validation data.
- Note: can use validation data to estimate risk at a new data point even if the decision rule didn't arise from training data

Empirical risk over validation data

Empirical risk over validation data

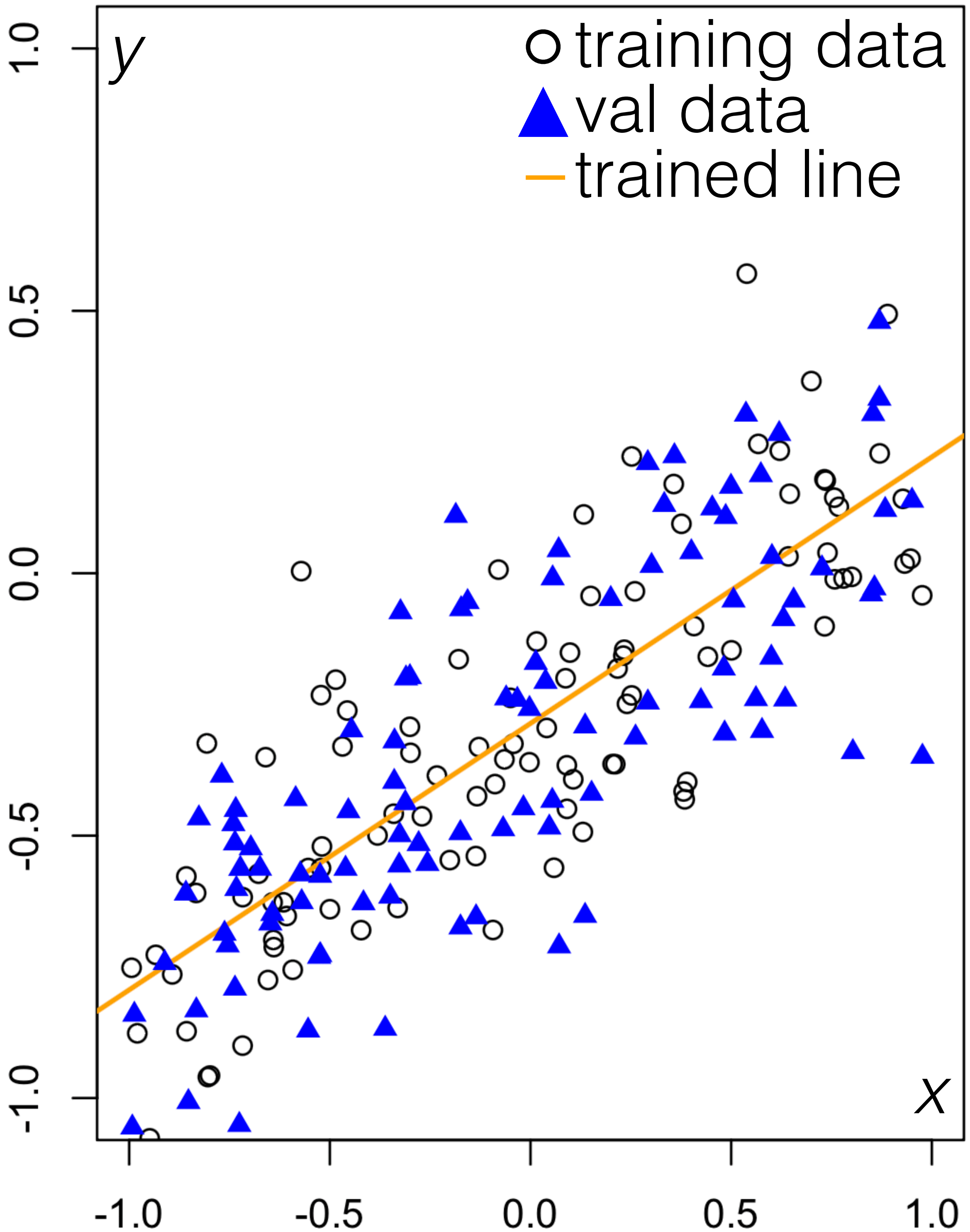
- Illustration: same setup as in demo from Lecture 6

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid

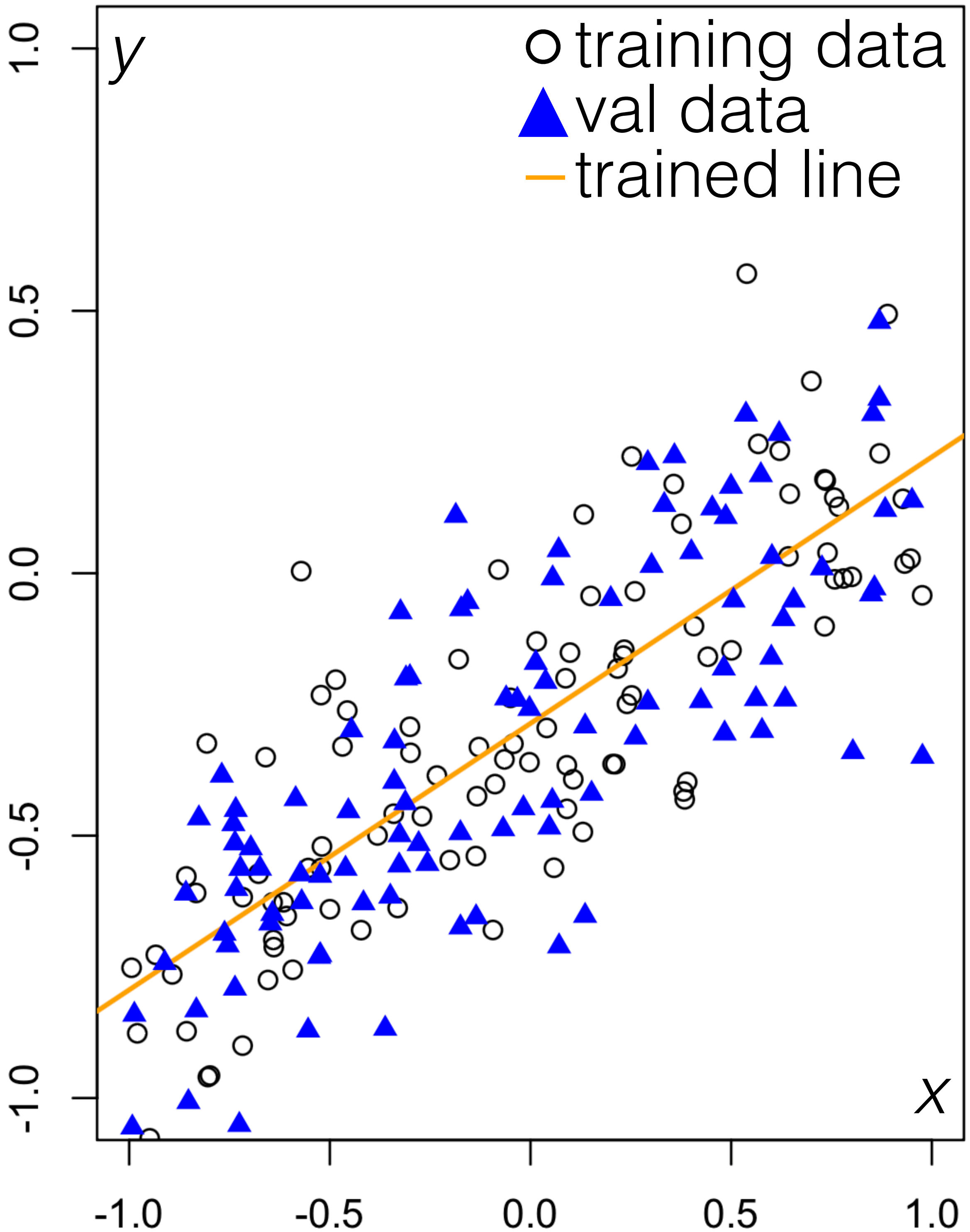
Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



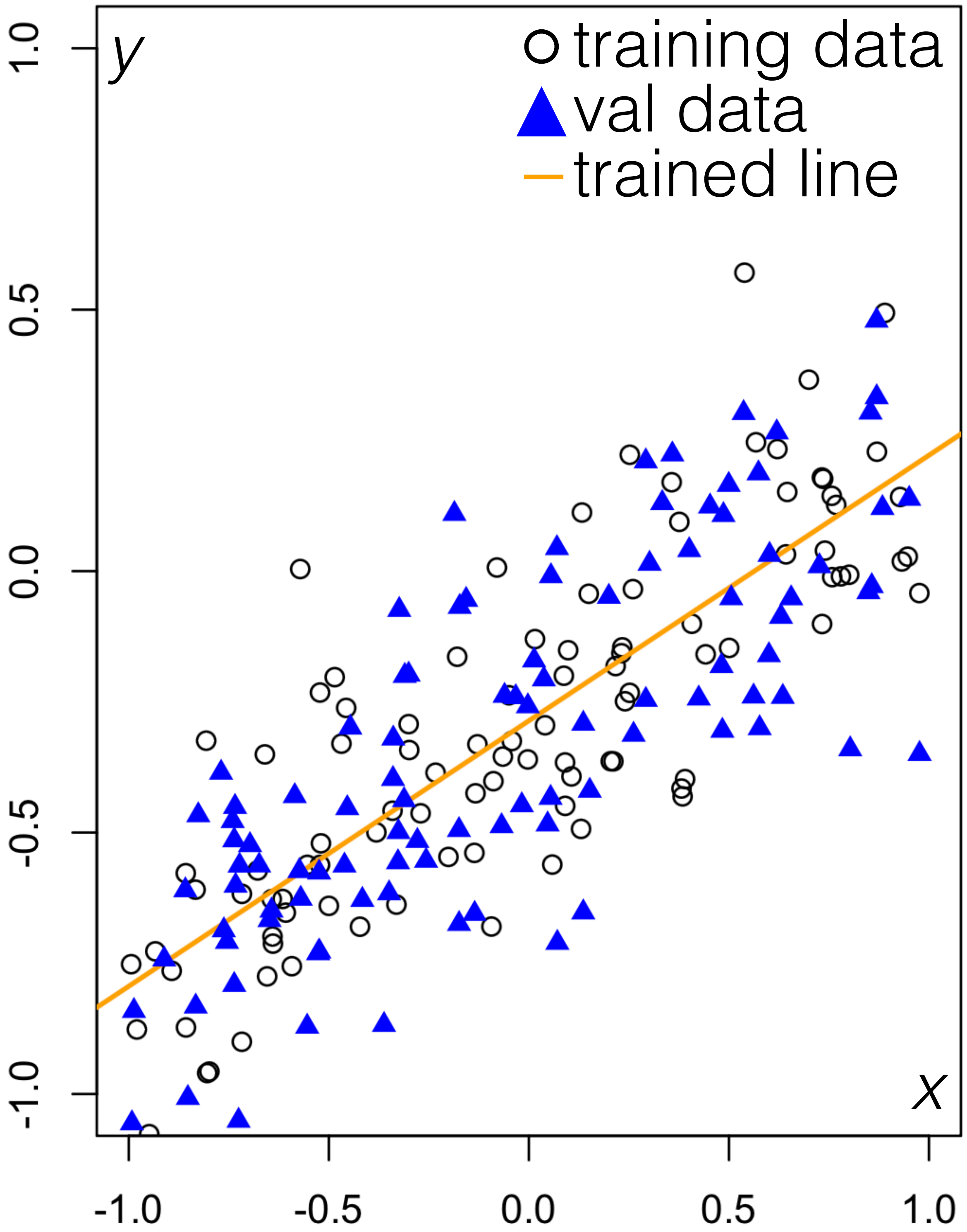
- I generated the data as:

$$X^{(n)} \overset{iid}{\sim} \text{Unif}[-1, 1]$$

$$(Y^{(n)} | X^{(n)} = x) \overset{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

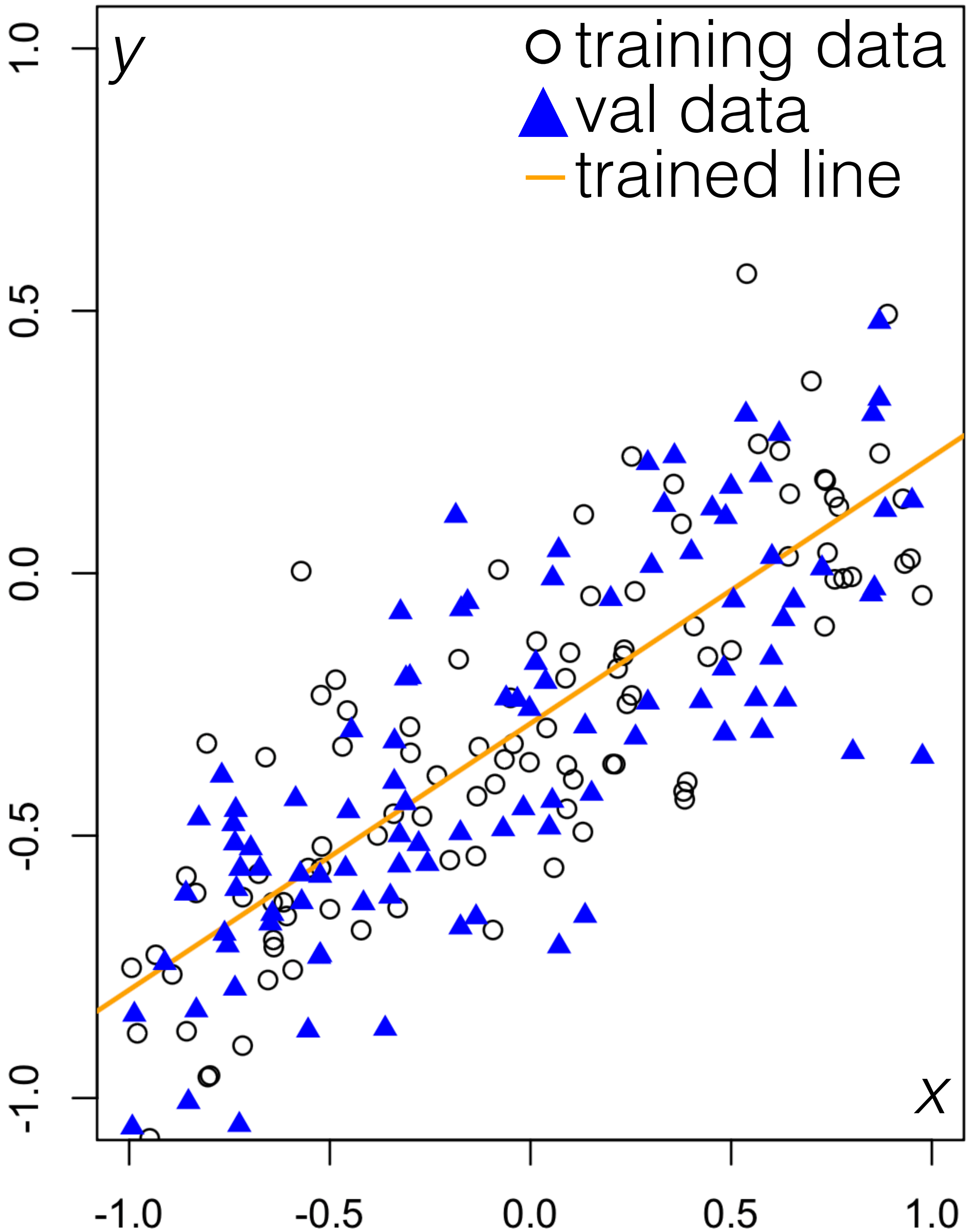
$$X^{(n)} \overset{iid}{\sim} \text{Unif}[-1, 1]$$

$$(Y^{(n)} | X^{(n)} = x) \overset{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

- Let's use square loss

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

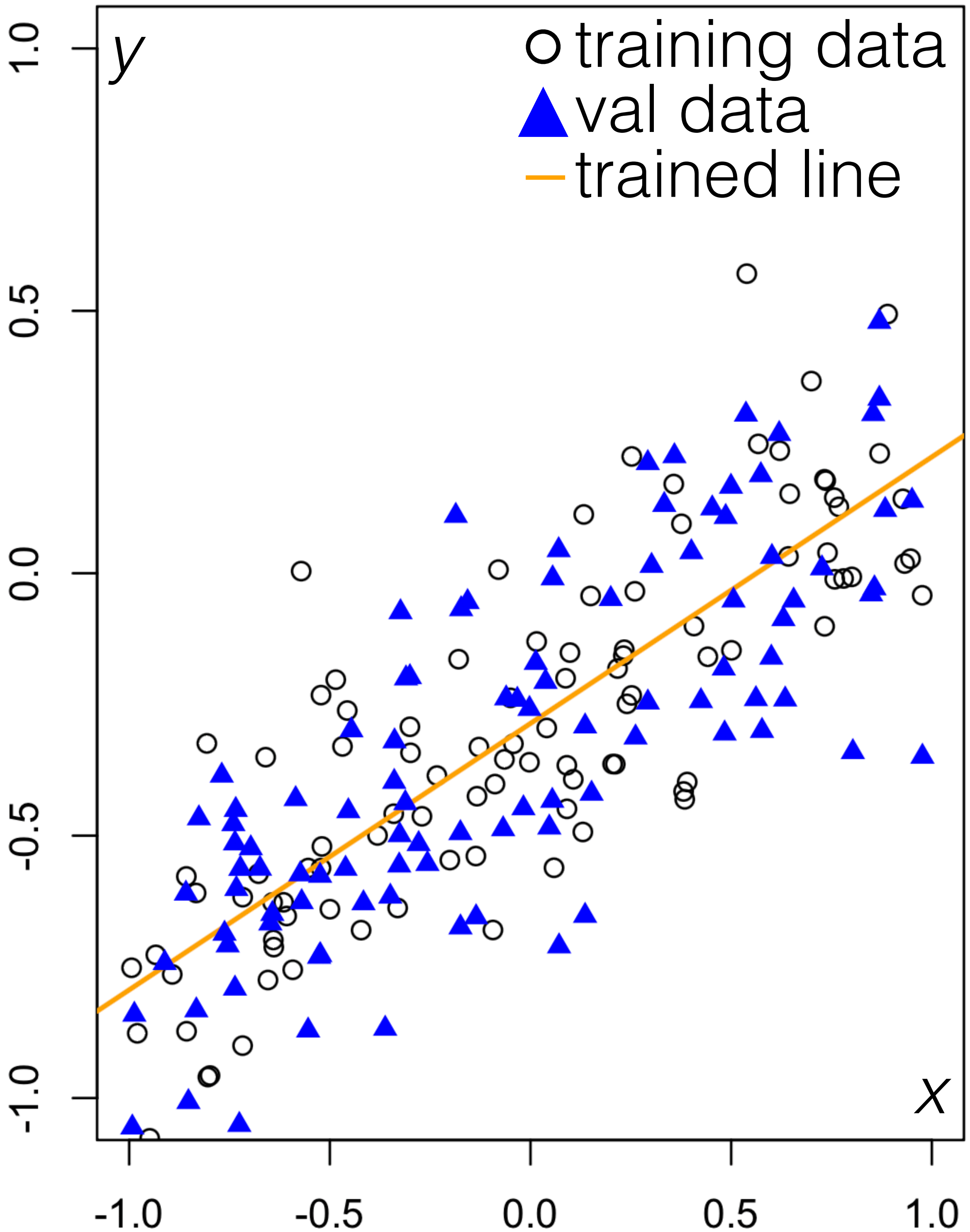
$$X^{(n)} \overset{iid}{\sim} \text{Unif}[-1, 1]$$

$$(Y^{(n)} | X^{(n)} = x) \overset{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

$$X^{(n)} \overset{iid}{\sim} \text{Unif}[-1, 1]$$

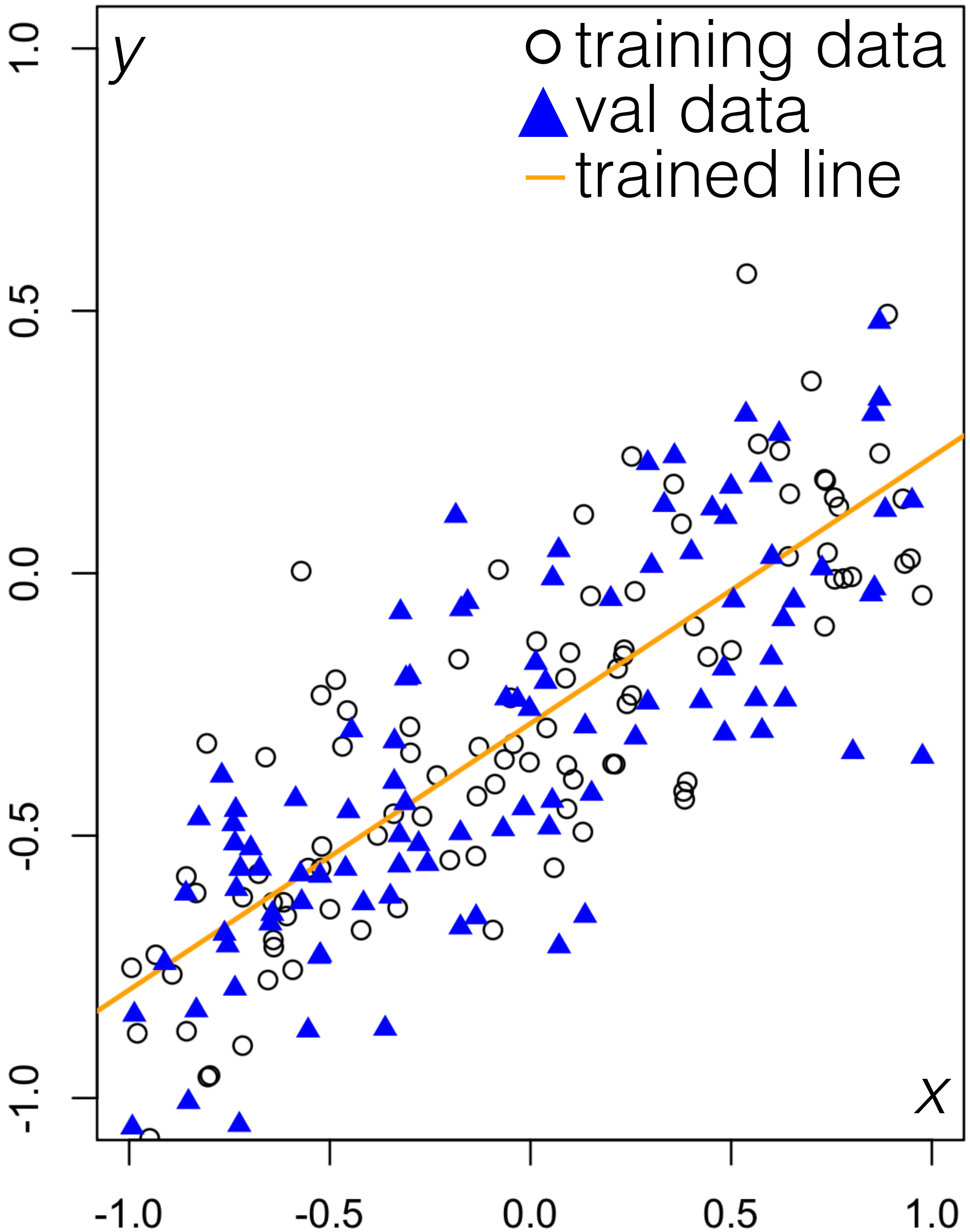
$$(Y^{(n)} | X^{(n)} = x) \overset{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

$$\mathbb{E}_{X,Y}[(Y - h_{\mathcal{D}}(X))^2]$$

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

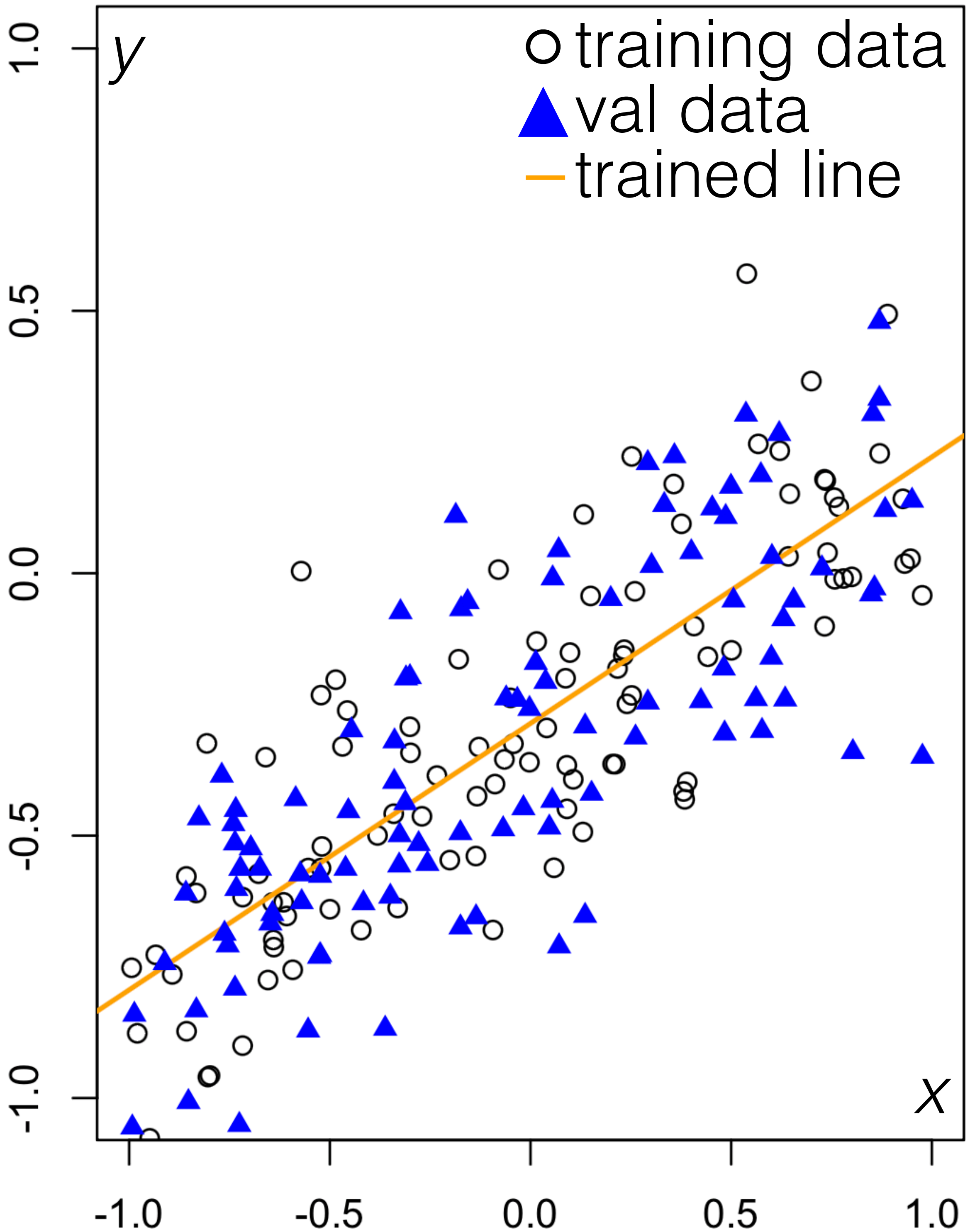
$$X^{(n)} \overset{iid}{\sim} \mathrm{Unif}[-1,1]$$
$$(Y^{(n)}|X^{(n)}=x) \overset{indep}{\sim} \mathcal{N}(-0.3+0.5x, 0.2^2)$$

- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

$$\mathbb{E}_{X,Y}[(Y - h_{\mathcal{D}}(X))^2]$$
$$= \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h_{\mathcal{D}}(X))^2]$$

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

$$X^{(n)} \overset{iid}{\sim} \mathrm{Unif}[-1, 1]$$

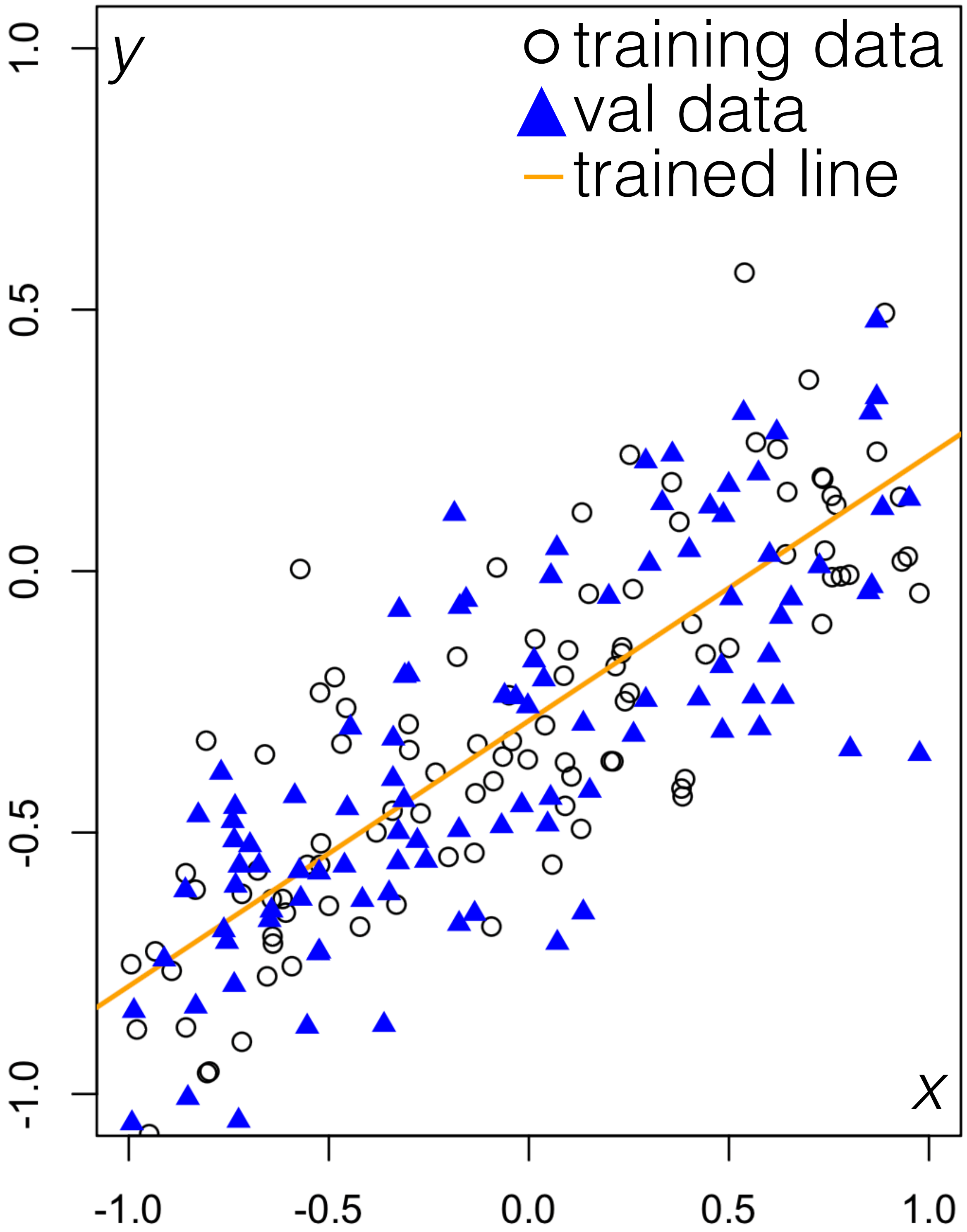
$$(Y^{(n)} | X^{(n)} = x) \overset{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

$$\mathbb{E}_{X,Y}[(Y - h_{\mathcal{D}}(X))^2]$$
$$= \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h_{\mathcal{D}}(X))^2]$$

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

$$X^{(n)} \stackrel{iid}{\sim} \mathrm{Unif}[-1, 1]$$
$$(Y^{(n)} | X^{(n)} = x) \stackrel{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

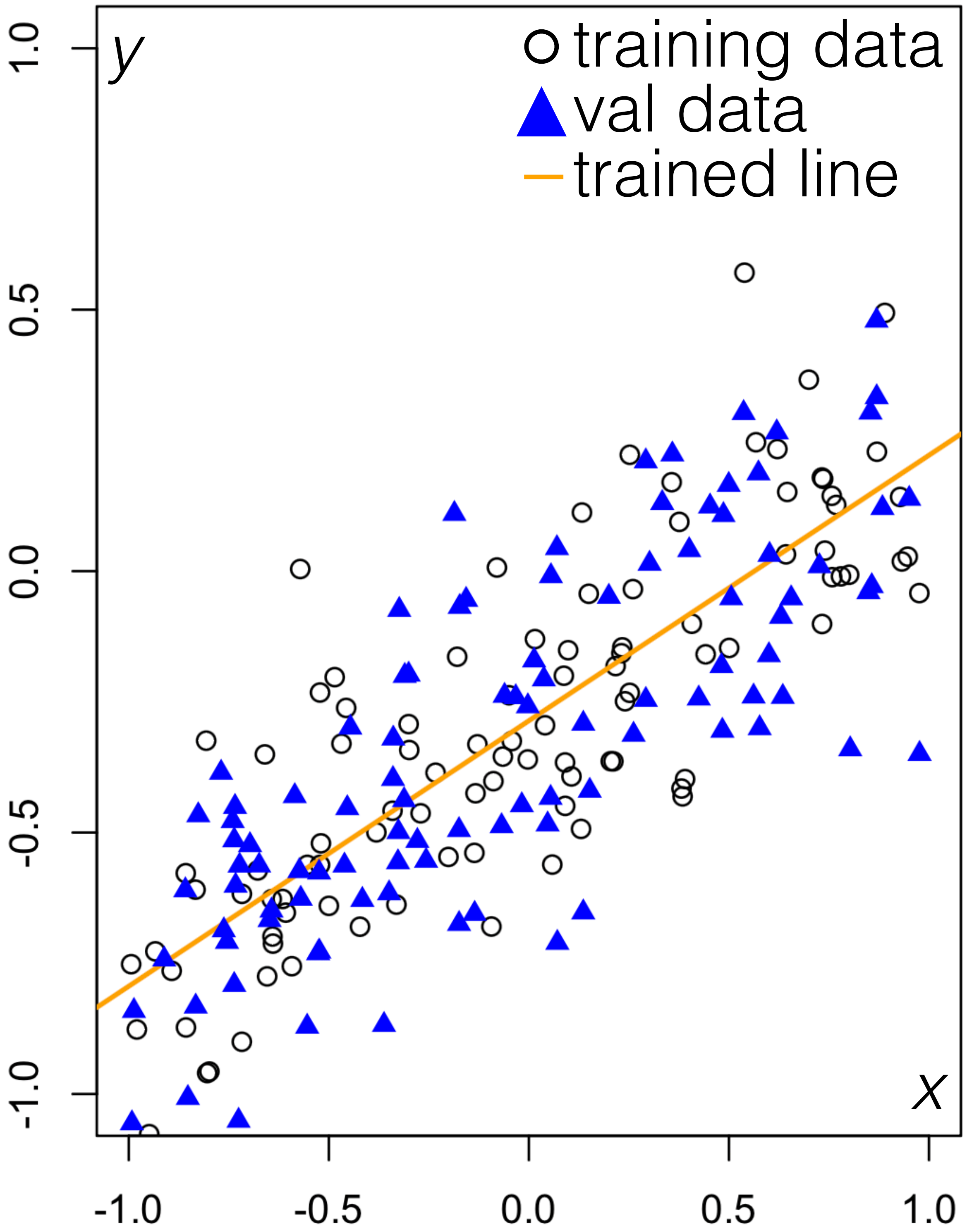
- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

$$\mathbb{E}_{X,Y}[(Y - h_{\mathcal{D}}(X))^2]$$
$$= \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h_{\mathcal{D}}(X))^2]$$
$$\approx \mathbb{E}_X \mathbb{E}_Z[Z^2]$$

with $Z \sim \mathcal{N}(0, 0.2^2)$

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

$$X^{(n)} \overset{iid}{\sim} \text{Unif}[-1, 1]$$

$$(Y^{(n)} | X^{(n)} = x) \overset{indep}{\sim} \mathcal{N}(-0.3 + 0.5x, 0.2^2)$$

- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

$$\mathbb{E}_{X,Y}[(Y - h_{\mathcal{D}}(X))^2]$$

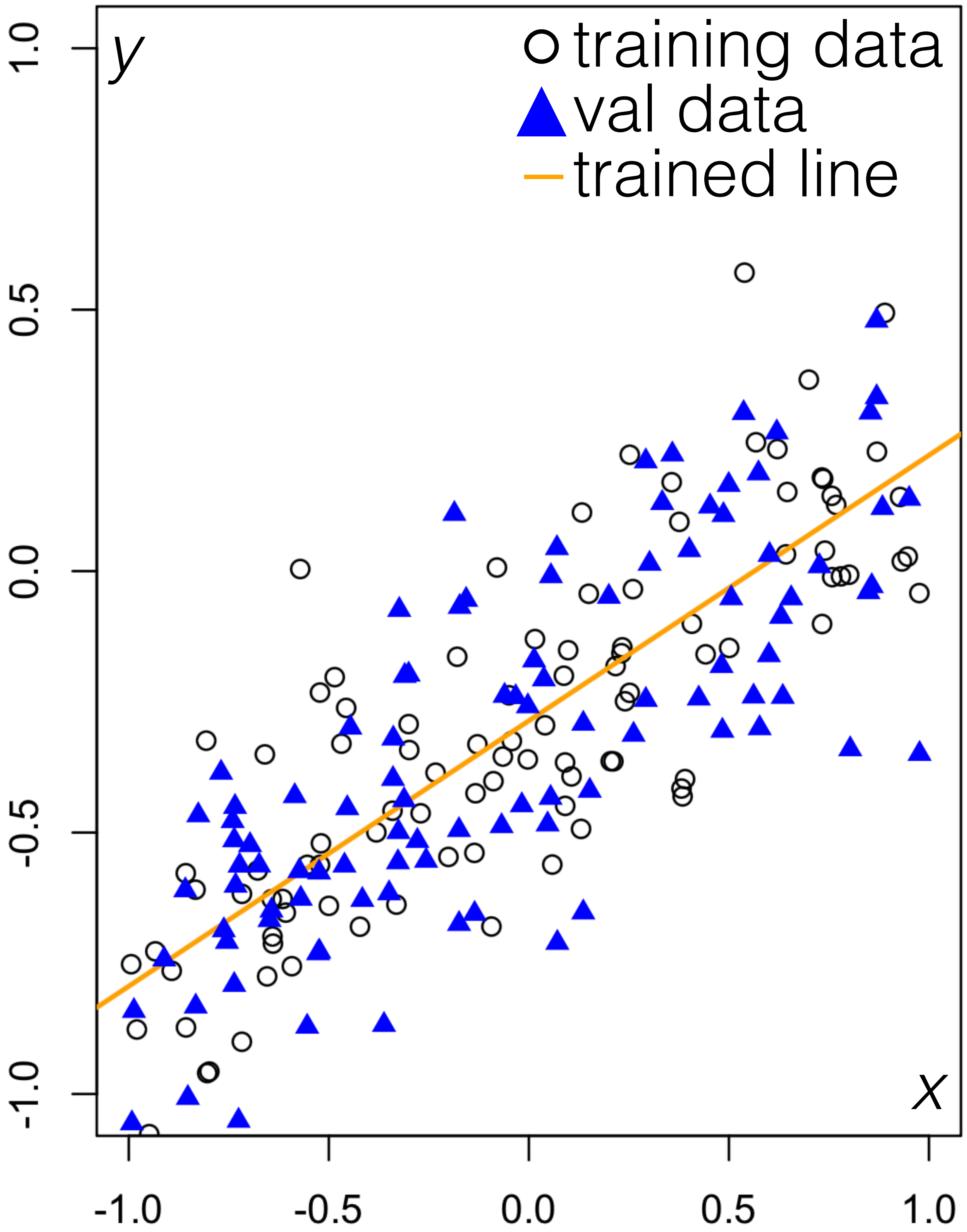
$$= \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h_{\mathcal{D}}(X))^2]$$

$$\approx \mathbb{E}_X \mathbb{E}_Z[Z^2]$$

with $Z \sim \mathcal{N}(0, 0.2^2)$

Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



- I generated the data as:

$$X^{(n)} \overset{iid}{\sim} \mathrm{Unif}[-1,1]$$

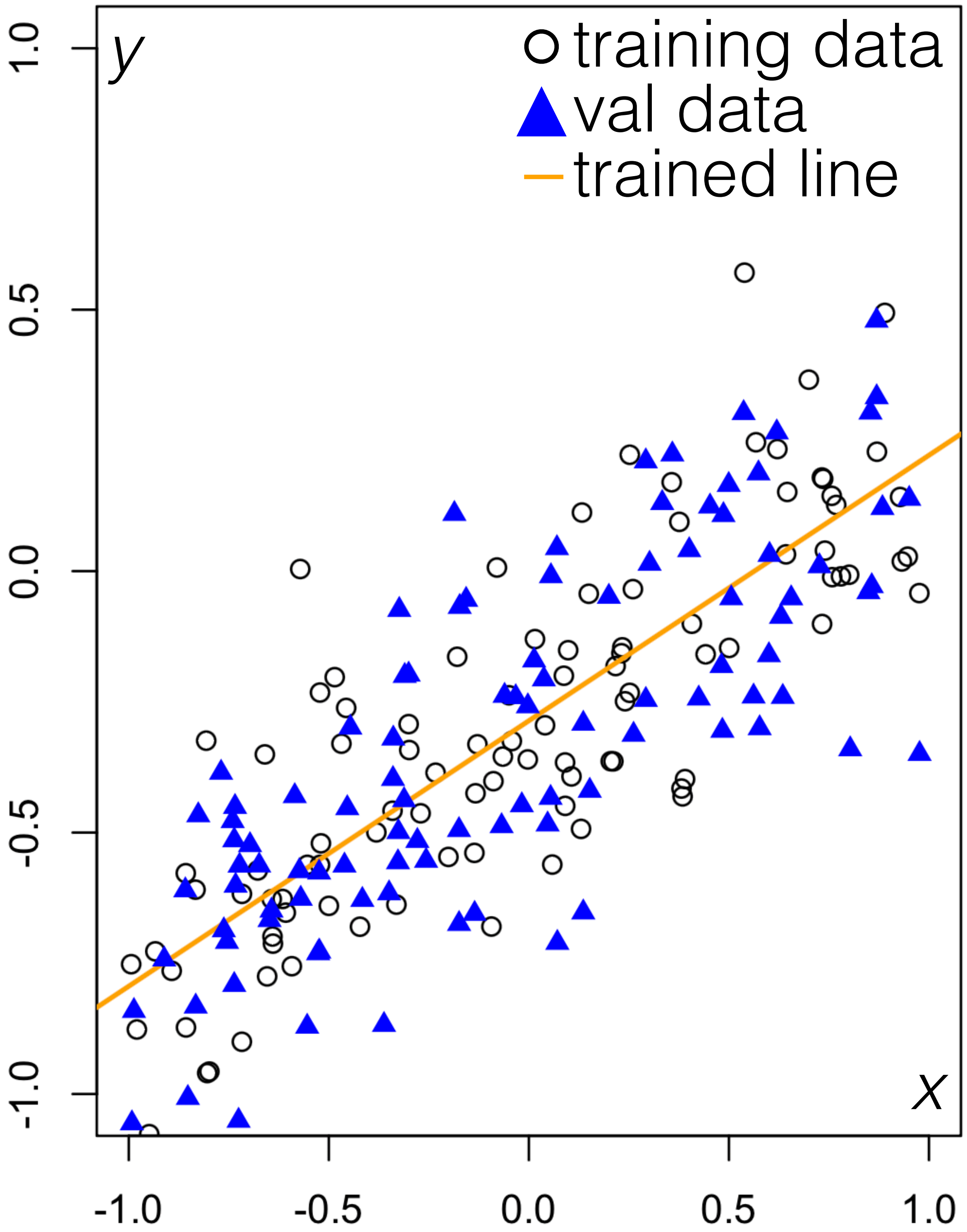
$$(Y^{(n)}|X^{(n)}=x) \overset{indep}{\sim} \mathcal{N}(-0.3+0.5x, 0.2^2)$$

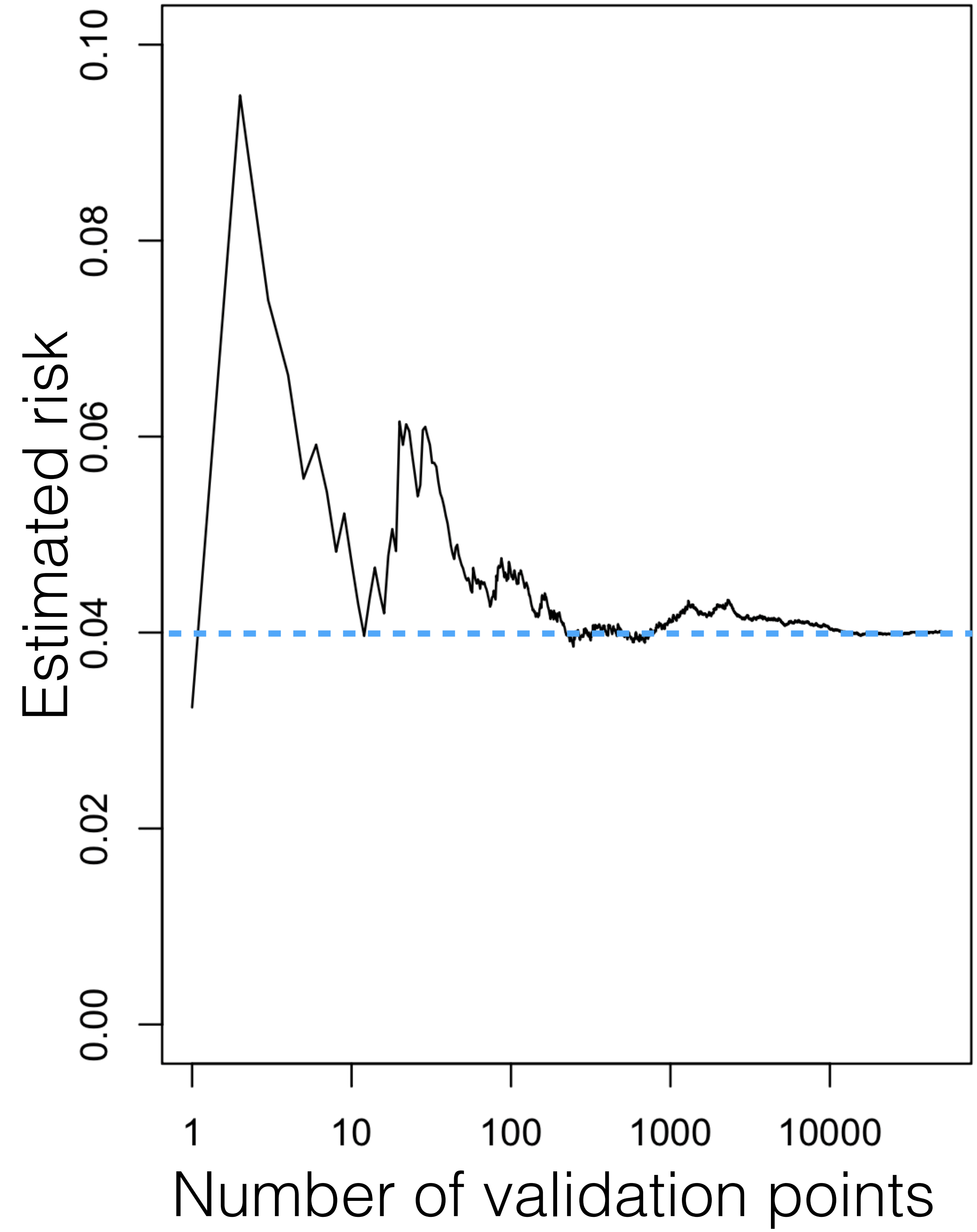
- Let's use square loss
- Approximately what is the true risk of new points generated the same way?

$$\mathbb{E}_{X,Y}[(Y-h_{\mathcal{D}}(X))^2]$$
$$= \mathbb{E}_X \mathbb{E}_{Y|X}[(Y-h_{\mathcal{D}}(X))^2]$$
$$\approx \mathbb{E}_X \mathbb{E}_Z[Z^2]$$
with $Z \sim \mathcal{N}(0, 0.2^2)$
$$= \mathbb{E}_X 0.2^2 = 0.2^2$$

Empirical risk over validation data

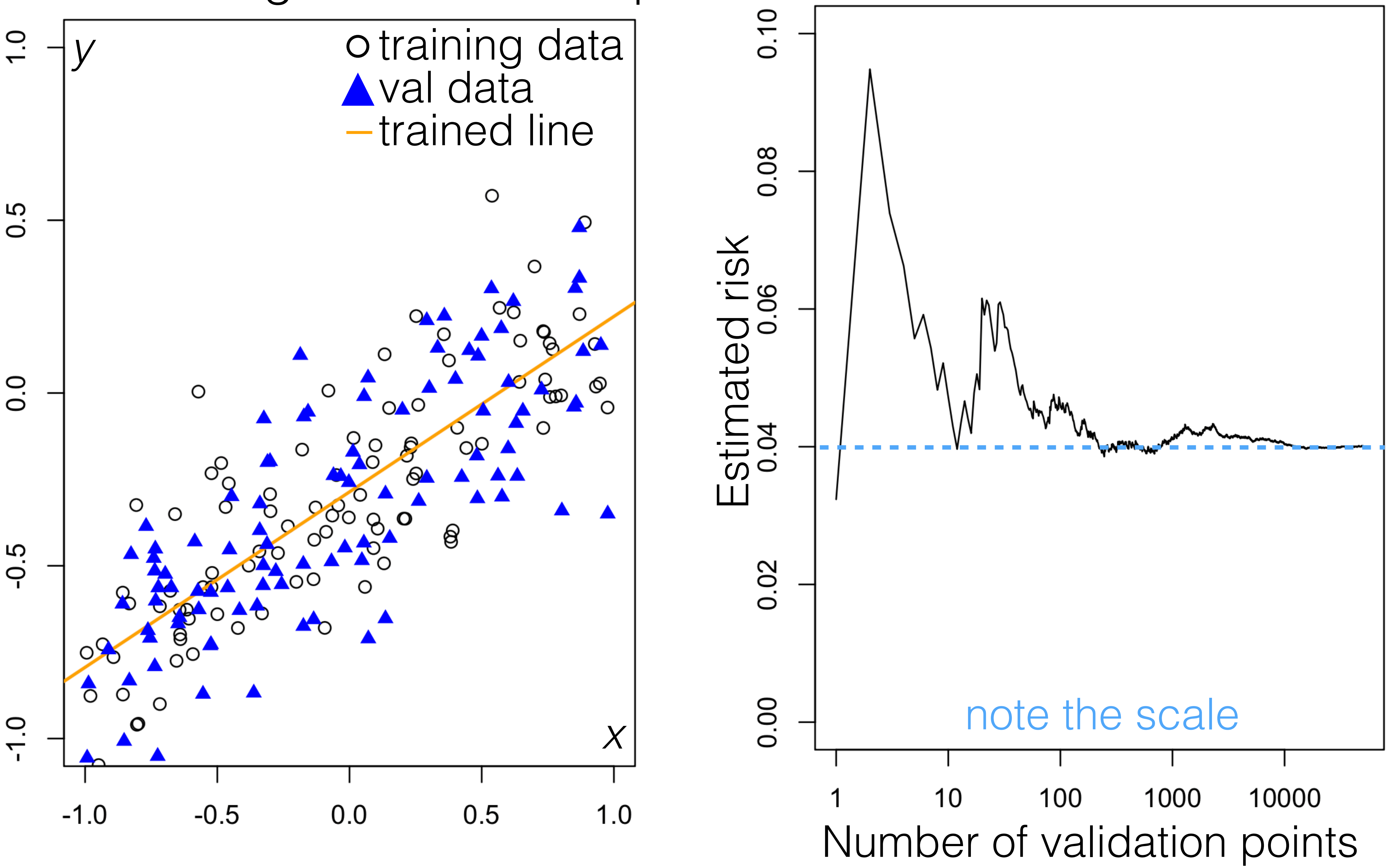
- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid





Empirical risk over validation data

- Illustration: same setup as in demo from Lecture 6
- Now training and validation points are iid



What’s the data distribution?

What’s the data distribution?

- I’m interested in predicting someone’s salary from their emails. Perhaps email-writing style predicts salary.

What’s the data distribution?

- I’m interested in predicting someone’s salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.

What’s the data distribution?

- I’m interested in predicting someone’s salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I’ve heard state-of-the-art machine learning needs a ton of data.

What’s the data distribution?

- I’m interested in predicting someone’s salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I’ve heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails.

What's the data distribution?

- I'm interested in predicting someone's salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I've heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the nth data point: I let x be one unique email from a volunteer and y be that person's salary. No x repeats.

What's the data distribution?

- I'm interested in predicting someone's salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I've heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the *n*th data point: I let *x* be one unique email from a volunteer and *y* be that person's salary. No *x* repeats.
- I uniformly at random partition my 100,000 points into training & validation sets. I use state-of-the-art ML to train $h_{\mathcal{D}}$.

What's the data distribution?

- I'm interested in predicting someone's salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I've heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the *n*th data point: I let *x* be one unique email from a volunteer and *y* be that person's salary. No *x* repeats.
- I uniformly at random partition my 100,000 points into training & validation sets. I use state-of-the-art ML to train $h_{\mathcal{D}}$.
- I report the empirical average of loss over the validation data.

What's the data distribution?

- I'm interested in predicting someone's salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I've heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the nth data point: I let x be one unique email from a volunteer and y be that person's salary. No x repeats.
- I uniformly at random partition my 100,000 points into training & validation sets. I use state-of-the-art ML to train $h_{\mathcal{D}}$.
- I report the empirical average of loss over the validation data.

Is this a good way to evaluate my predictor?

What's the data distribution?

- I'm interested in predicting someone's salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I've heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the *n*th data point: I let *x* be one unique email from a volunteer and *y* be that person's salary. No *x* repeats.
- I uniformly at random partition my 100,000 points into training & validation sets. I use state-of-the-art ML to train $h_{\mathcal{D}}$.
- I report the empirical average of loss over the validation data.
- The training & validation data might be seen as roughly iid with distribution: (1) choose one of my existing volunteers with probability 1/100 & (2) generate an email from that person

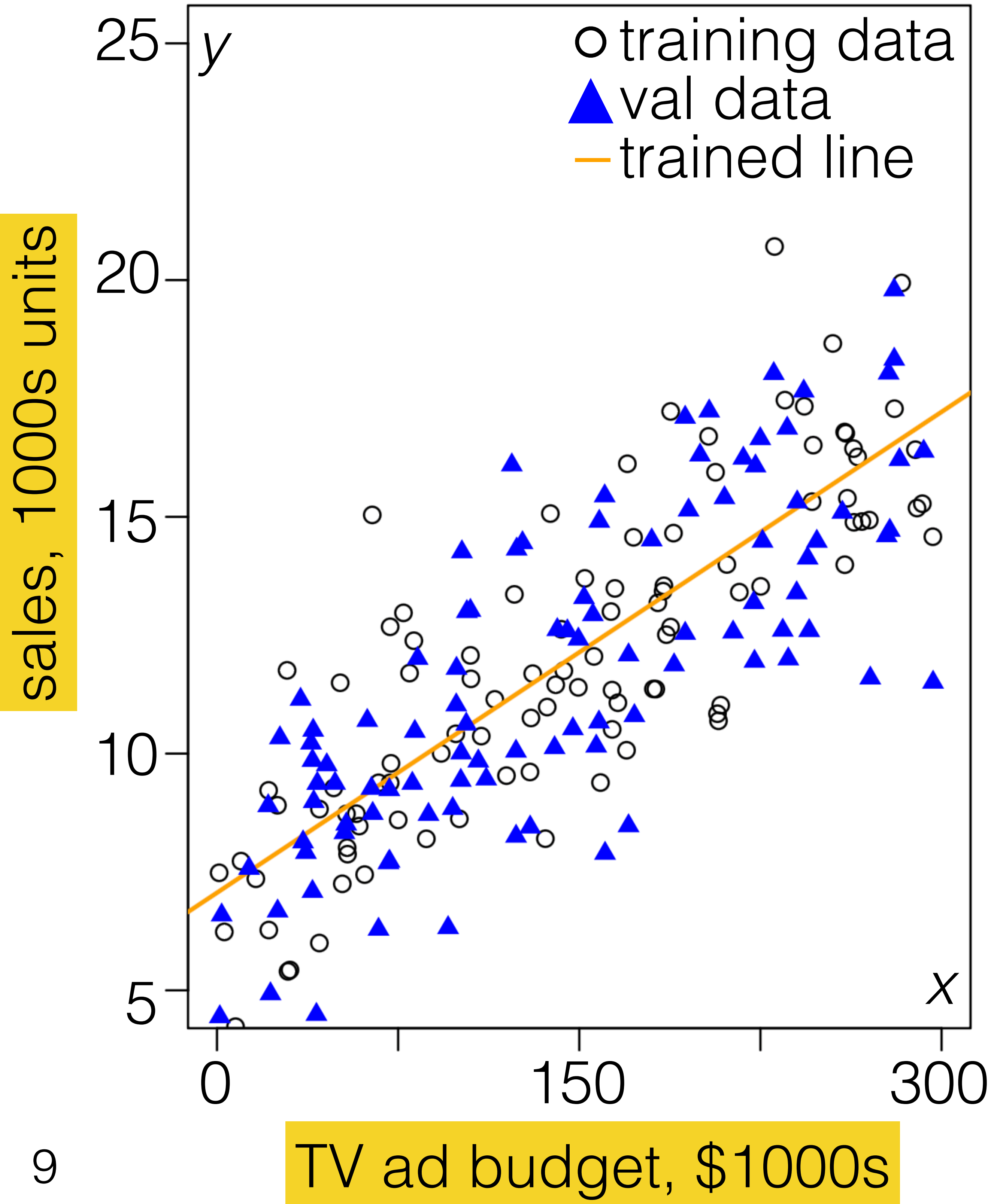
What's the data distribution?

- I'm interested in predicting someone's salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I've heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the *n*th data point: I let *x* be one unique email from a volunteer and *y* be that person's salary. No *x* repeats.
- I uniformly at random partition my 100,000 points into training & validation sets. I use state-of-the-art ML to train $h_{\mathcal{D}}$.
- I report the empirical average of loss over the validation data.
- The training & validation data might be seen as roughly iid with distribution: (1) choose one of my existing volunteers with probability 1/100 & (2) generate an email from that person
  - I might get low validation risk by: take whichever name appears in the email signature and then report their salary

What’s the data distribution?

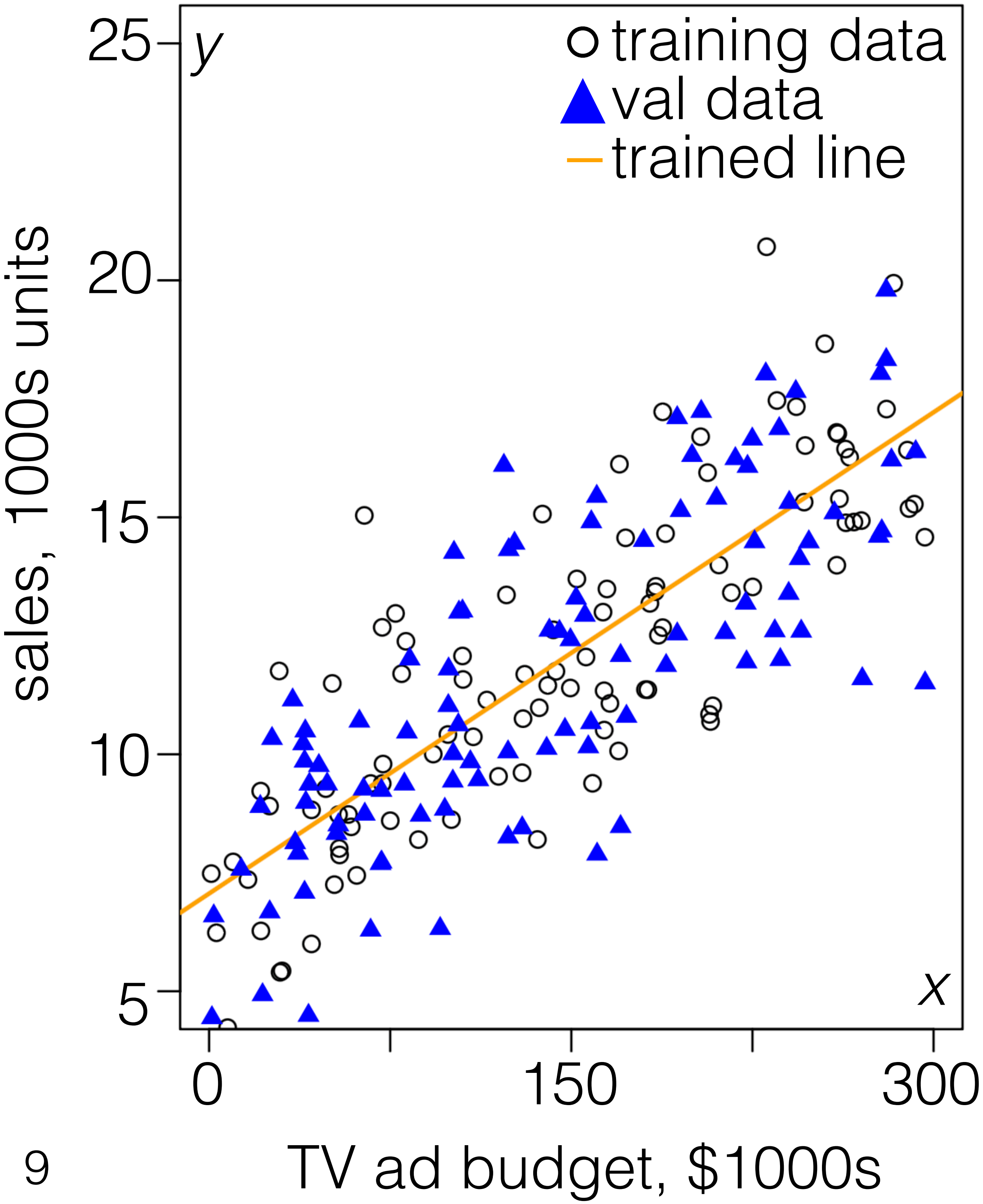
- I’m interested in predicting someone’s salary from their emails. Perhaps email-writing style predicts salary.
- I put in a *ton* of work and found 100 volunteers.
- I’ve heard state-of-the-art machine learning needs a ton of data. So for each volunteer, I use 1000 separate/unique emails. To make the *n*th data point: I let *x* be one unique email from a volunteer and *y* be that person’s salary. No *x* repeats.
- I uniformly at random partition my 100,000 points into training & validation sets. I use state-of-the-art ML to train $h_{\mathcal{D}}$.
- I report the empirical average of loss over the validation data.
- The training & validation data might be seen as roughly iid with distribution: (1) choose one of my existing volunteers with probability 1/100 & (2) generate an email from that person
  - I might get low validation risk by: take whichever name appears in the email signature and then report their salary
- I care about the risk (expected loss) for a new person.

What's the data distribution?



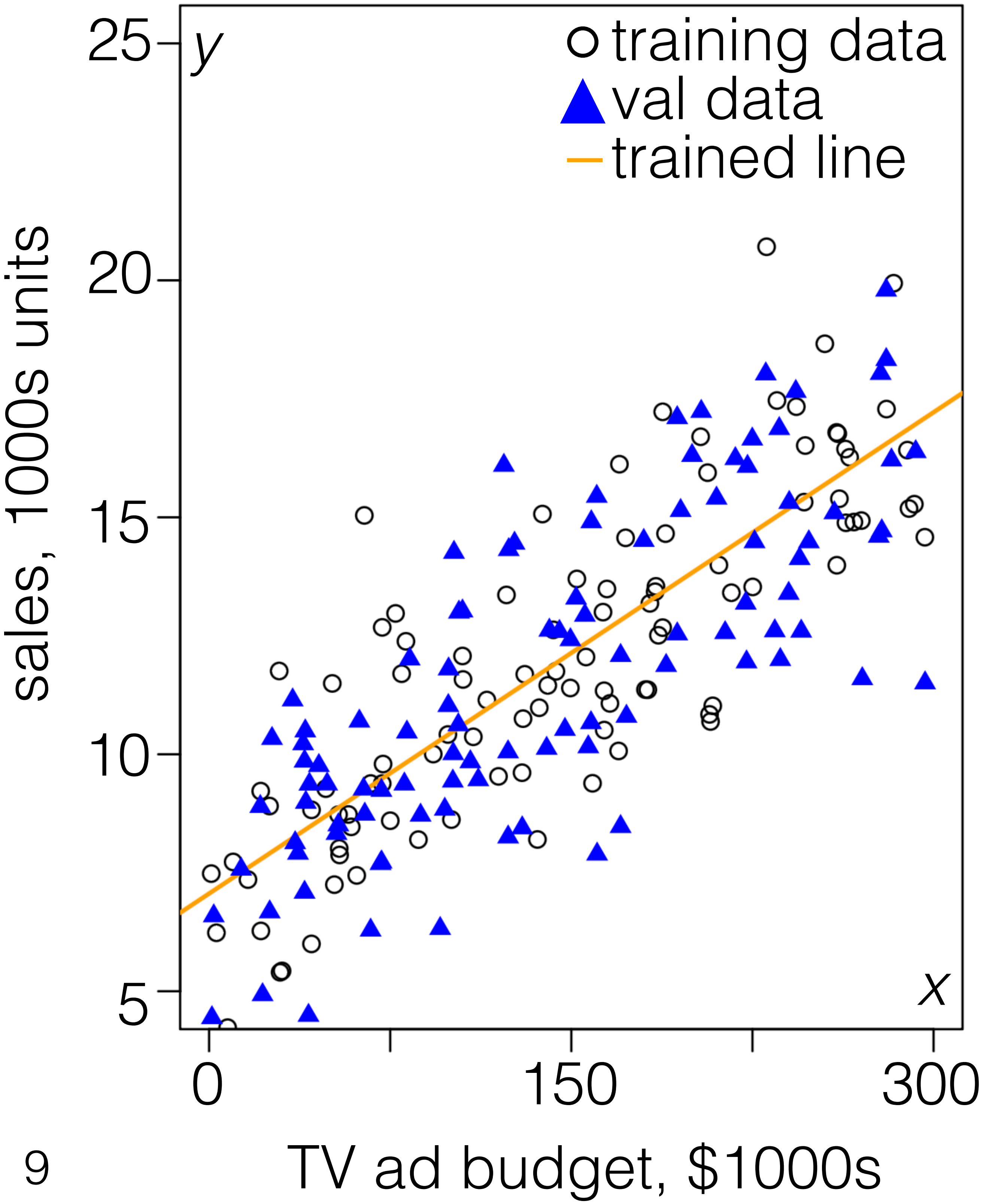
What's the data distribution?

- My boss wants to know the expected loss of my prediction for TV ad budgets in the $300,000 to $400,000 range



What's the data distribution?

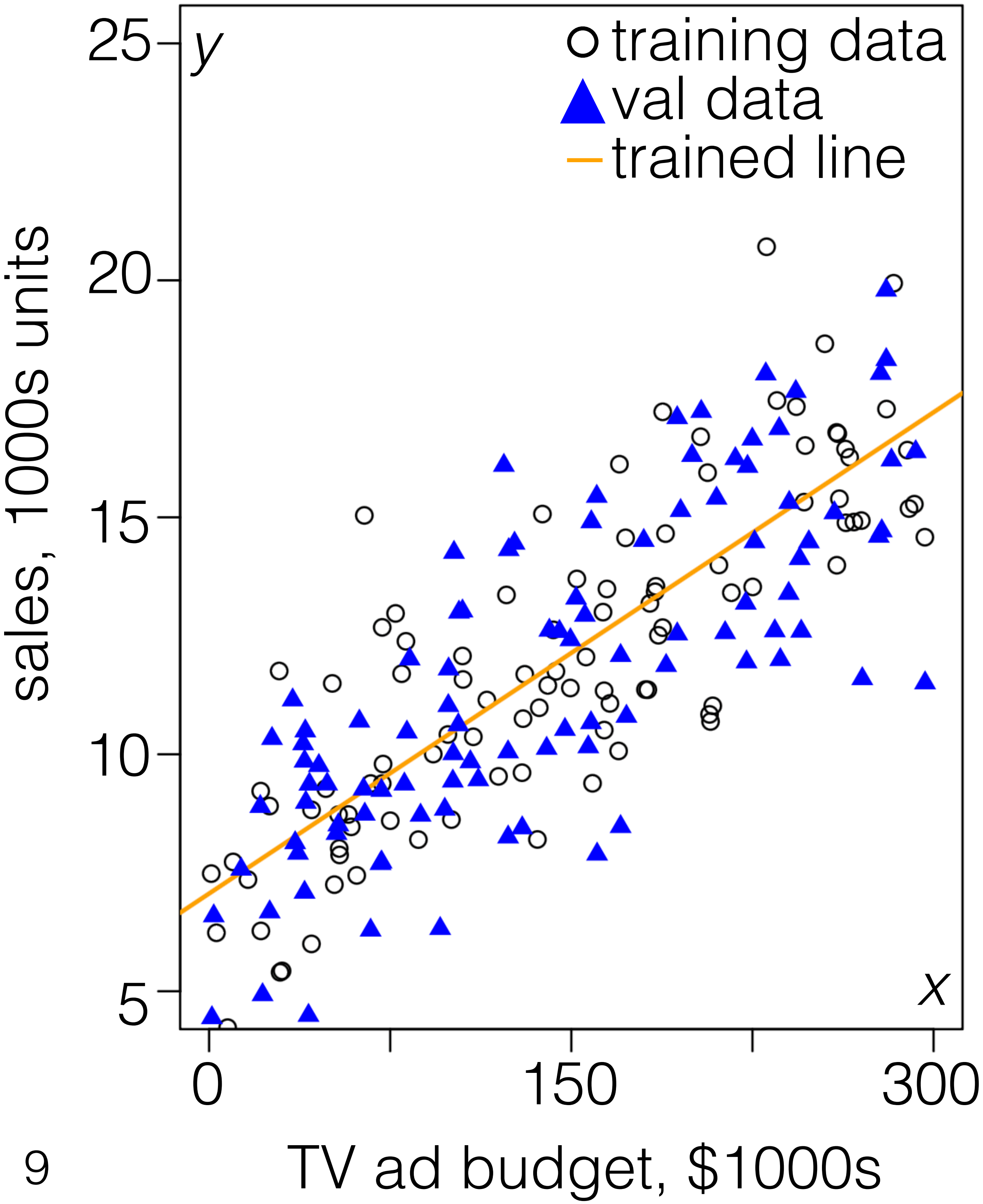
- My boss wants to know the expected loss of my prediction for TV ad budgets in the \$300,000 to \$400,000 range



- Does the empirical average of the loss over validation data estimate this value?

What's the data distribution?

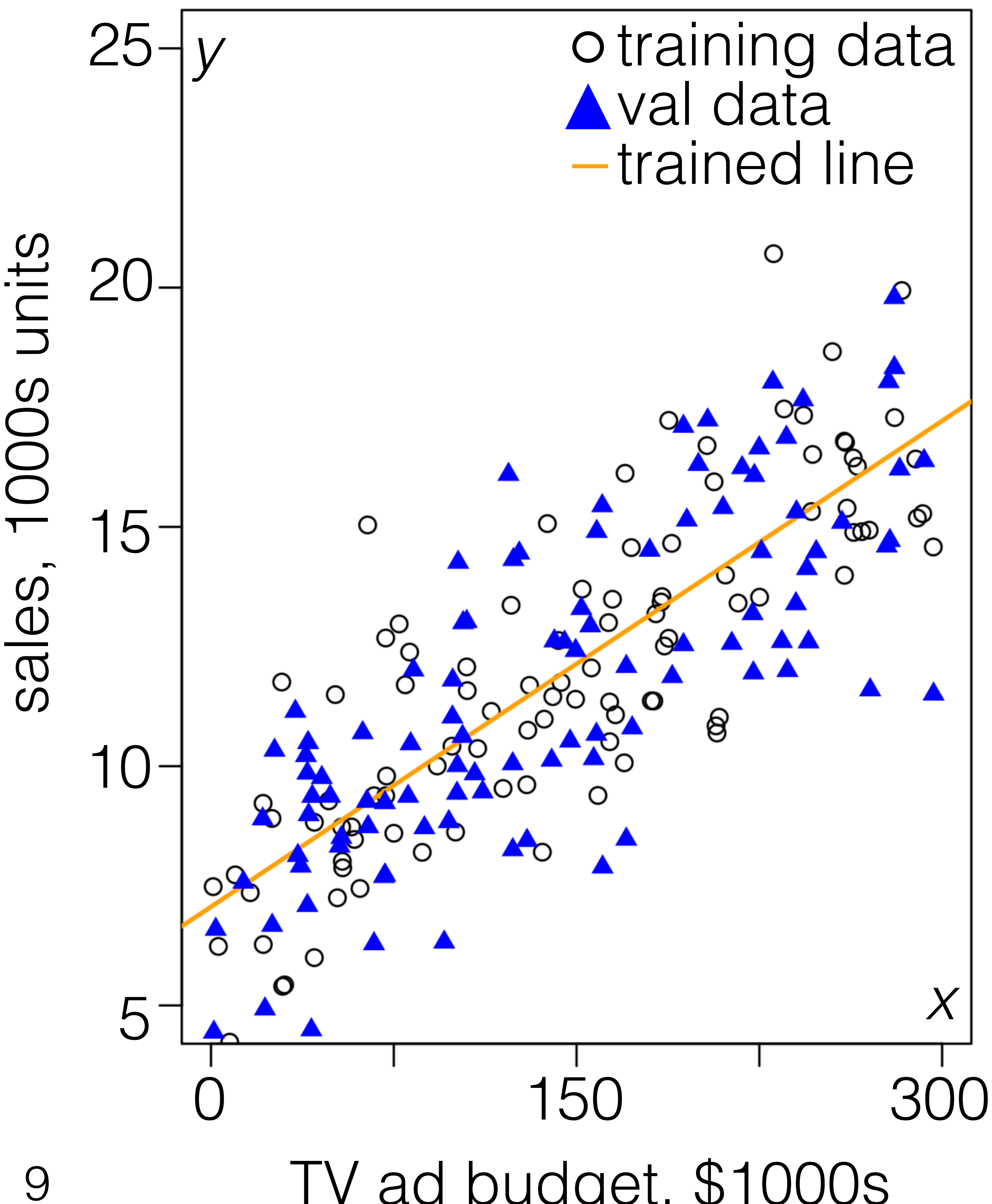
- My boss wants to know the expected loss of my prediction for TV ad budgets in the $300,000 to $400,000 range



- Does the empirical average of the loss over validation data estimate this value?
- The *x* values my boss proposes aren't from the same distribution as the validation *x*

What's the data distribution?

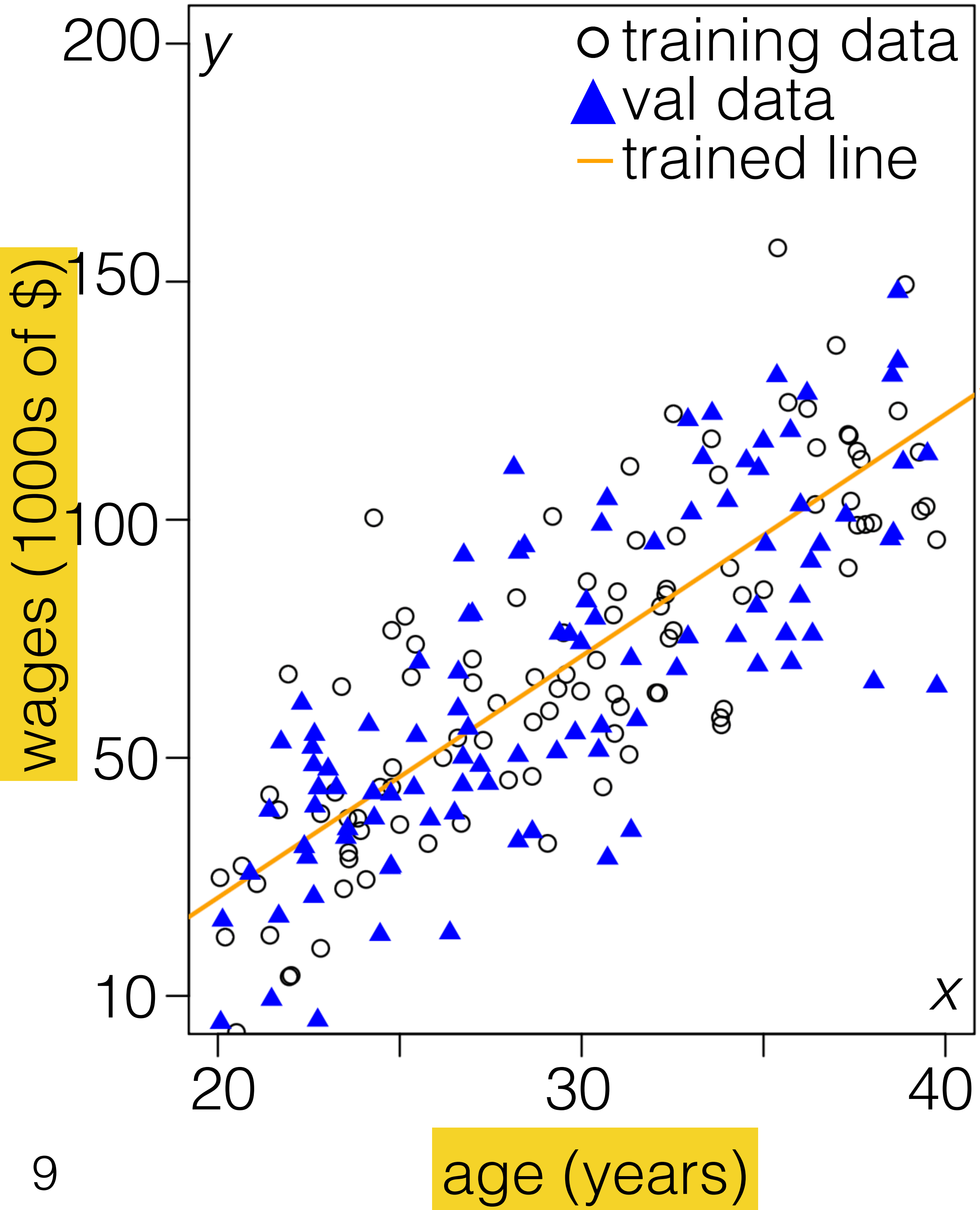
- My boss wants to know the expected loss of my prediction for TV ad budgets in the \$300,000 to \$400,000 range



- Does the empirical average of the loss over validation data estimate this value?
- The *x* values my boss proposes aren't from the same distribution as the validation *x*
- Aside: It's difficult to predict or evaluate outside the observed data without making any other assumptions

What's the data distribution?

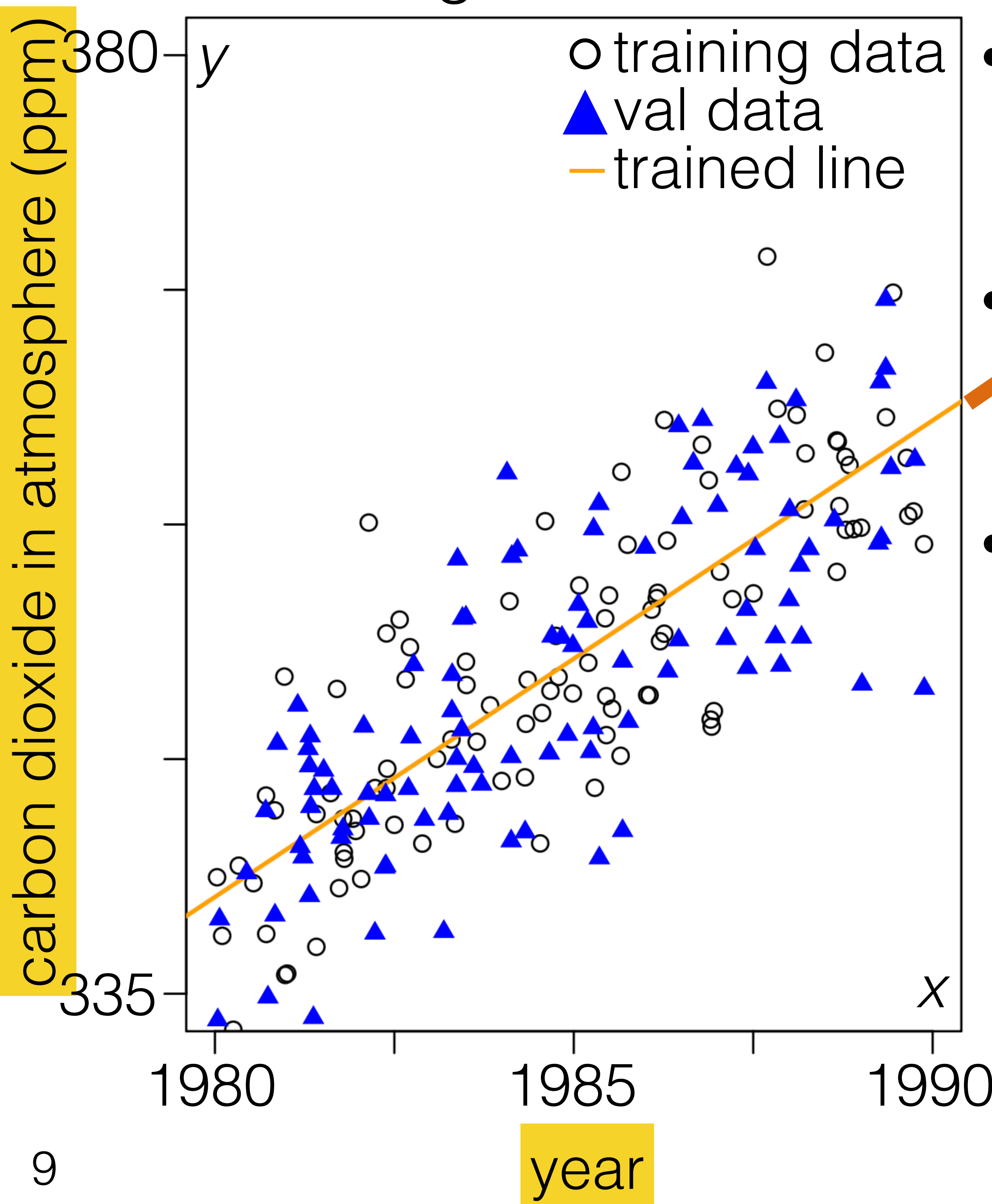
- My boss wants to know the expected loss of my prediction for TV ad budgets in the \$300,000 to \$400,000 range



- Does the empirical average of the loss over validation data estimate this value?
- The x values my boss proposes aren't from the same distribution as the validation x
- Aside: It's difficult to predict or evaluate outside the observed data without making any other assumptions

What's the data distribution?

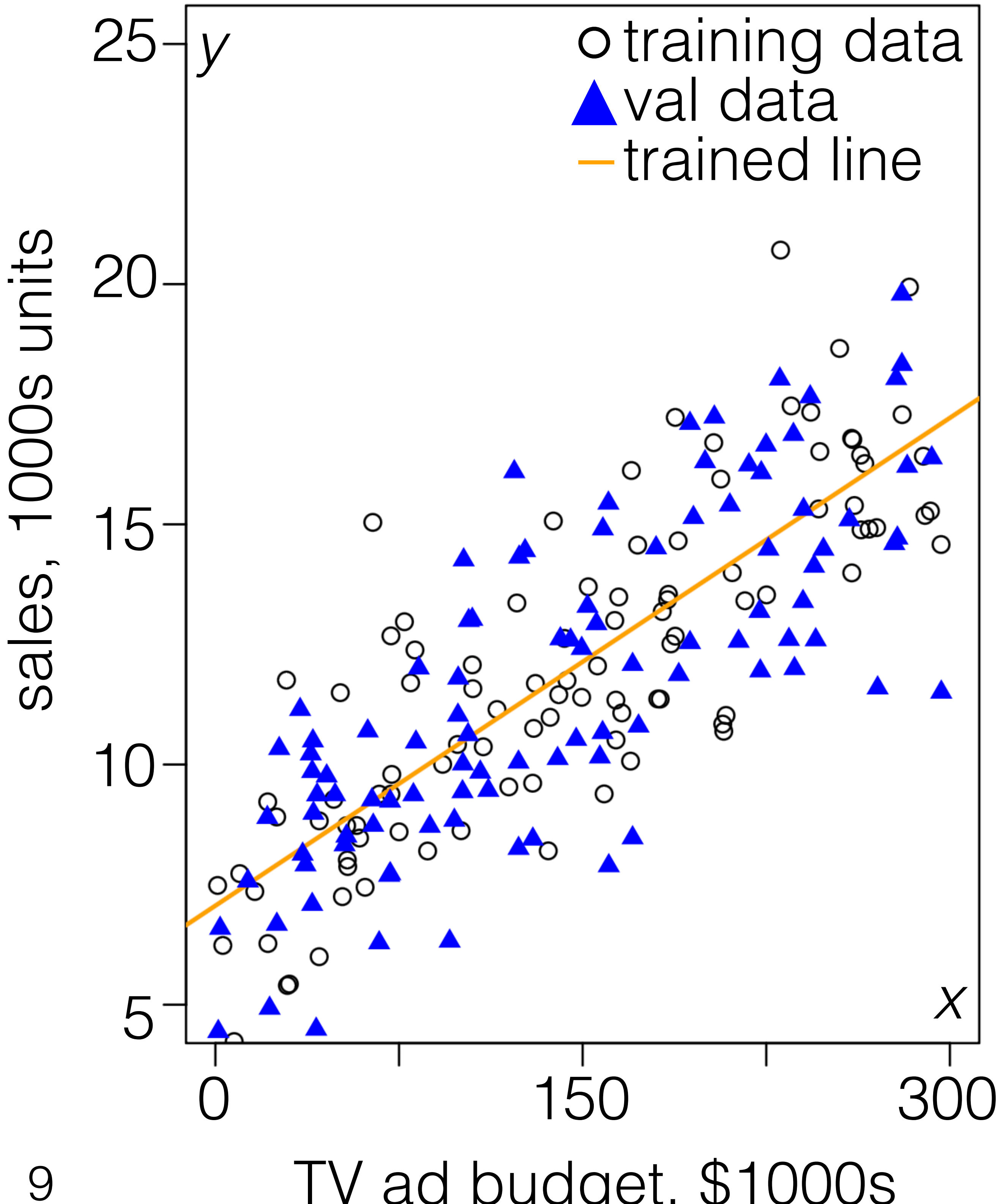
- My boss wants to know the expected loss of my prediction for TV ad budgets in the $300,000 to $400,000 range



- Does the empirical average of the loss over validation data estimate this value?
- The x values my boss proposes aren't from the same distribution as the validation x
- Aside: It's difficult to predict or evaluate outside the observed data without making any other assumptions

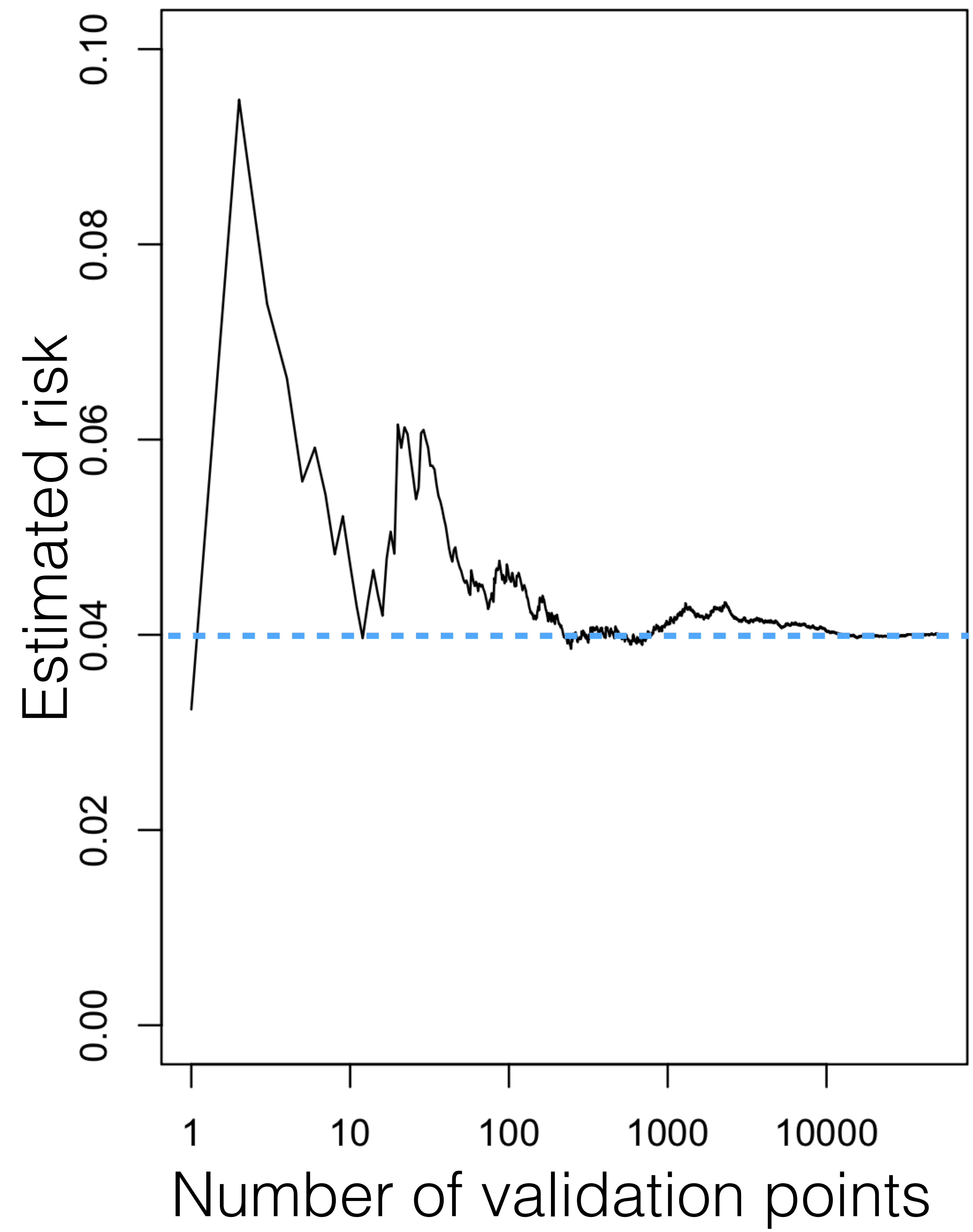
What’s the data distribution?

- My boss wants to know the expected loss of my prediction for TV ad budgets in the $300,000 to $400,000 range



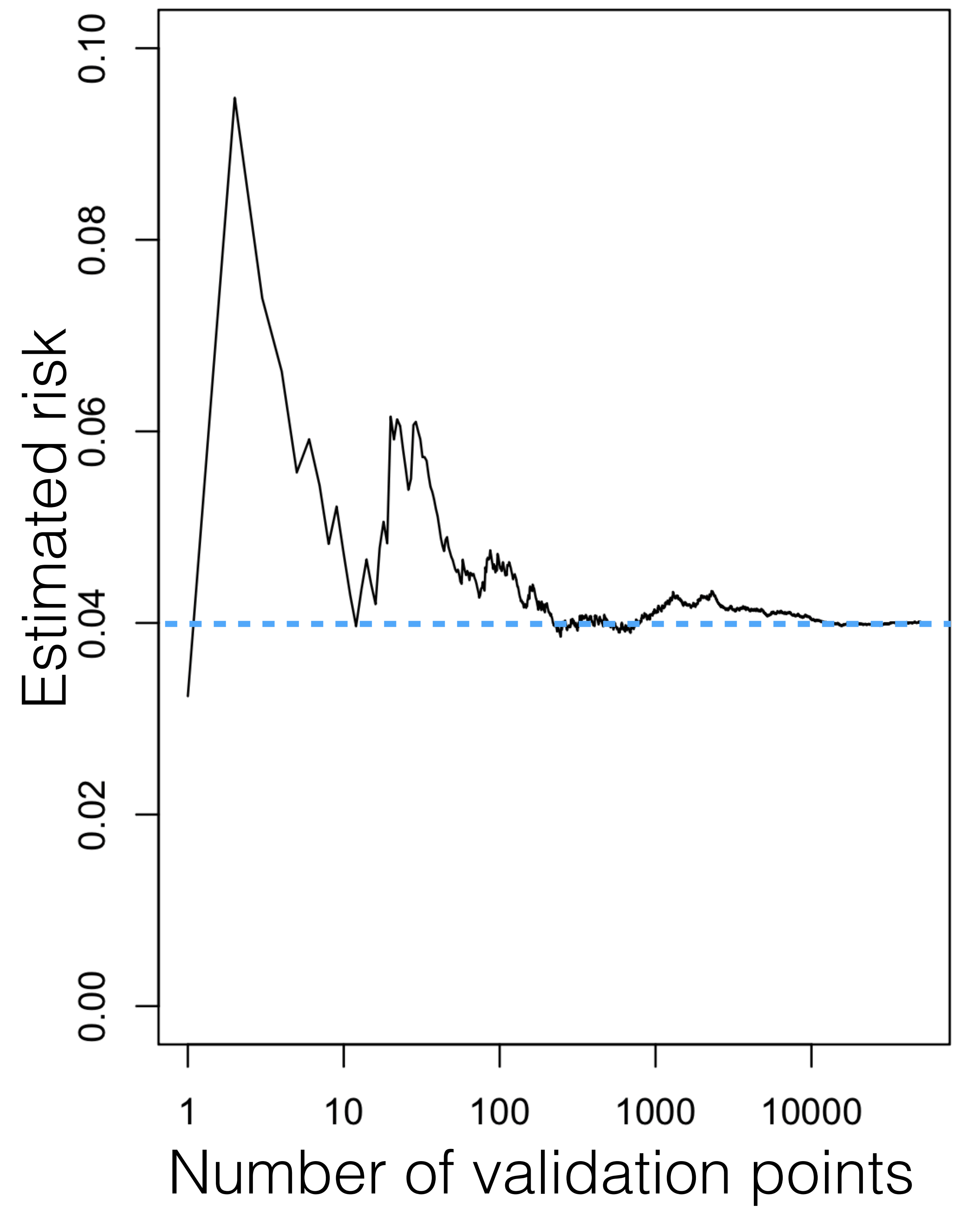
- Does the empirical average of the loss over validation data estimate this value?
- The *x* values my boss proposes aren’t from the same distribution as the validation *x*
- Aside: It’s difficult to predict or evaluate outside the observed data without making any other assumptions
  - People make this kind of prediction all the time. Ask: What assumptions are they making?

Did I use enough validation points?



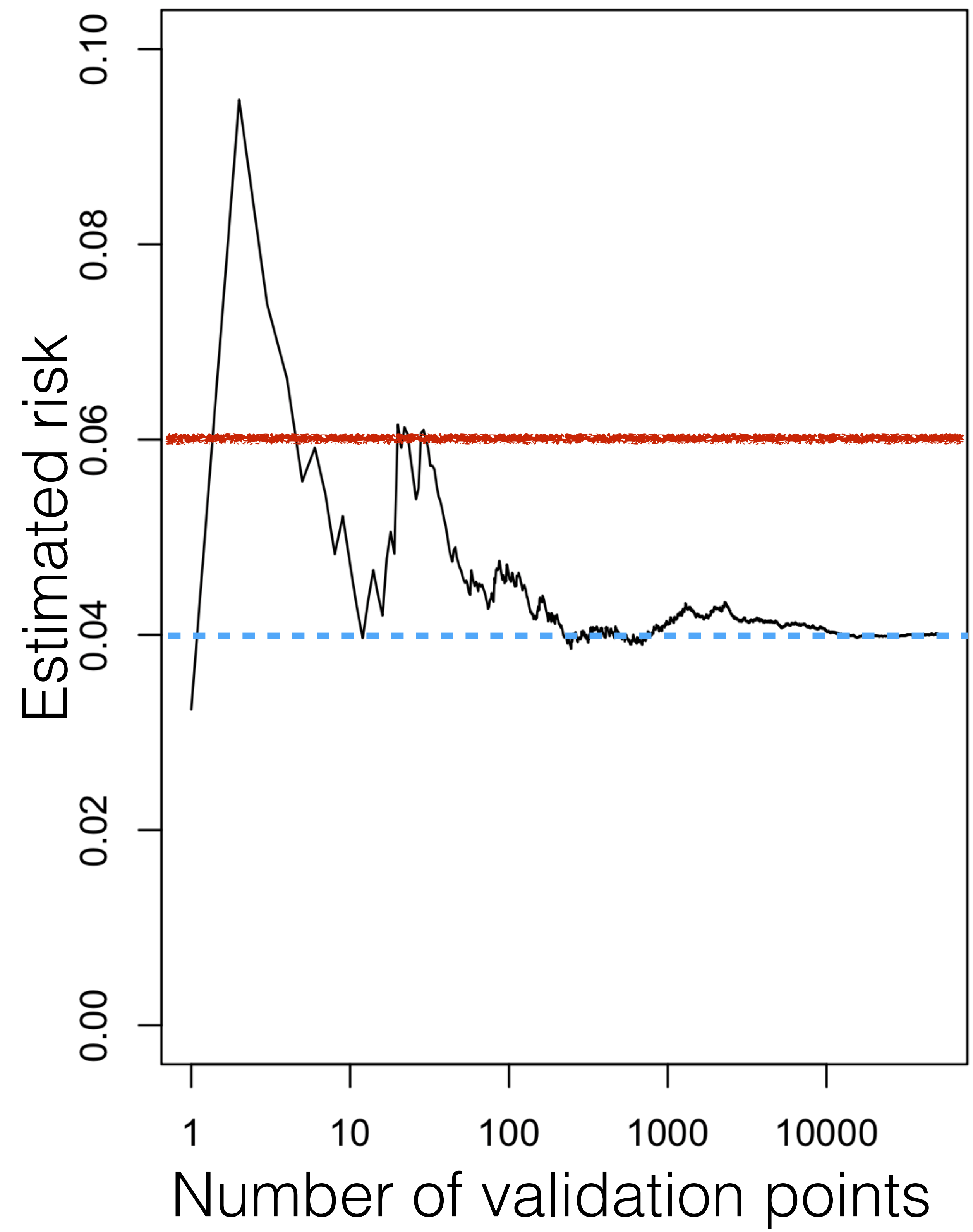
Did I use enough validation points?

- The answer depends on how precisely you need to know risk



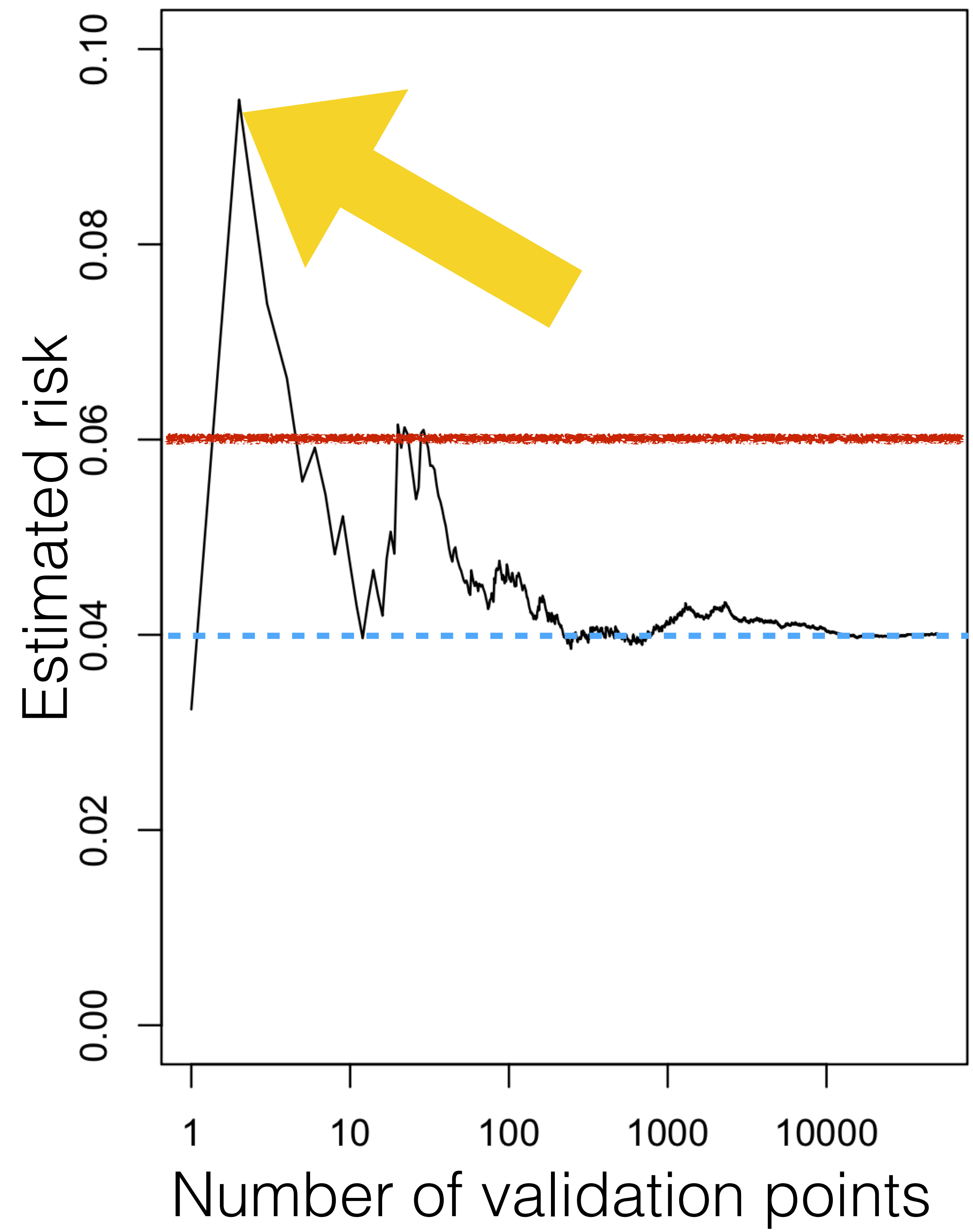
Did I use enough validation points?

- The answer depends on how precisely you need to know risk



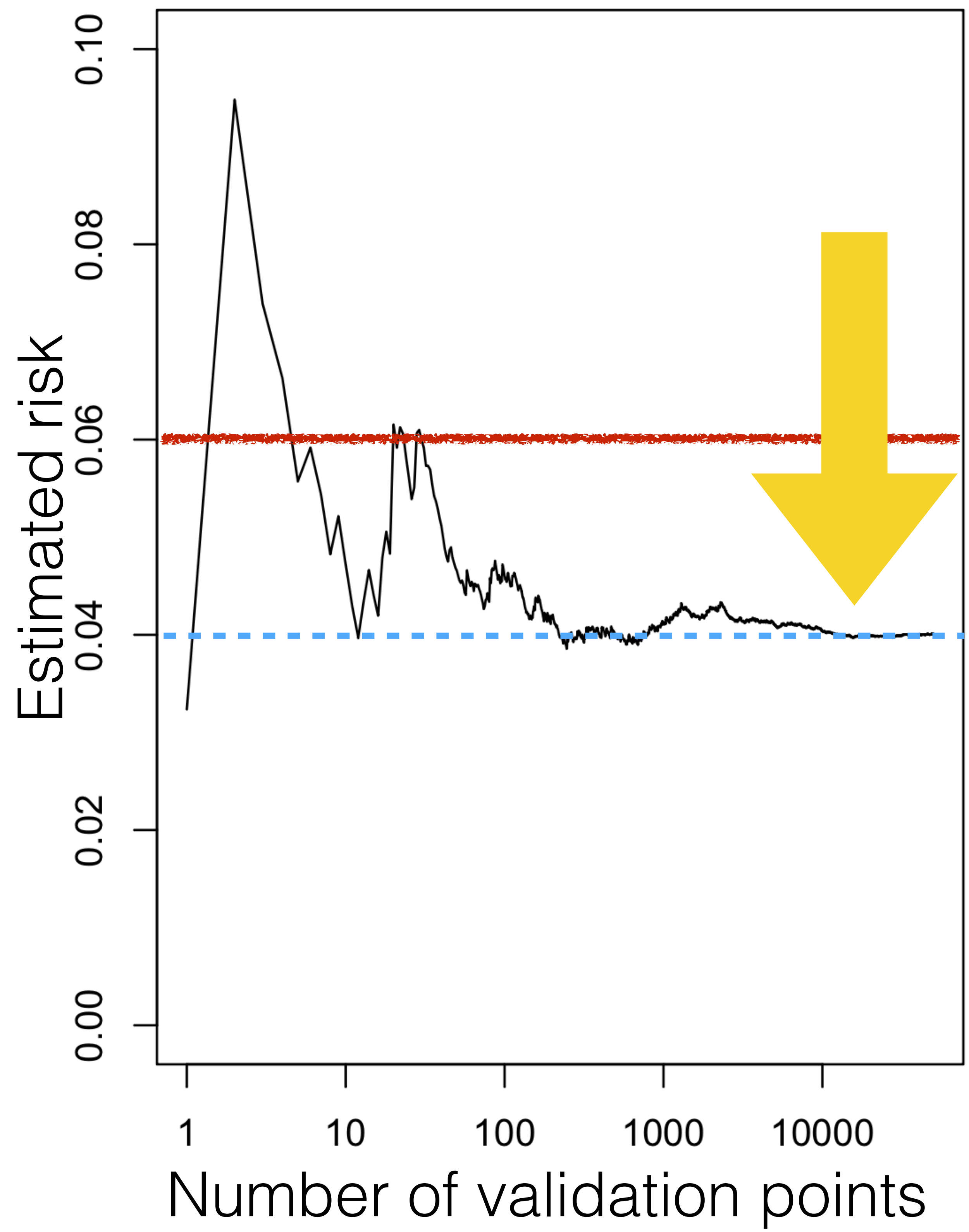
Did I use enough validation points?

- The answer depends on how precisely you need to know risk



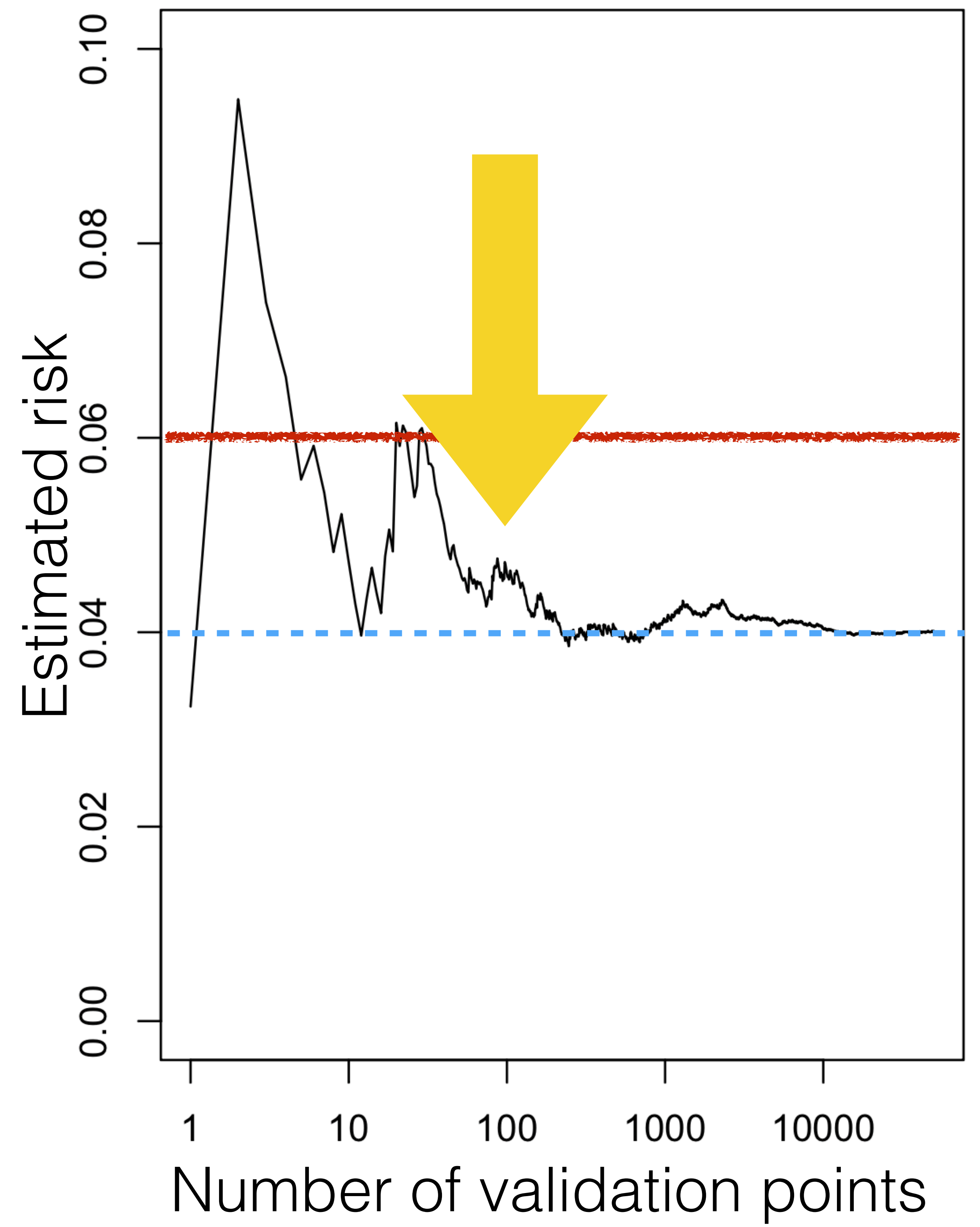
Did I use enough validation points?

- The answer depends on how precisely you need to know risk



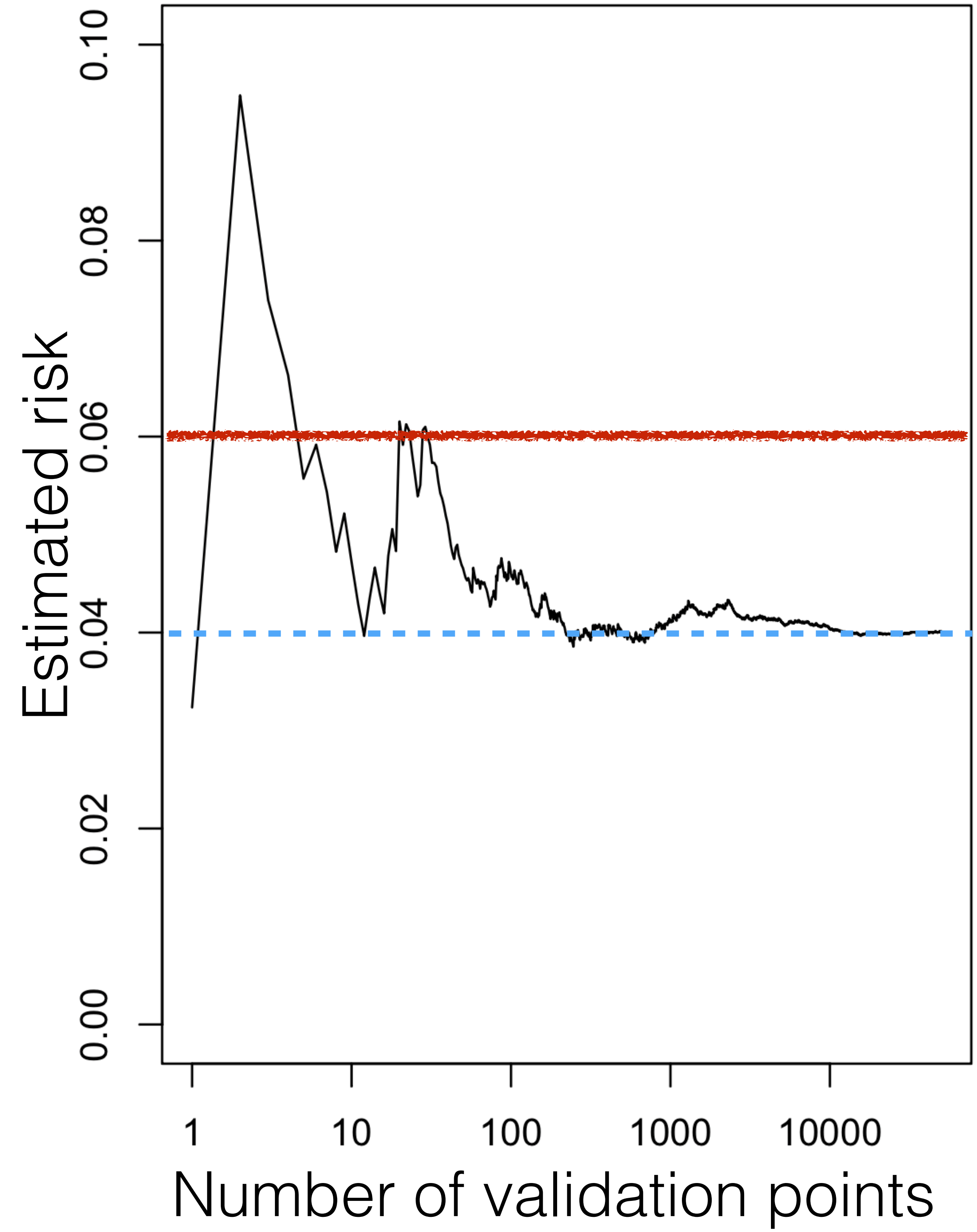
Did I use enough validation points?

- The answer depends on how precisely you need to know risk



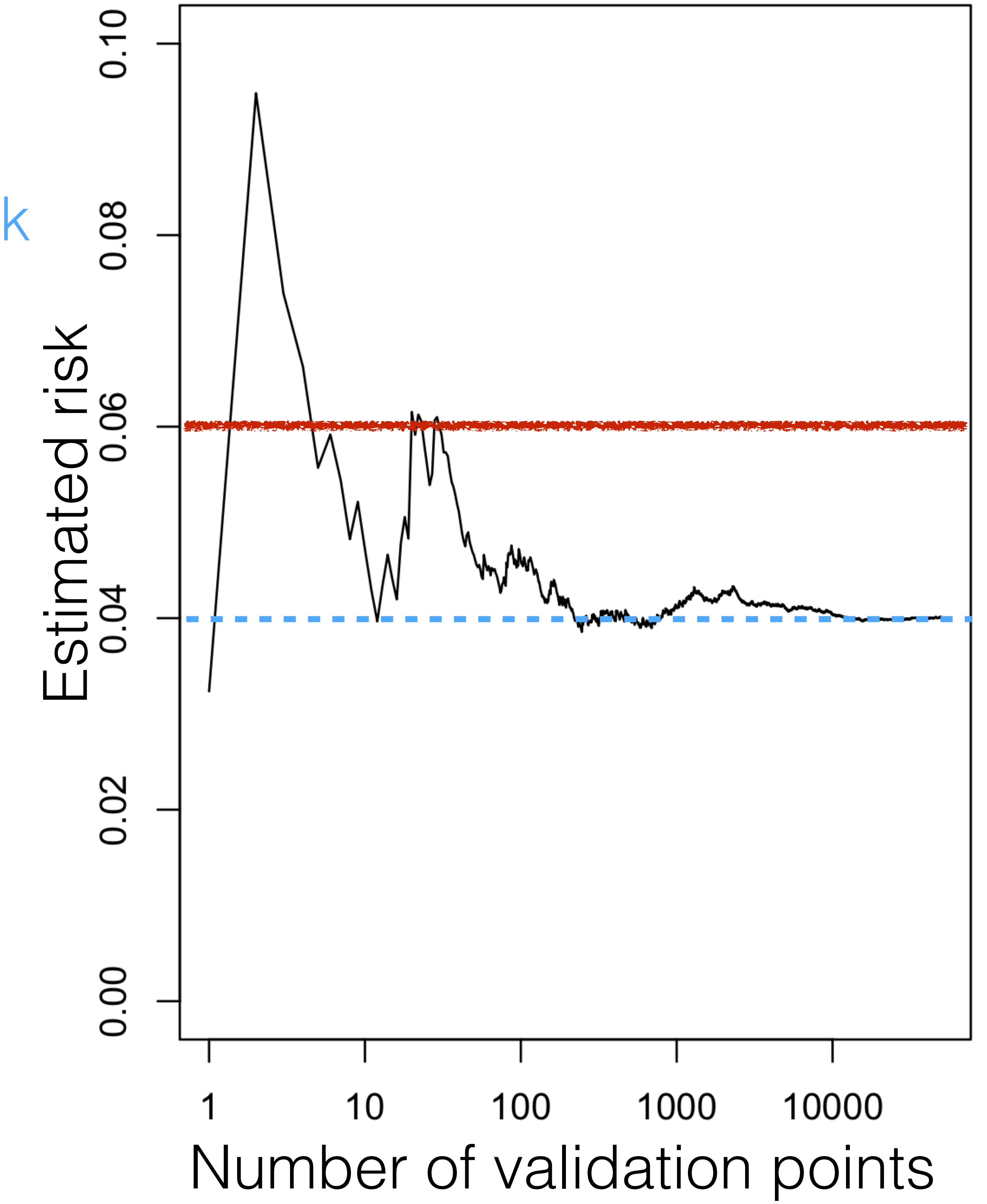
Did I use enough validation points?

- The answer depends on how precisely you need to know risk
- If $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are iid random variables with finite mean and variance, then the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$



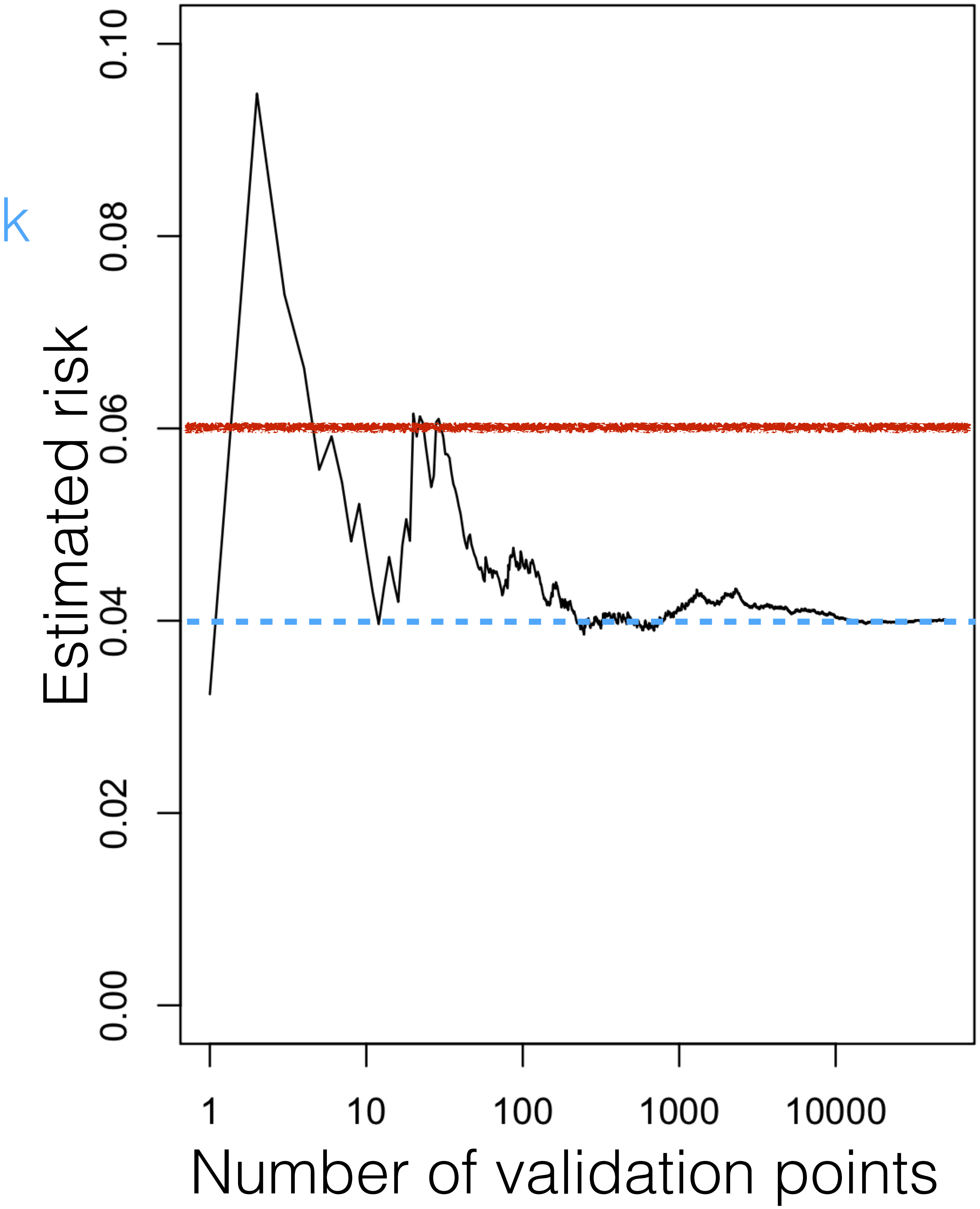
Did I use enough validation points?

- The answer depends on how precisely you need to know risk
- If $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are iid random variables with finite mean and variance, then the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$ check



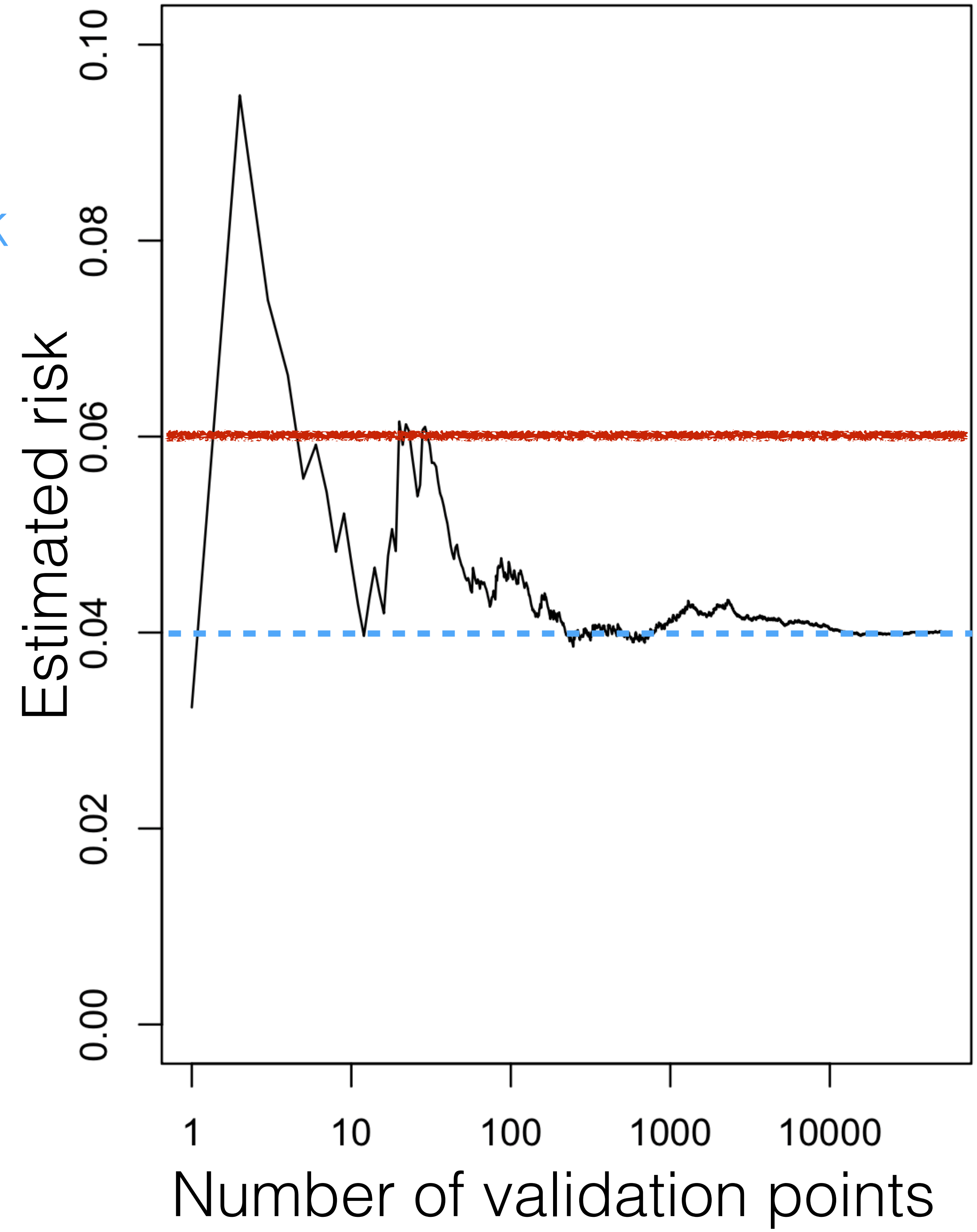
Did I use enough validation points?

- The answer depends on how precisely you need to know risk
- If $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are iid random variables with finite mean and variance, then the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$ check
  - The square root is often called the **standard error**



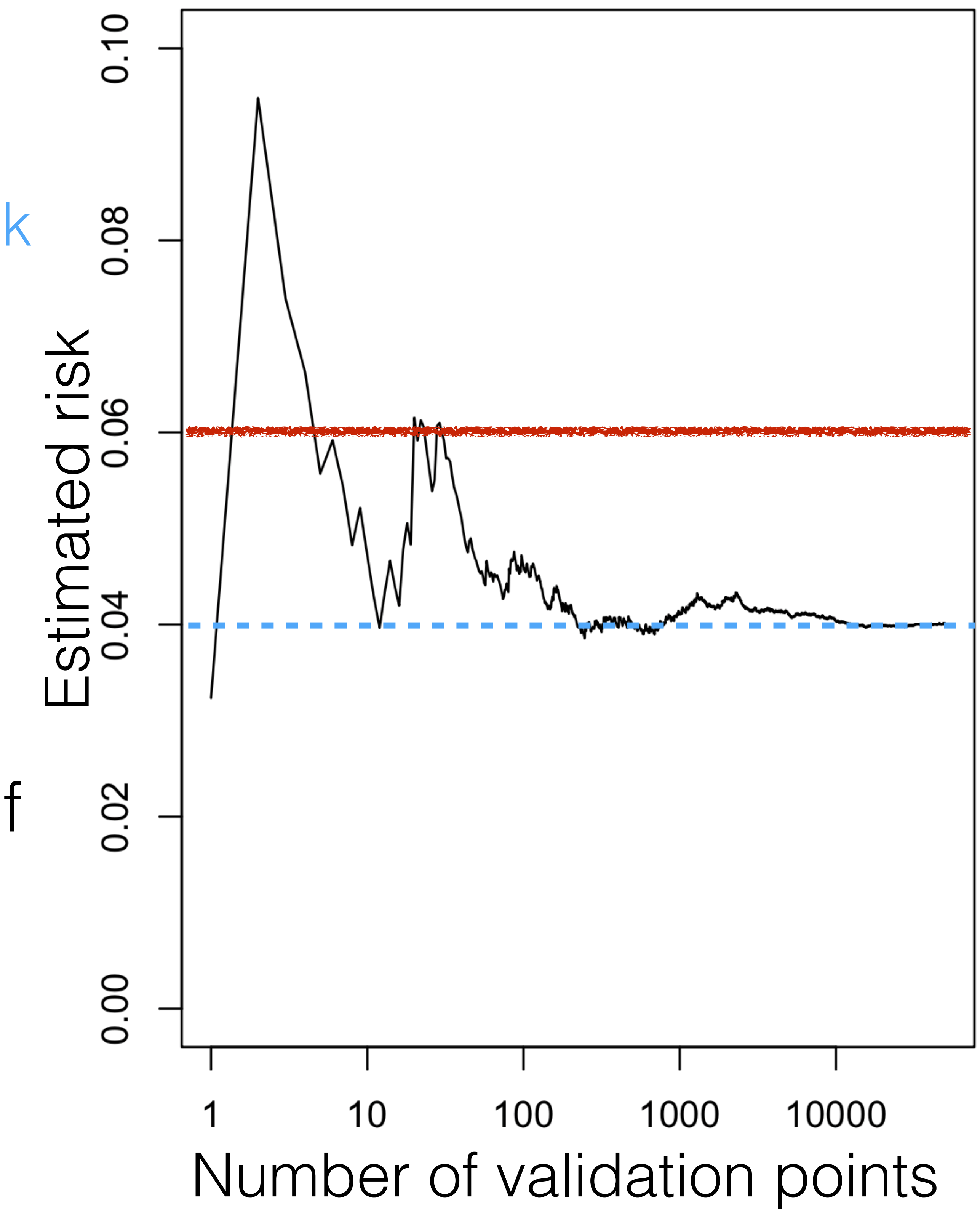
Did I use enough validation points?

- The answer depends on how precisely you need to know risk
- If $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are iid random variables with finite mean and variance, then the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$ check
  - The square root is often called the **standard error**
  - An estimate of $\mathrm{Var}(Z^{(1)})$ is the empirical var of the Z's.



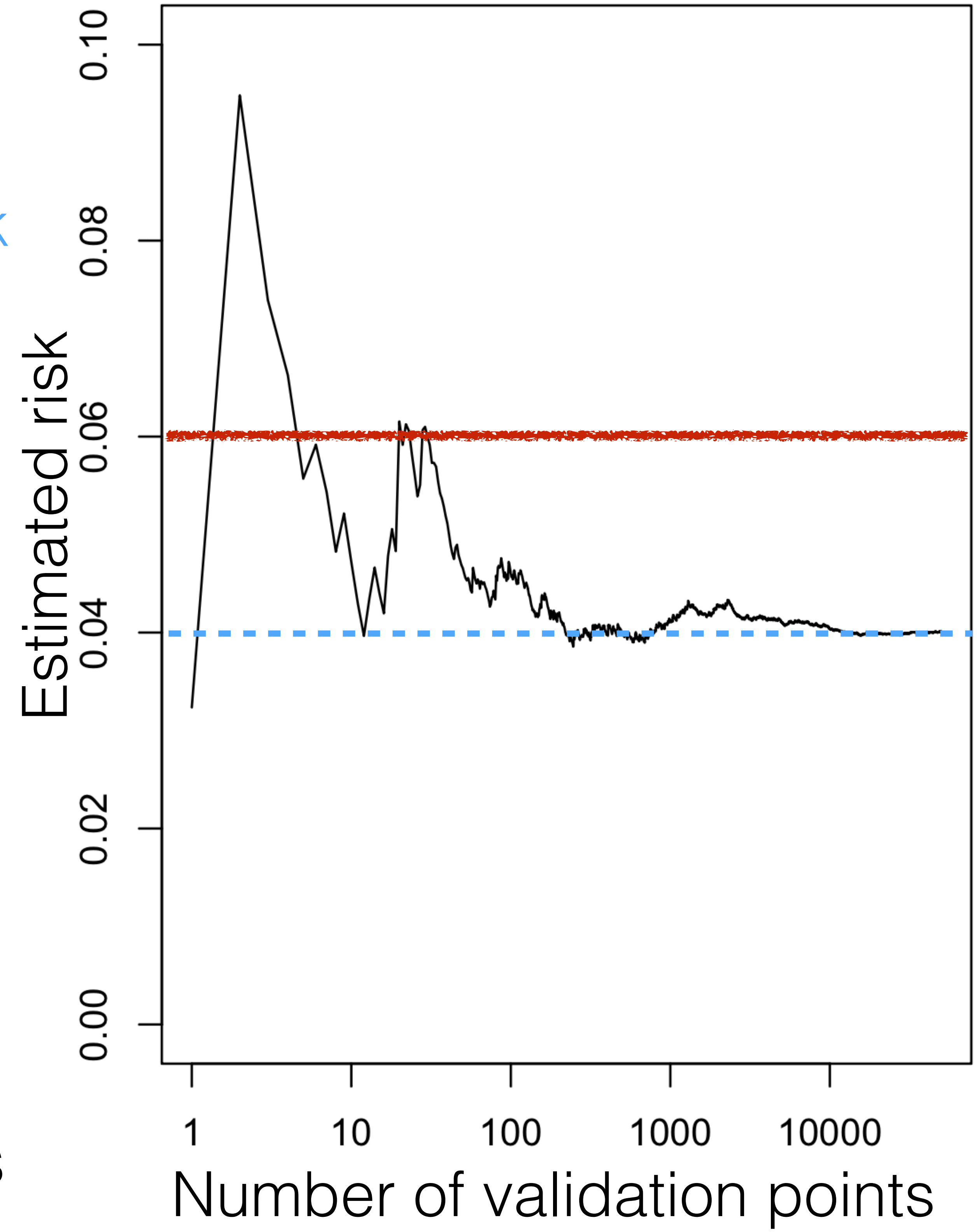
Did I use enough validation points?

- The answer depends on how precisely you need to know risk
- If $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are iid random variables with finite mean and variance, then the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$ check
  - The square root is often called the **standard error**
  - An estimate of $\mathrm{Var}(Z^{(1)})$ is the empirical var of the Z's.
- So an estimate of the std dev of the risk estimate is: the empirical standard deviation of the observed losses divided by the square root of the # of validation points



Did I use enough validation points?

- The answer depends on how precisely you need to know risk
- If $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are iid random variables with finite mean and variance, then the variance of their empirical average is $\mathrm{Var}(Z^{(1)})/N$ check
  - The square root is often called the **standard error**
  - An estimate of $\mathrm{Var}(Z^{(1)})$ is the empirical var of the Z's.
- So an estimate of the std dev of the risk estimate is: the empirical standard deviation of the observed losses divided by the square root of the # of validation points
- Can use to compare predictors



Heavy-tailed data

Heavy-tailed data

- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.

Heavy-tailed data

- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.
- The expectation and variance exist for Gaussian distributions.

Heavy-tailed data

- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.
- The expectation and variance exist for Gaussian distributions.
- But they don’t have to exist for **heavy-tailed distributions**, which often occur in real life.
    - Heavy-tailed: mass in tails doesn’t decay exponentially fast

Heavy-tailed data

- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.
- The expectation and variance exist for Gaussian distributions.
- But they don't have to exist for **heavy-tailed distributions**, which often occur in real life.
  - Heavy-tailed: mass in tails doesn't decay exponentially fast
- Applications where variance or mean at least sometimes does not exist:

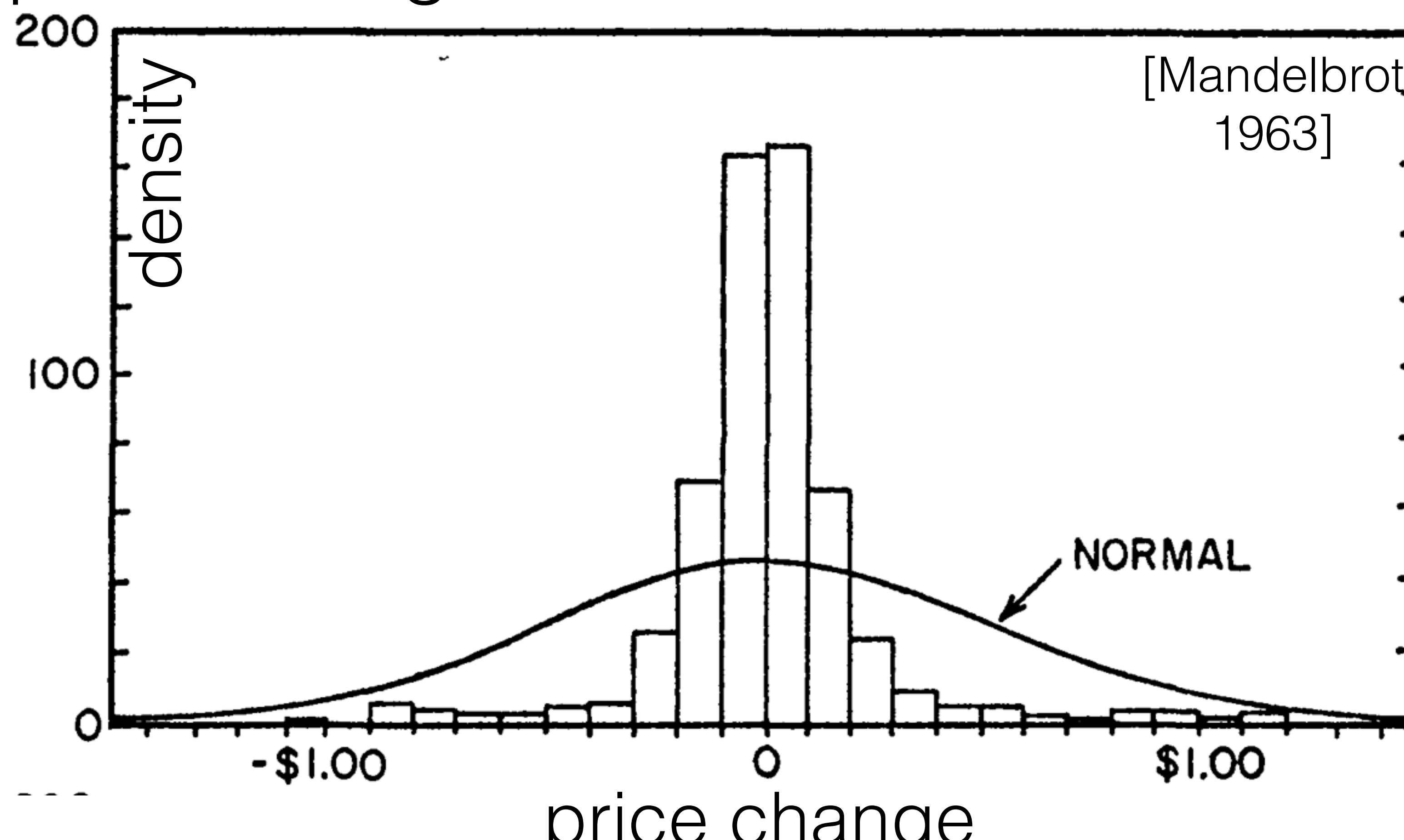
Heavy-tailed data

- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.
- The expectation and variance exist for Gaussian distributions.
- But they don't have to exist for **heavy-tailed distributions**, which often occur in real life.
  - Heavy-tailed: mass in tails doesn't decay exponentially fast
- Applications where variance or mean at least sometimes does not exist: historical daily price changes in cotton

[Meerschaert & Scheffler]

Heavy-tailed data

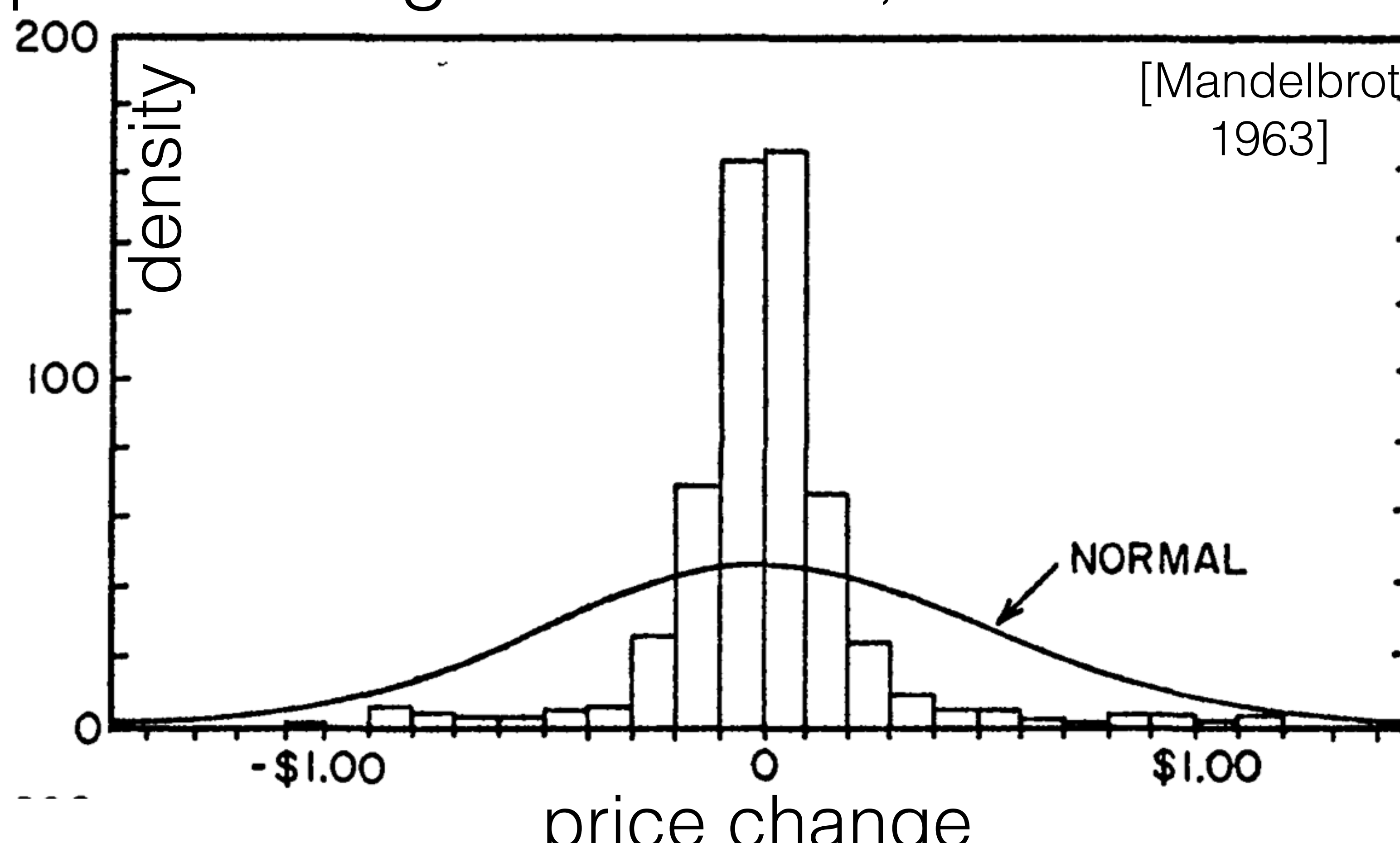
- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.
- The expectation and variance exist for Gaussian distributions.
- But they don't have to exist for **heavy-tailed distributions**, which often occur in real life.
  - Heavy-tailed: mass in tails doesn't decay exponentially fast
- Applications where variance or mean at least sometimes does not exist: historical daily price changes in cotton



[Meerschaert & Scheffler]

Heavy-tailed data

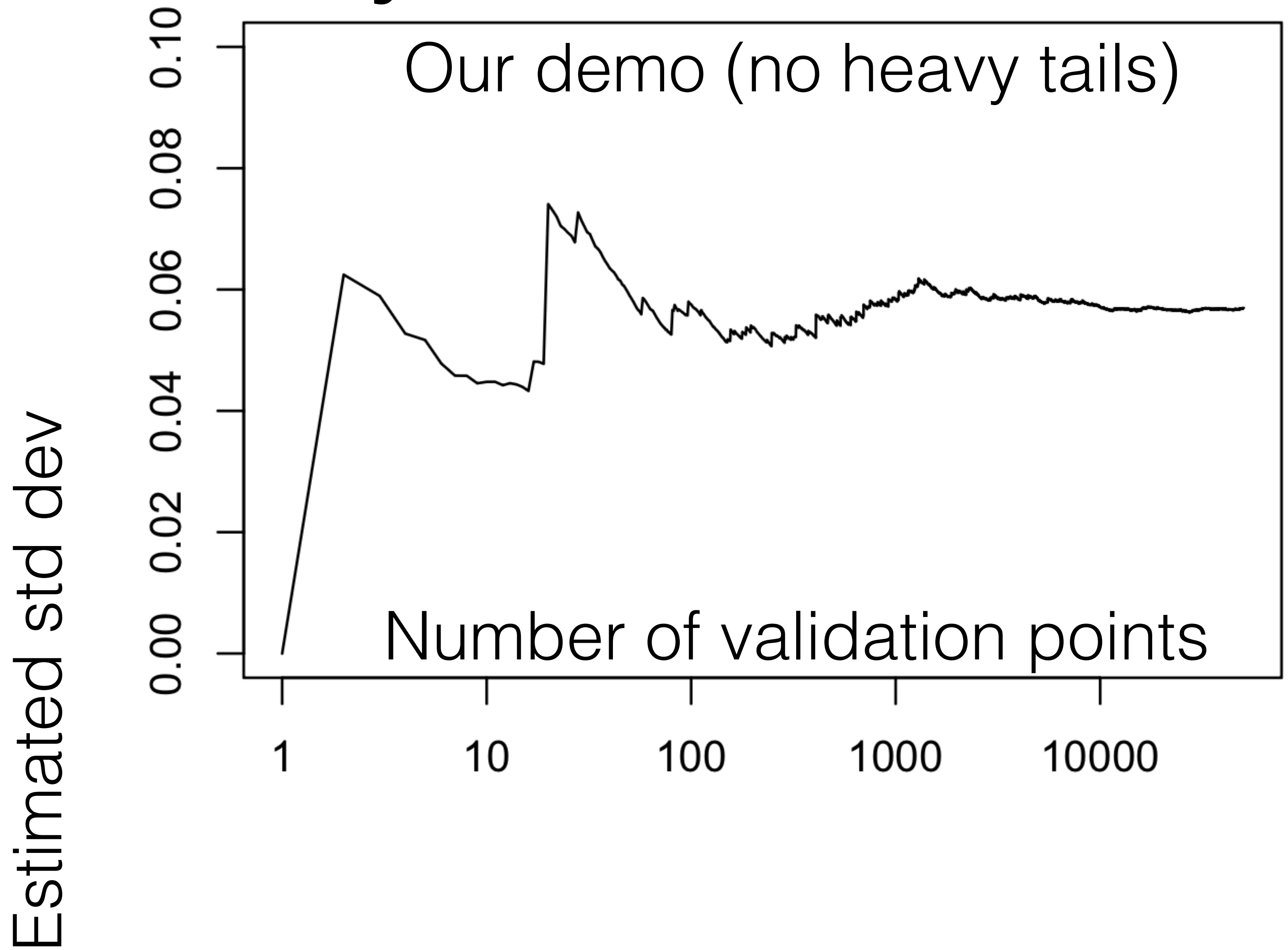
- On the board, we assumed the expected loss existed. On the previous slide, we assumed the variance of the loss existed.
- The expectation and variance exist for Gaussian distributions.
- But they don't have to exist for **heavy-tailed distributions**, which often occur in real life.
  - Heavy-tailed: mass in tails doesn't decay exponentially fast
- Applications where variance or mean at least sometimes does not exist: historical daily price changes in cotton, stock indices, exchange rates, groundwater, physics (e.g. diffusions), quiet periods between transmissions for a networked computer terminal [Meerschaert & Scheffler]



Heavy-tailed data: what happens?

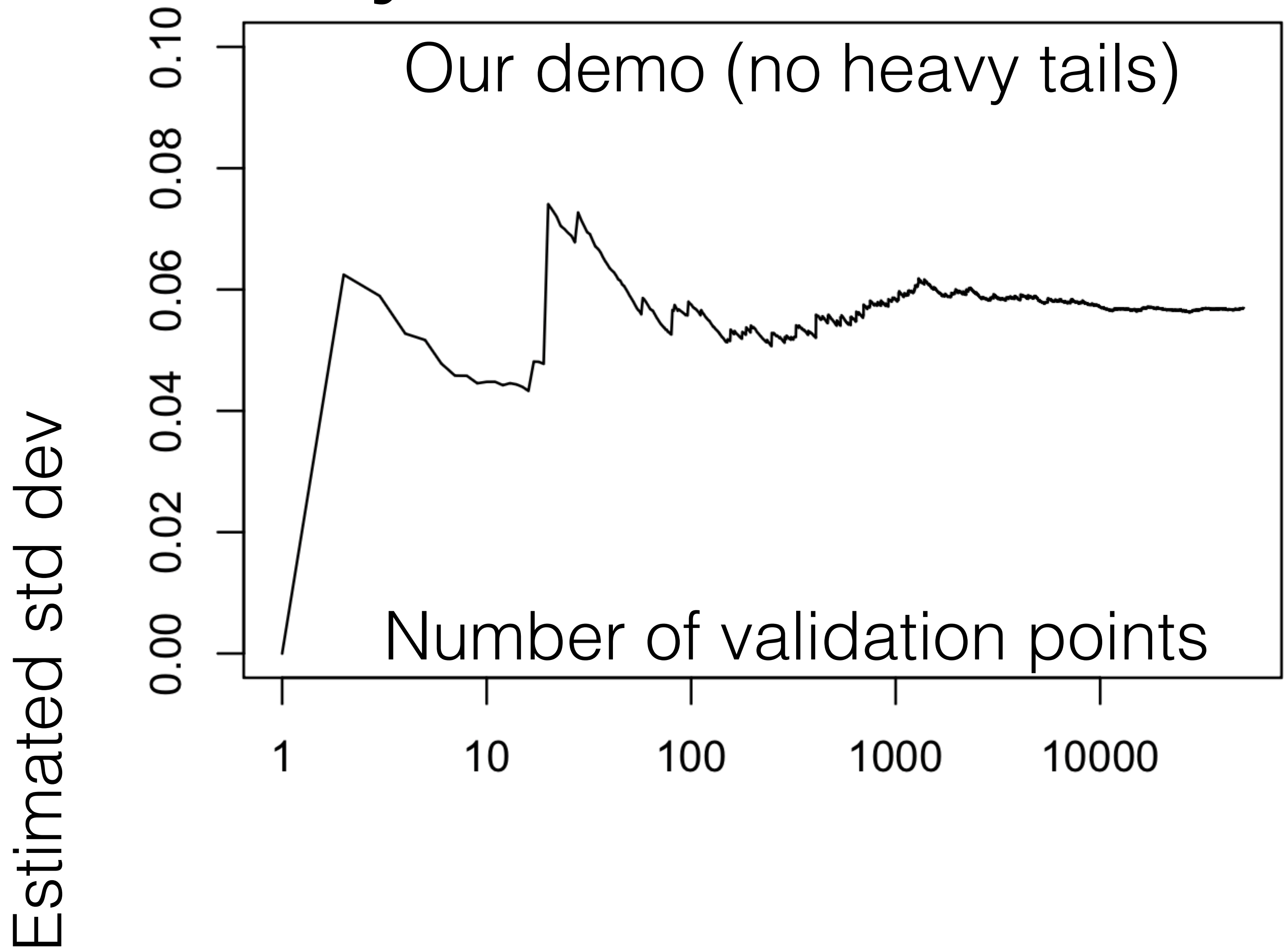
- Same setup as our demo from Lecture 6

Heavy-tailed data: what happens?



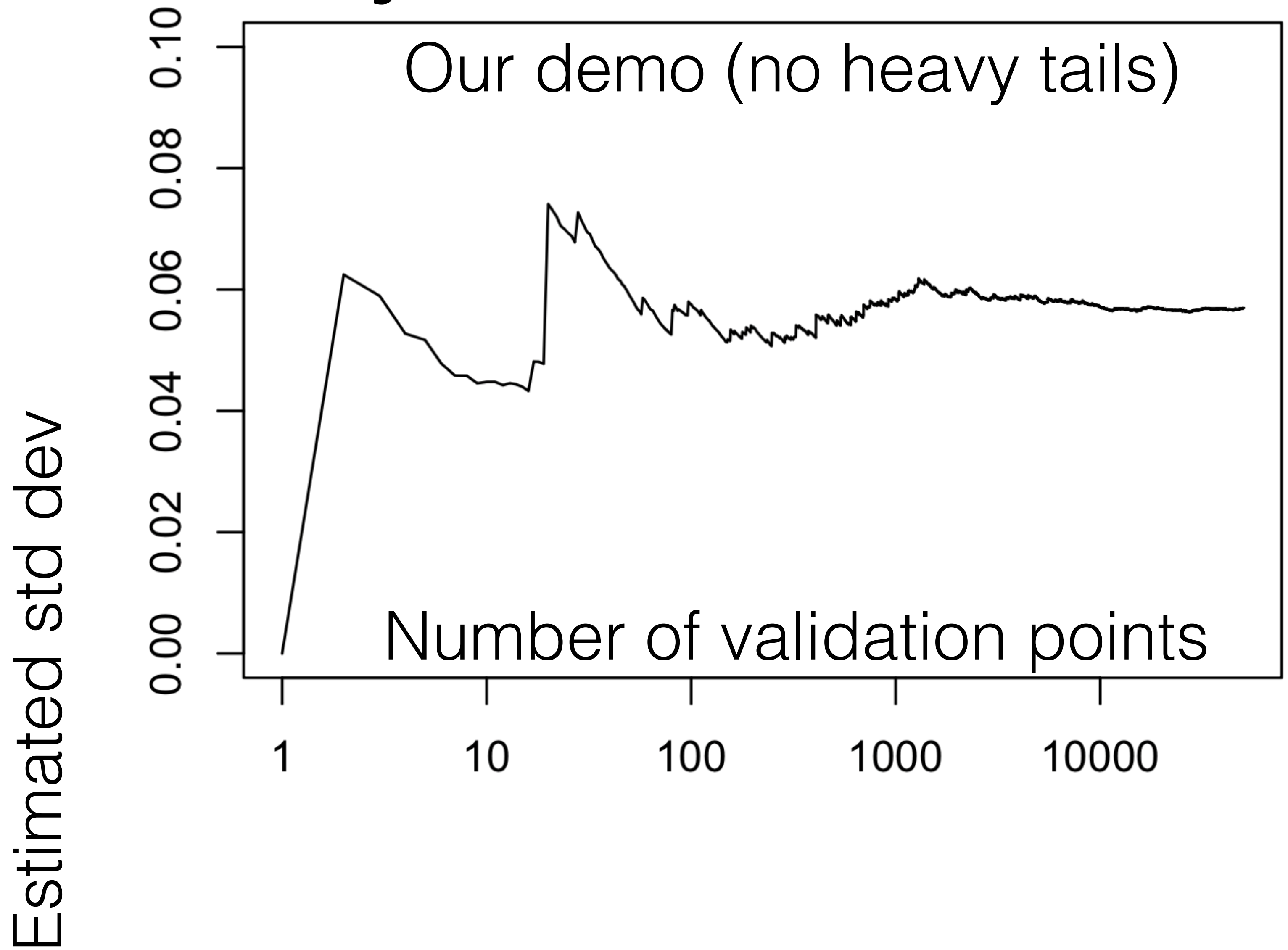
- Same setup as our demo from Lecture 6

Heavy-tailed data: what happens?



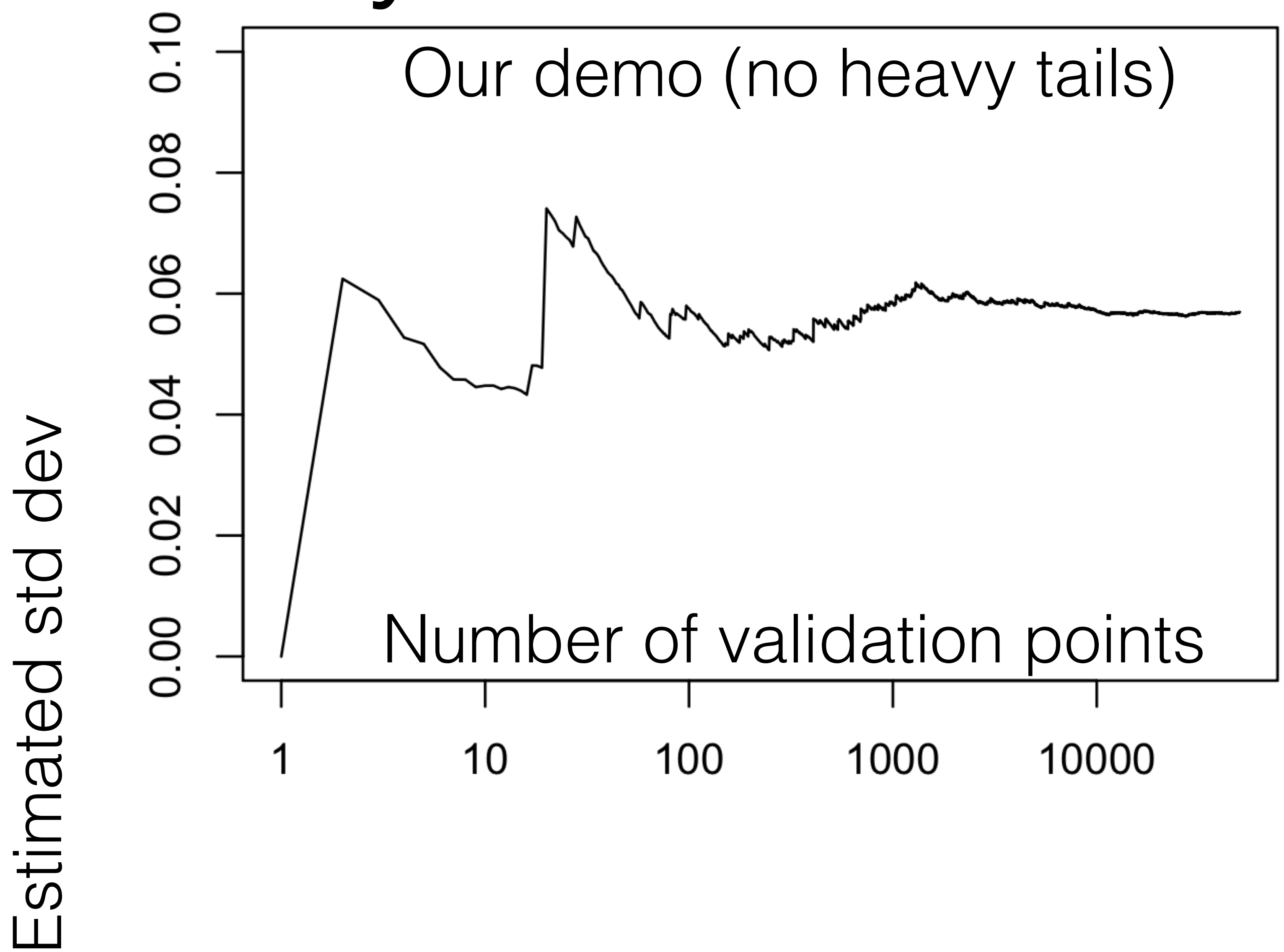
- Same setup as our demo from Lecture 6
- Mean and variance exist

Heavy-tailed data: what happens?



- Same setup as our demo from Lecture 6
- Mean and variance exist
- Estimate of standard deviation converges to the exact standard deviation as we get more data

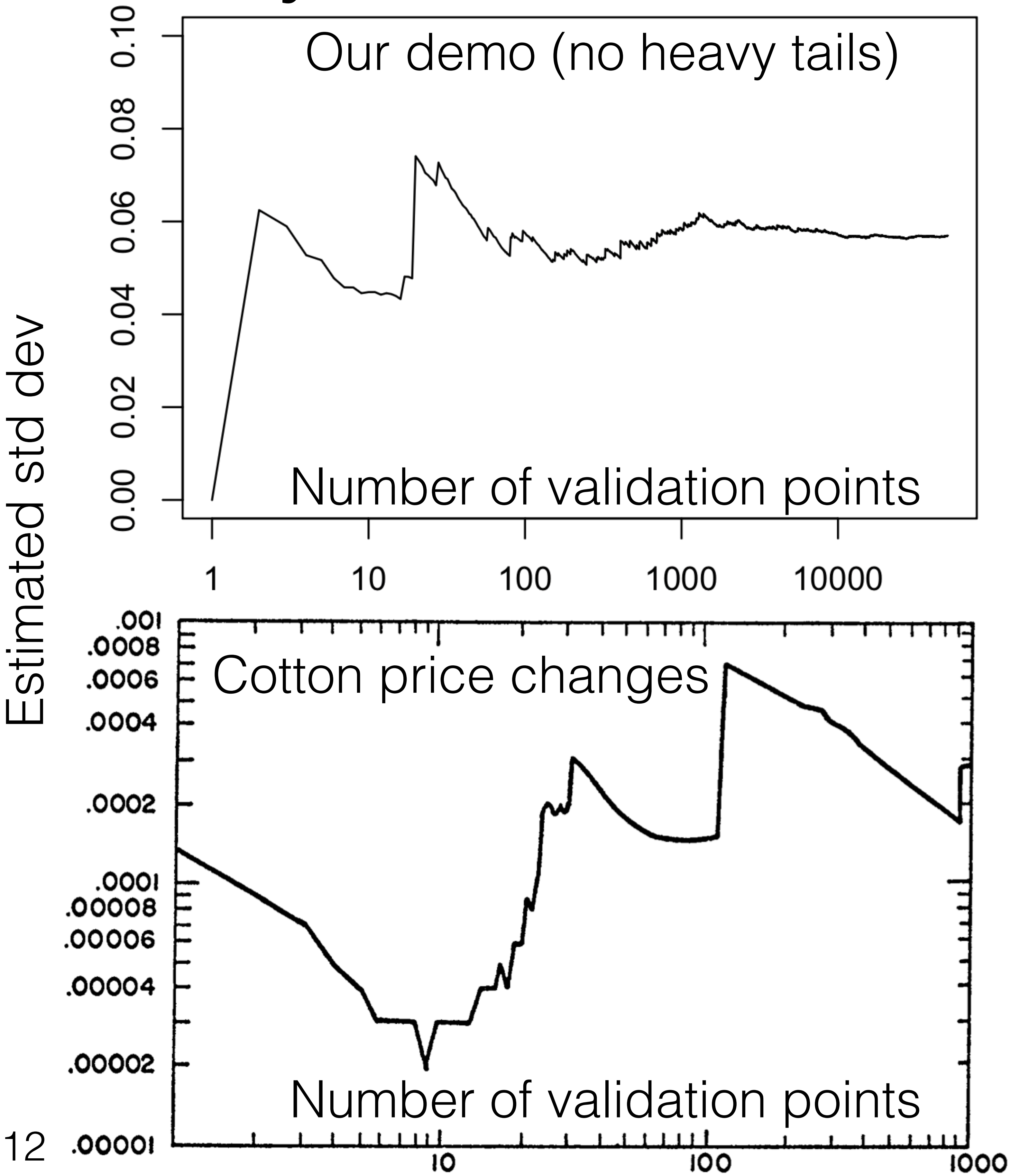
Heavy-tailed data: what happens?



- Same setup as our demo from Lecture 6
- Mean and variance exist
- Estimate of standard deviation converges to the exact standard deviation as we get more data

- Real cotton data

Heavy-tailed data: what happens?

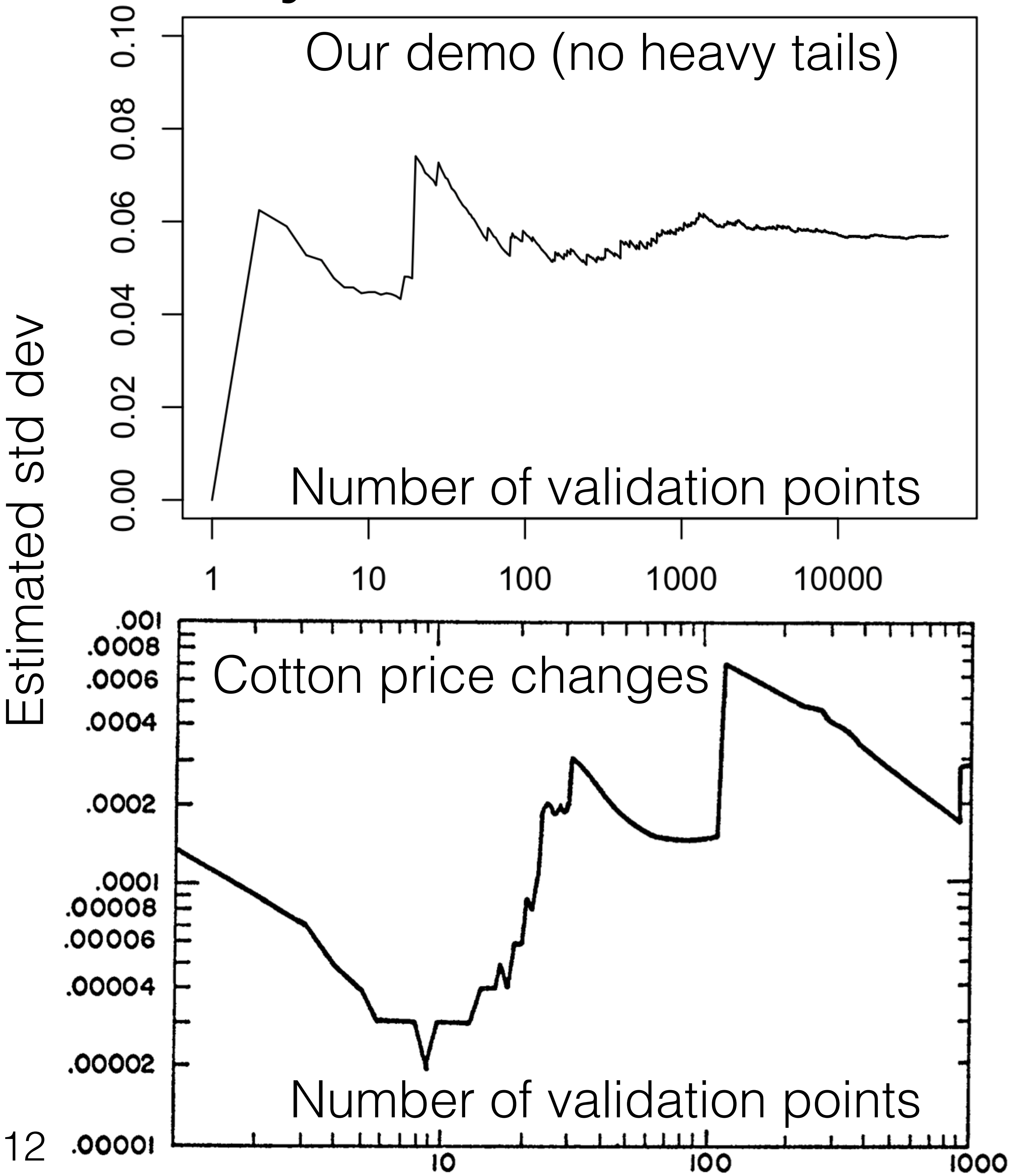


- Same setup as our demo from Lecture 6
- Mean and variance exist
- Estimate of standard deviation converges to the exact standard deviation as we get more data

- Real cotton data

[Mandelbrot 1963; technically the cotton plot is of the second moment estimate, but the point is the same]

Heavy-tailed data: what happens?

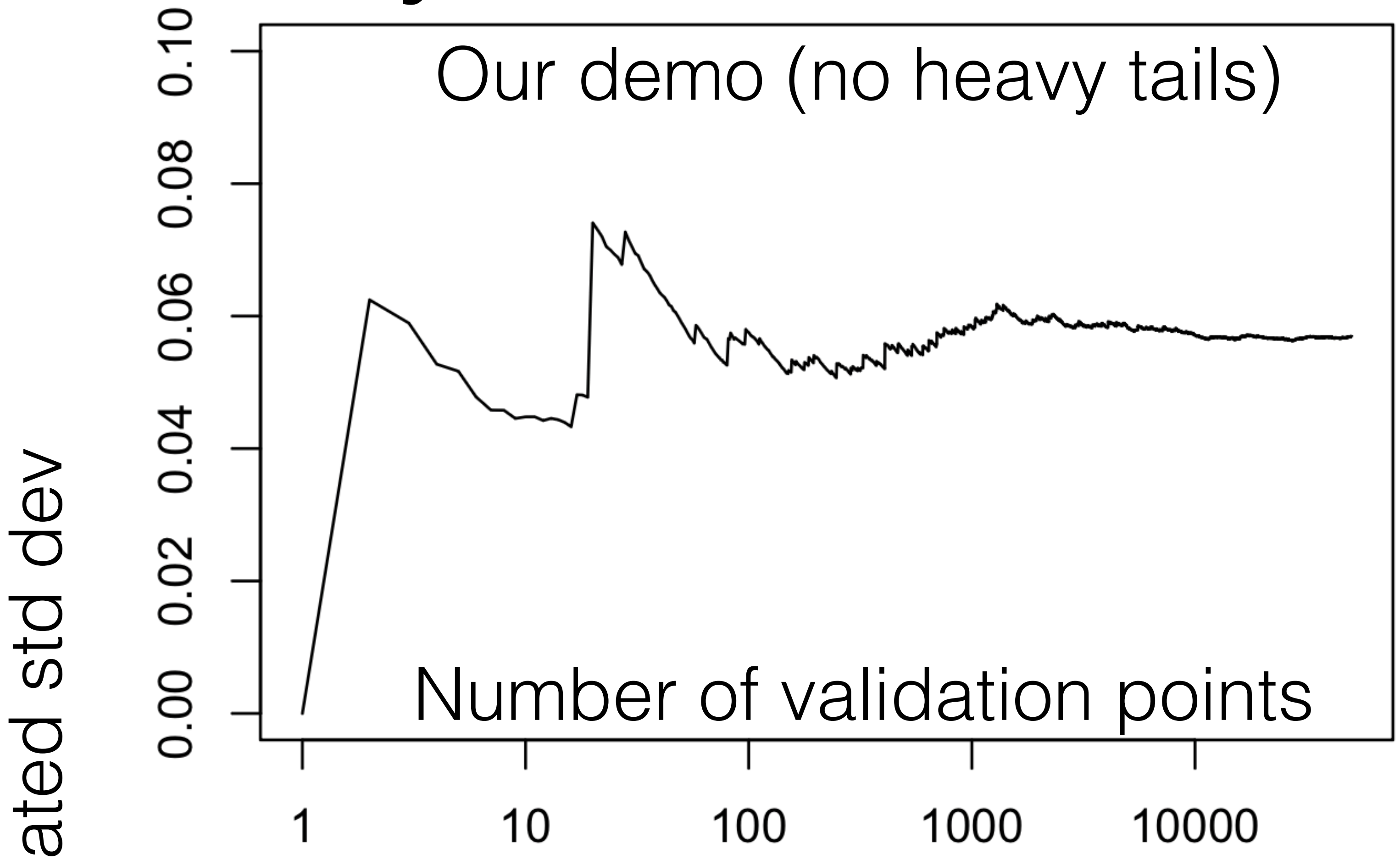


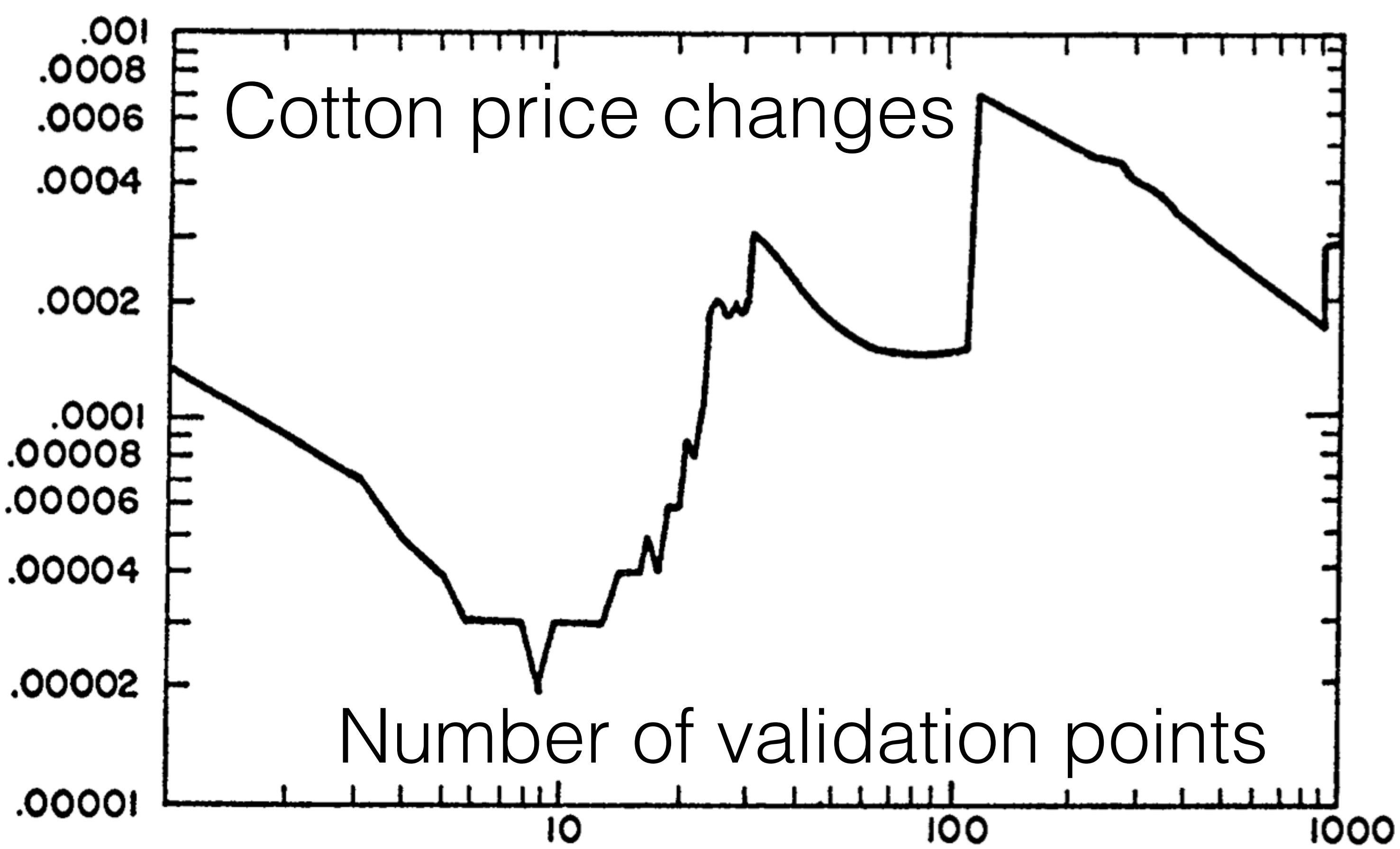
- Same setup as our demo from Lecture 6
- Mean and variance exist
- Estimate of standard deviation converges to the exact standard deviation as we get more data

- Real cotton data (notice the axes)

[Mandelbrot 1963; technically the cotton plot is of the second moment estimate, but the point is the same]

Heavy-tailed data: what happens?



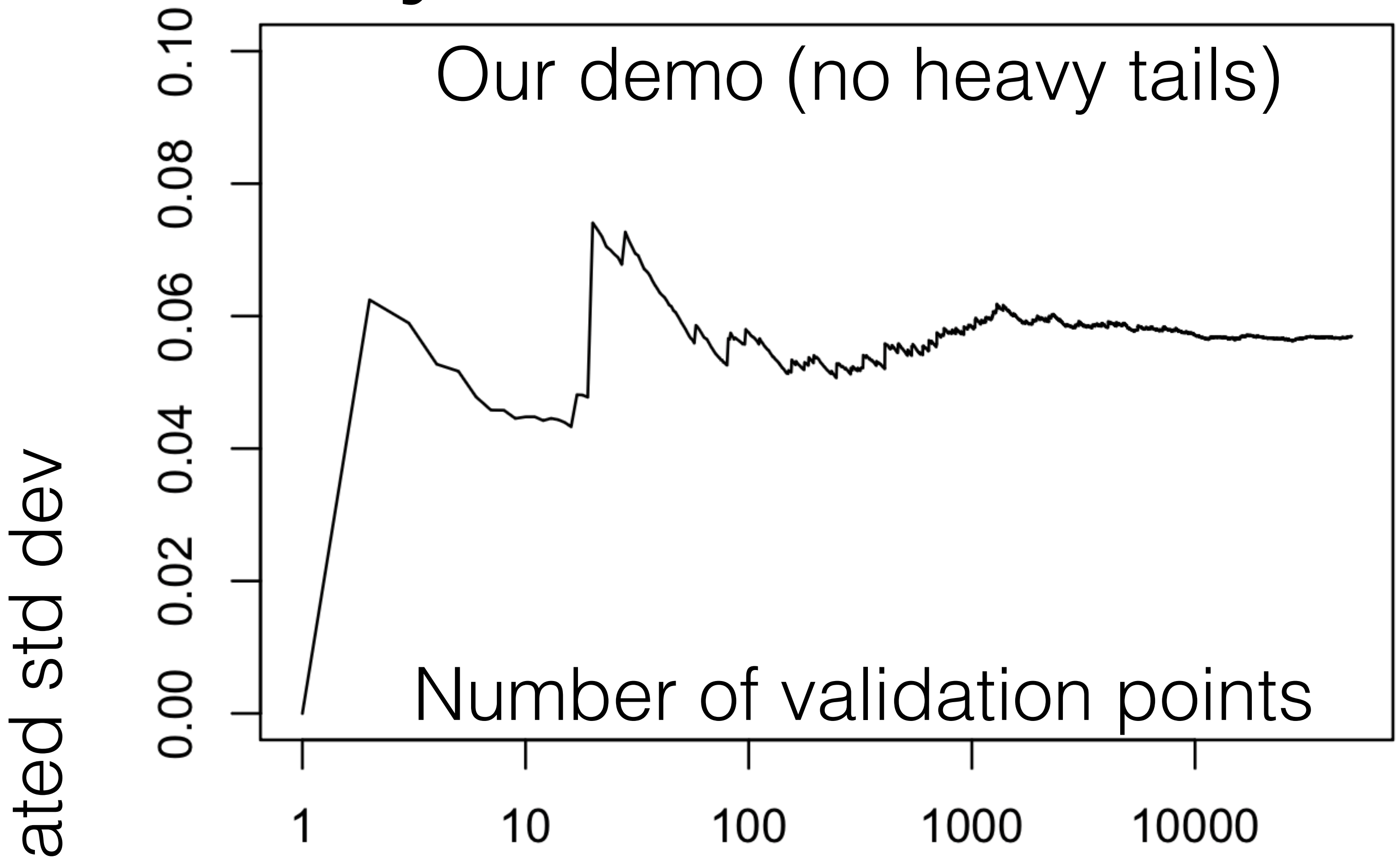


- Same setup as our demo from Lecture 6
- Mean and variance exist
- Estimate of standard deviation converges to the exact standard deviation as we get more data

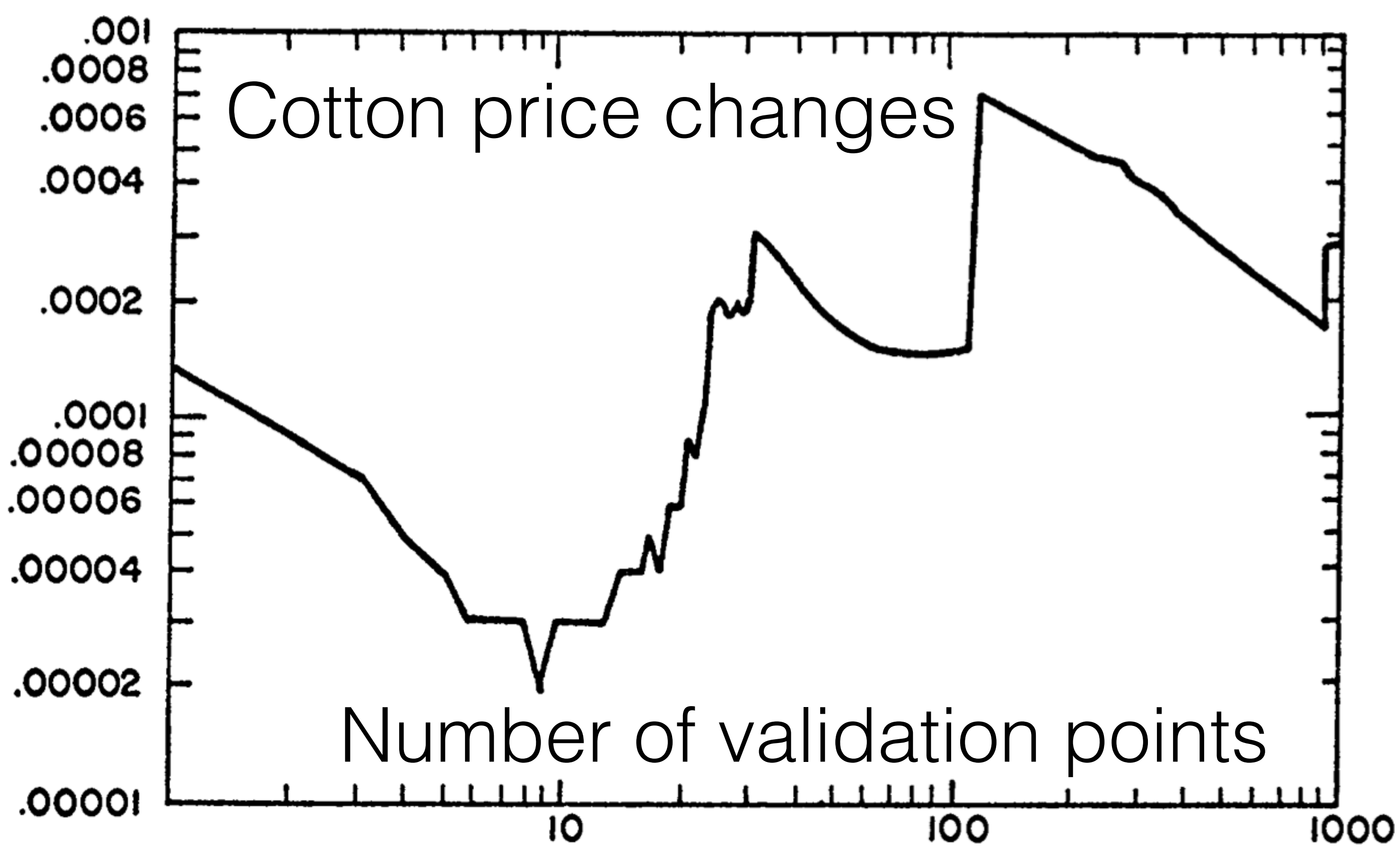
- Real cotton data (notice the axes)
- Variance doesn't exist

[Mandelbrot 1963; technically the cotton plot is of the second moment estimate, but the point is the same]

Heavy-tailed data: what happens?



- Same setup as our demo from Lecture 6
- Mean and variance exist
- Estimate of standard deviation converges to the exact standard deviation as we get more data



- Real cotton data (notice the axes)
- Variance doesn't exist
- Estimate jumps around with more data, does not converge

[Mandelbrot 1963; technically the cotton plot is of the second moment estimate, but the point is the same]

References (1/1)

Meerschaert, M. M. and Hans-Peter Scheffler. Nonparametric methods for heavy tailed vector data: A survey with applications from finance and hydrology.

Mandelbrot, Benoit B. “The variation of certain speculative prices.” The Journal of Business, 1963.