What uncertainty are we quantifying?

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
    - E.g. “significance”, “bias”, “generalization”

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. “significance”, “bias”, “generalization”
  - Every precise use of “uncertainty” has this issue
    - E.g. frequentist sampling, Bayesian, etc.

What uncertainty are we quantifying?

- It's worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. "significance", "bias", "generalization"
  - Every precise use of "uncertainty" has this issue
    - E.g. frequentist sampling, Bayesian, etc.
  - We should always make sure we can distinguish what is, and what is not, covered by the term of art

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. “significance”, “bias”, “generalization”
  - Every precise use of “uncertainty” has this issue
    - E.g. frequentist sampling, Bayesian, etc.
  - We should always make sure we can distinguish what is, and what is not, covered by the term of art
- A standard setup (our setup so far):

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. “significance”, “bias”, “generalization”
  - Every precise use of “uncertainty” has this issue
    - E.g. frequentist sampling, Bayesian, etc.
  - We should always make sure we can distinguish what is, and what is not, covered by the term of art
- A standard setup (our setup so far):
  - We model the data as generated according to a GP with squared exponential kernel and observation noise

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. “significance”, “bias”, “generalization”
  - Every precise use of “uncertainty” has this issue
    - E.g. frequentist sampling, Bayesian, etc.
  - We should always make sure we can distinguish what is, and what is not, covered by the term of art
- A standard setup (our setup so far):
  - We model the data as generated according to a GP with squared exponential kernel and observation noise
  - We fit the hyperparameters (the signal variance, the length scale(s), and the noise variance) to single values

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. “significance”, “bias”, “generalization”
  - Every precise use of “uncertainty” has this issue
    - E.g. frequentist sampling, Bayesian, etc.
  - We should always make sure we can distinguish what is, and what is not, covered by the term of art
- A standard setup (our setup so far):
  - We model the data as generated according to a GP with squared exponential kernel and observation noise
  - We fit the hyperparameters (the signal variance, the length scale(s), and the noise variance) to single values
  - The reported uncertainties are what result when the GP model and fitted hyperparameters are exactly correct

What uncertainty are we quantifying?

- It’s worth being aware that data science (ML/stats/AI) often overloads common colloquial terms with terms of art
  - E.g. “significance”, “bias”, “generalization”
  - Every precise use of “uncertainty” has this issue
    - E.g. frequentist sampling, Bayesian, etc.
  - We should always make sure we can distinguish what is, and what is not, covered by the term of art
- A standard setup (our setup so far):
  - We model the data as generated according to a GP with squared exponential kernel and observation noise
  - We fit the hyperparameters (the signal variance, the length scale(s), and the noise variance) to single values
  - The reported uncertainties are what result when the GP model and fitted hyperparameters are exactly correct

Are there other uncertainties that aren’t being quantified here?

Some other sources of uncertainty

[demo1,2,3]

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do?

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
    - What can we do? First: unit test, plot, sense check!
        - Ask what is possible to learn with the data available
        - Multiple random restarts: plot the results
        - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data (is your model what you think it is? check defaults!)

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data (is your model what you think it is? check defaults!)
  - Box: "All models are wrong, but some are useful"

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data (is your model what you think it is? check defaults!)
  - Box: "All models are wrong, but some are useful"
  - What can we do?

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data (is your model what you think it is? check defaults!)
  - Box: "All models are wrong, but some are useful"
  - What can we do? First: unit test, plot, sense check!

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data (is your model what you think it is? check defaults!)
  - Box: “All models are wrong, but some are useful”
  - What can we do? First: unit test, plot, sense check!
    - Can change the mean and/or kernel
      - E.g. local/heteroskedastic models, periodic kernels, linear mean function, many many more

Some other sources of uncertainty

[demo1,2,3]

- There may be multiple sets of substantively different hyperparameter values that are both plausible and consistent with the observed data
  - What can we do? First: unit test, plot, sense check!
    - Ask what is possible to learn with the data available
    - Multiple random restarts: plot the results
    - Bayesian model of the hyperparameters → be careful: expense & interpretation

[demo1,2]

- A GP with your mean & kernel may be meaningfully misspecified for the data (is your model what you think it is? check defaults!)
  - Box: “All models are wrong, but some are useful”
  - What can we do? First: unit test, plot, sense check!
    - Can change the mean and/or kernel
      - E.g. local/heteroskedastic models, periodic kernels, linear mean function, many many more

Extrapolation

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
  - Compare to *interpolation*: estimation/prediction within the observed data

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
  - Compare to *interpolation*: estimation/prediction within the observed data [demo1,2]

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
    - Compare to *interpolation*: estimation/prediction within the observed data [demo1,2]
- When using GPs with a squared exponential kernel:
    - Data points that are more than a handful of length scales from other data points will revert to prior behavior

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
  - Compare to *interpolation*: estimation/prediction within the observed data [demo1,2]
- When using GPs with a squared exponential kernel:
  - Data points that are more than a handful of length scales from other data points will revert to prior behavior
- Note: extrapolation isn't a special issue unique to GPs. It's a fundamentally hard problem for all data analysis methods

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
  - Compare to *interpolation*: estimation/prediction within the observed data [demo1,2]
- When using GPs with a squared exponential kernel:
  - Data points that are more than a handful of length scales from other data points will revert to prior behavior
- Note: extrapolation isn't a special issue unique to GPs. It's a fundamentally hard problem for all data analysis methods
  - To extrapolate, you need to make assumptions

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
  - Compare to *interpolation*: estimation/prediction within the observed data [demo1,2]
- When using GPs with a squared exponential kernel:
  - Data points that are more than a handful of length scales from other data points will revert to prior behavior
- Note: extrapolation isn't a special issue unique to GPs. It's a fundamentally hard problem for all data analysis methods
  - To extrapolate, you need to make assumptions
    - When you have domain knowledge of a system, you might be able to use it to extrapolate

Extrapolation

- *Extrapolation*: Estimation/prediction beyond the observed data
  - Compare to *interpolation*: estimation/prediction within the observed data [demo1,2]
- When using GPs with a squared exponential kernel:
  - Data points that are more than a handful of length scales from other data points will revert to prior behavior
- Note: extrapolation isn't a special issue unique to GPs. It's a fundamentally hard problem for all data analysis methods
  - To extrapolate, you need to make assumptions
    - When you have domain knowledge of a system, you might be able to use it to extrapolate
    - When you're letting a machine learning method use its defaults, it's making assumptions. Do you know what those assumptions are?

More than one input

More than one input

- Our illustrations have almost all been for one input so far

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it’s typical to have more than one input

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it’s typical to have more than one input
- What could go wrong?

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it's typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it’s typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it’s typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are “far away” in high dimensions. Illustration:

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it's typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are "far away" in high dimensions. Illustration:
  - Uniformly randomly sample 10,000 points on $[0,1]^D$

[Binois, Wycoff 2022]

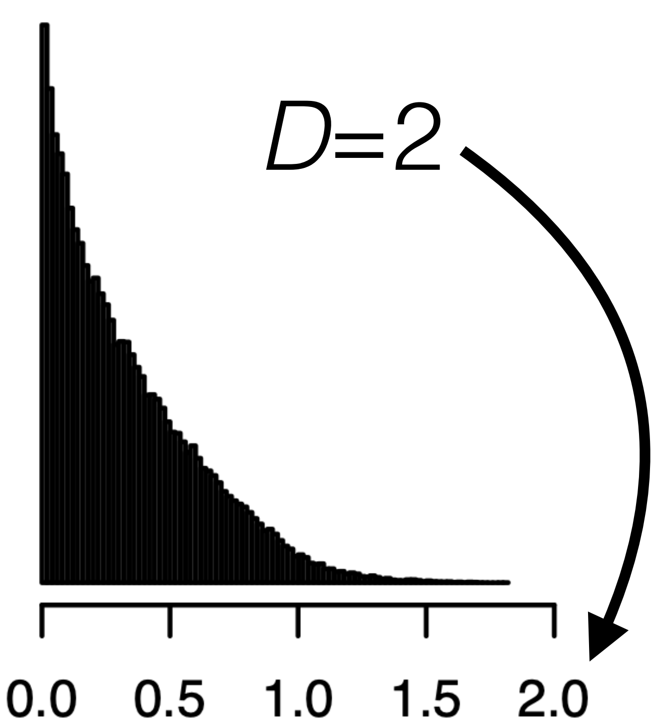
More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it's typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are "far away" in high dimensions. Illustration:
  - Uniformly randomly sample 10,000 points on $[0,1]^D$
  - Make a histogram of squared inter-point distances

[Binois, Wycoff 2022]

More than one input

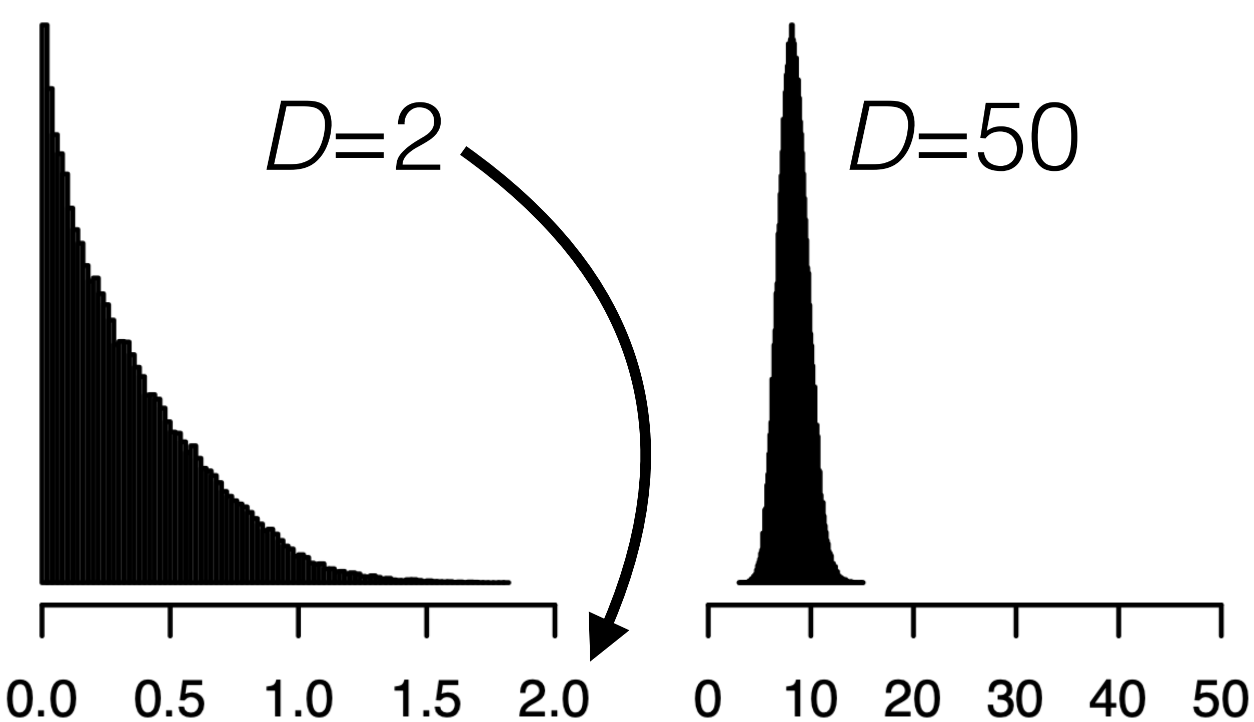
- Our illustrations have almost all been for one input so far
- But in real life, it's typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are "far away" in high dimensions. Illustration:
  - Uniformly randomly sample 10,000 points on $[0,1]^D$
  - Make a histogram of squared inter-point distances



[Binois, Wycoff 2022]

More than one input

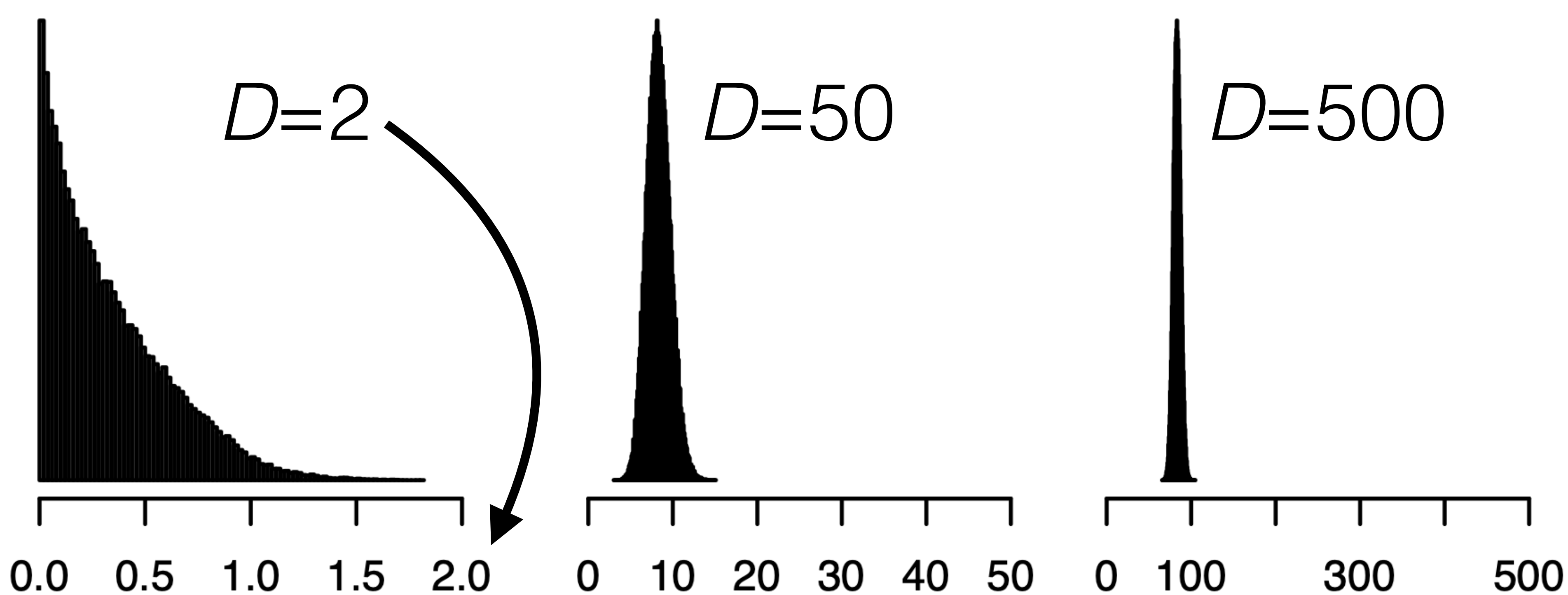
- Our illustrations have almost all been for one input so far
- But in real life, it’s typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are “far away” in high dimensions. Illustration:
  - Uniformly randomly sample 10,000 points on $[0,1]^D$
  - Make a histogram of squared inter-point distances



[Binois, Wycoff 2022]

More than one input

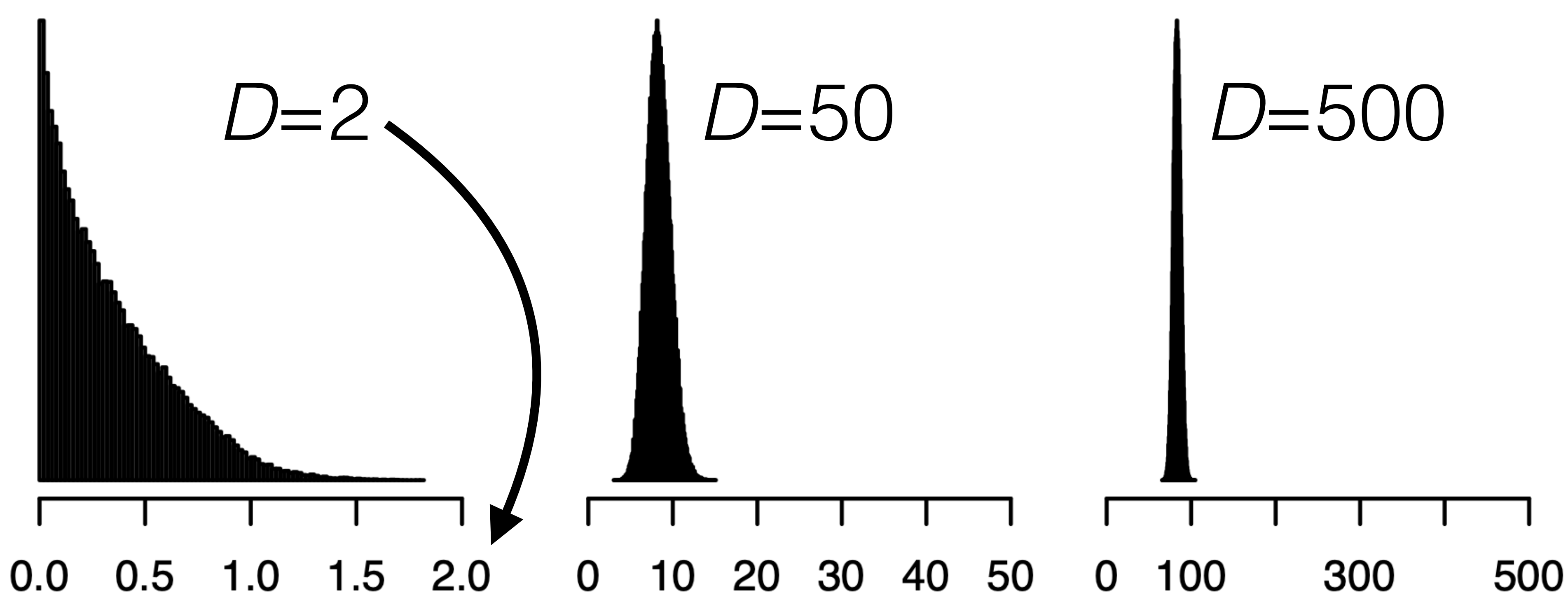
- Our illustrations have almost all been for one input so far
- But in real life, it's typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are "far away" in high dimensions. Illustration:
  - Uniformly randomly sample 10,000 points on $[0,1]^D$
  - Make a histogram of squared inter-point distances



[Binois, Wycoff 2022]

More than one input

- Our illustrations have almost all been for one input so far
- But in real life, it's typical to have more than one input
- What could go wrong? Previous lessons apply, but also:
  - Possibly different lengthscales. Check defaults.
  - Regression in high dimensions is a fundamentally hard problem (without additional assumptions)
- All points are "far away" in high dimensions. Illustration:
  - Uniformly randomly sample 10,000 points on $[0,1]^D$
  - Make a histogram of squared inter-point distances



[Binois, Wycoff 2022]

- Recall: points "far" from data default to the prior mean and variance

Some high points of what got cut for time

- We ran out of time! Here are some high-level summary points beyond what we discussed together:
  - There are other challenges with many inputs, both conceptual and practical
  - Running time for GP regression can be an issue with a large number of training data points
    - In particular, the matrix inverse can be expensive
    - There are incredibly many papers about fast approximations to the exact Gaussian process
      - Each approximation has pros and cons
- Bayesian optimization inherits many of the pros and cons of Gaussian processes for regression
  - Exercise: once you learn about Bayesian optimization, think about how the pros and cons we discussed together might translate there

Roadmap

- Bayesian modeling and inference
- Gaussian process model
  - Popular version using a squared exponential kernel
- Gaussian process inference
  - Prediction & uncertainty quantification
- Observation noise
- What uncertainty are we quantifying?
- What can go wrong?
- Bayesian optimization

- Goals:
  - Learn the mechanism behind standard GPs to identify benefits and pitfalls (also in BayesOpt)
  - Learn the skills to be responsible users of standard GPs (transferable to other ML/AI methods)