# CS761 Spring 2017 Homework 3

Assigned Apr. 6, due Apr. 20

Instructions:

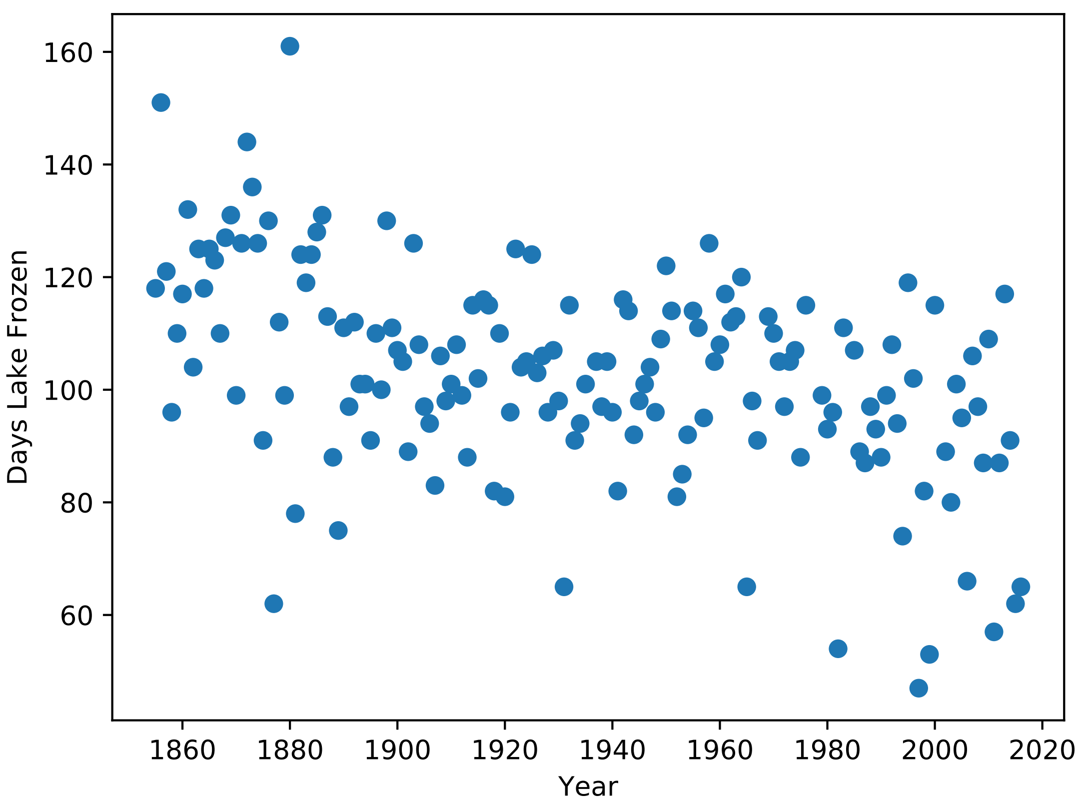
- Homeworks are to be done individually.
- Typeset your homework in latex using this file as template (e.g. use pdflatex). Show your derivations.
- Hand in the compiled pdf (not the latex file) online. Instructions will be provided. We do not accept hand-written homeworks.
- Homework will no longer be accepted once the lecture starts.
- Fill in your name and email below.

Name: Max Horowitz-Gelb
Email: horowitzgelb@wisc.edu

(4 questions, 25 points each)

1. The Wisconsin State Climatology Office keeps a record on the number of days Lake Mendota was covered by ice at `http://www.aos.wisc.edu/~sco/lakes/Mendota-ice.html`. The article DETERMINING THE ICE COVER ON MADISON LAKES at `http://www.aos.wisc.edu/~sco/lakes/msn-lakes_instruc.html` serves as a fine example of the Wisconsin tradition to integrate science with beer.

   (a) As with any real problems, the data is not as clean nor as organized as one would like for machine learning. Produce a clean data set starting from 1855-56 and ending in 2016-17 for the output variable DAYS. You do not need to attach your data set, but please produce a scatter plot of year vs. DAYs. Show us the sample mean and sample variance (round to 5 digits after decimal point).




**Mean:** = 102.80769 **Variance:** = 343.57840

(b) Perform ordinary least squares to estimate a linear model

$$y = \alpha + \beta x$$

where $y$ is DAYS and $x$ is the year. For example, for 1855-56 the year is 1855. Show us $\hat{\alpha}, \hat{\beta}$, and an estimate of the standard error on $\beta$: $\widehat{s.e}(\hat{\beta})$.

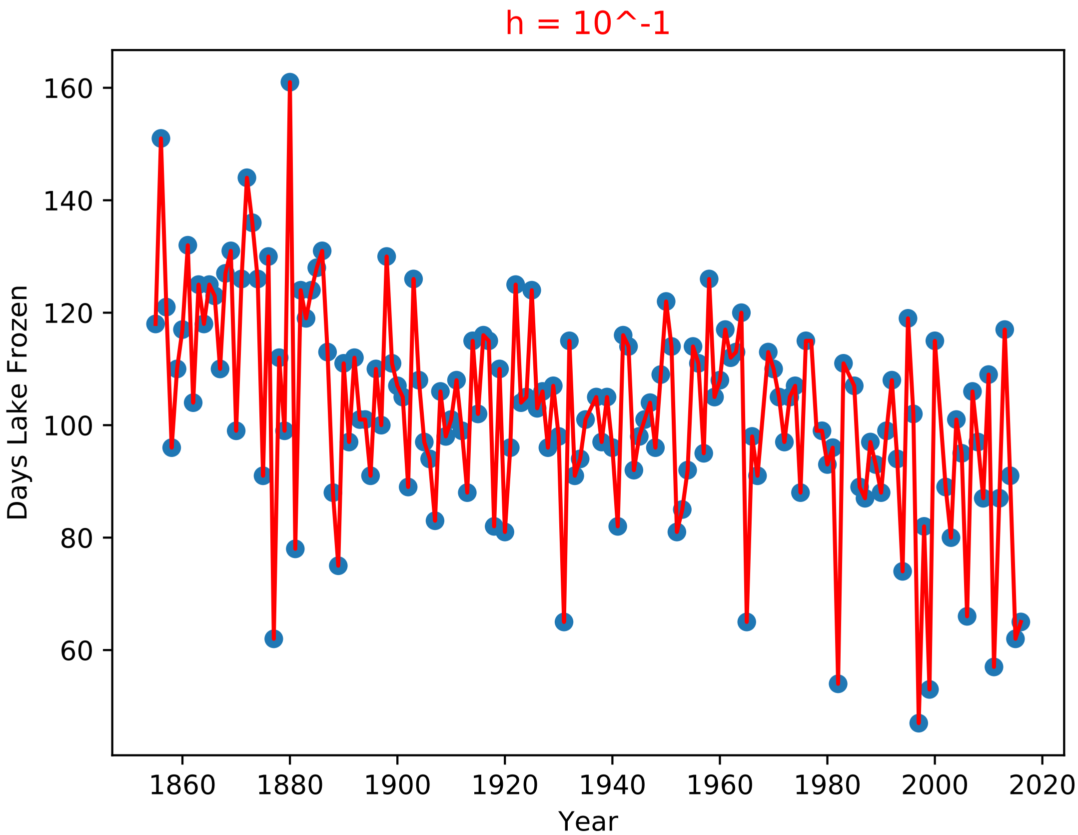
$\hat{\beta} = -0.18561 \qquad \hat{\alpha} = 461.78577 \qquad s\hat{.}e\{\hat{\beta}\} = \sqrt{s^2(X^\top X)^{-1}} = 0.00068$

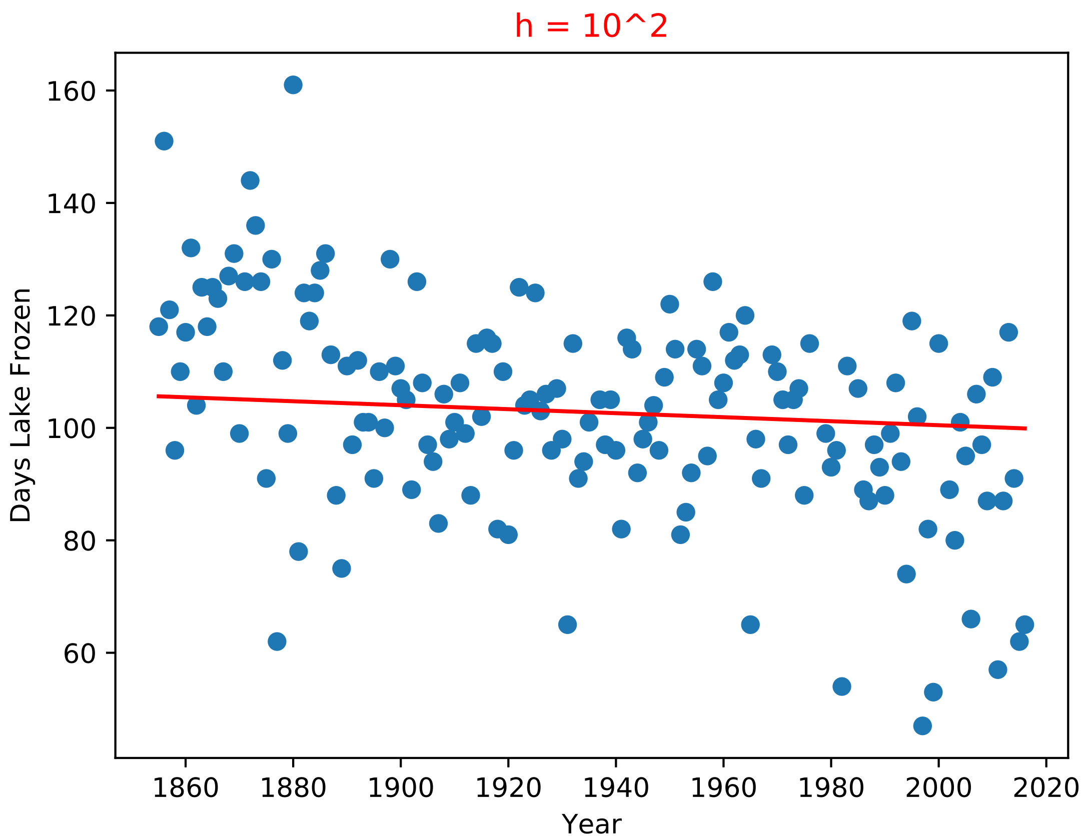
(c) Perform nonparametric kernel regression using the Nadaraya-Watson estimator on this data set (input: year, output: days). Use the Gaussian kernel. Write your own code for the Nadaraya-Watson estimator. Show us the leave-one-out score (Equation 23 in lecture notes `http://pages.cs.wisc.edu/~jerryzhu/cs761/kde.pdf`) for bandwidth $h = 10^{-1}, 10^{-0.9}, 10^{-0.8}, \ldots, 10^{2}$, respectively.

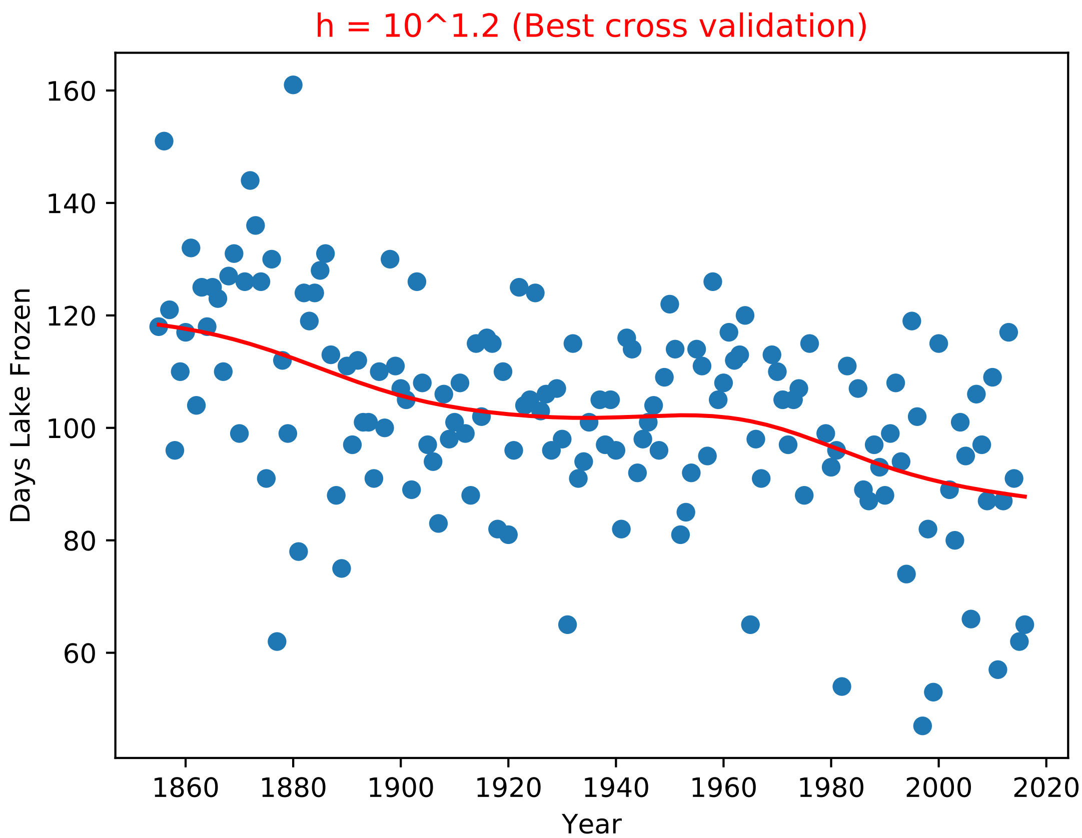
```
h = 10^-1, LeaveOneOut Score = NaN
h = 10^-0.9, LeaveOneOut Score = 470.15966
h = 10^-0.8, LeaveOneOut Score = 472.28524
h = 10^-0.7, LeaveOneOut Score = 472.28526
h = 10^-0.6, LeaveOneOut Score = 472.28526
h = 10^-0.5, LeaveOneOut Score = 472.28511
h = 10^-0.4, LeaveOneOut Score = 472.24836
h = 10^-0.3, LeaveOneOut Score = 471.07843
h = 10^-0.2, LeaveOneOut Score = 461.79288
h = 10^-0.1, LeaveOneOut Score = 435.06801
h = 10^0, LeaveOneOut Score = 398.43935
h = 10^0.1, LeaveOneOut Score = 366.57279
h = 10^0.2, LeaveOneOut Score = 343.70665
h = 10^0.3, LeaveOneOut Score = 327.49724
h = 10^0.4, LeaveOneOut Score = 315.25907
h = 10^0.5, LeaveOneOut Score = 305.71688
h = 10^0.6, LeaveOneOut Score = 298.58151
h = 10^0.7, LeaveOneOut Score = 292.96769
h = 10^0.8, LeaveOneOut Score = 287.56533
h = 10^0.9, LeaveOneOut Score = 282.08339
h = 10^1, LeaveOneOut Score = 277.3443
h = 10^1.1, LeaveOneOut Score = 274.17402
h = 10^1.2, LeaveOneOut Score = 272.96093
h = 10^1.3, LeaveOneOut Score = 273.91815
h = 10^1.4, LeaveOneOut Score = 277.05608
h = 10^1.5, LeaveOneOut Score = 281.81211
h = 10^1.6, LeaveOneOut Score = 287.40215
```

```
h = 10^1.7, LeaveOneOut Score = 294.11911
h = 10^1.8, LeaveOneOut Score = 302.82929
h = 10^1.9, LeaveOneOut Score = 312.90299
```

(d) For $h = 10^{-1}, 10^2$ and the optimal $h$ you found, respectively, plot the function estimated by Nadaraya-Watson.




h = 10^2
160
140
120
100
80
60
Days Lake Frozen
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year




2. Consider a Gaussian Process $f \sim GP(m, k)$ over $\mathbb{R}$ with mean function

$$m(x) = \sin(\frac{\pi x}{100}) + \frac{x}{100}$$

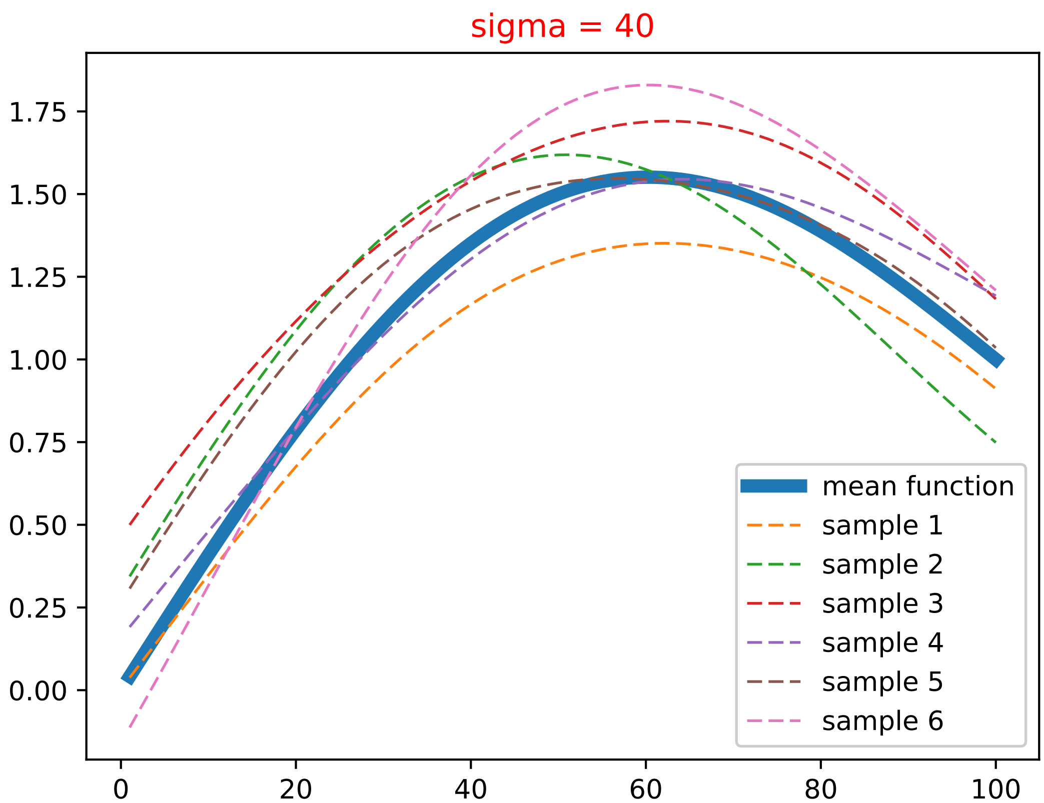
and kernel function

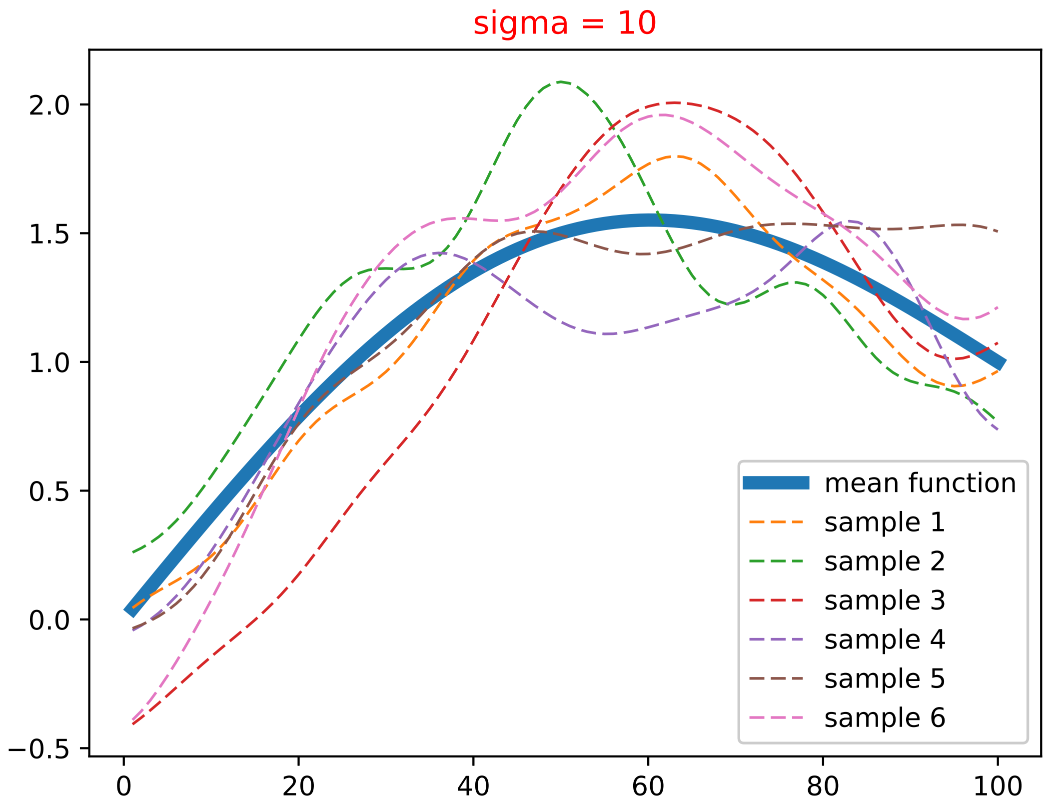
$$k(x, x') = \frac{1}{16} \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right).$$

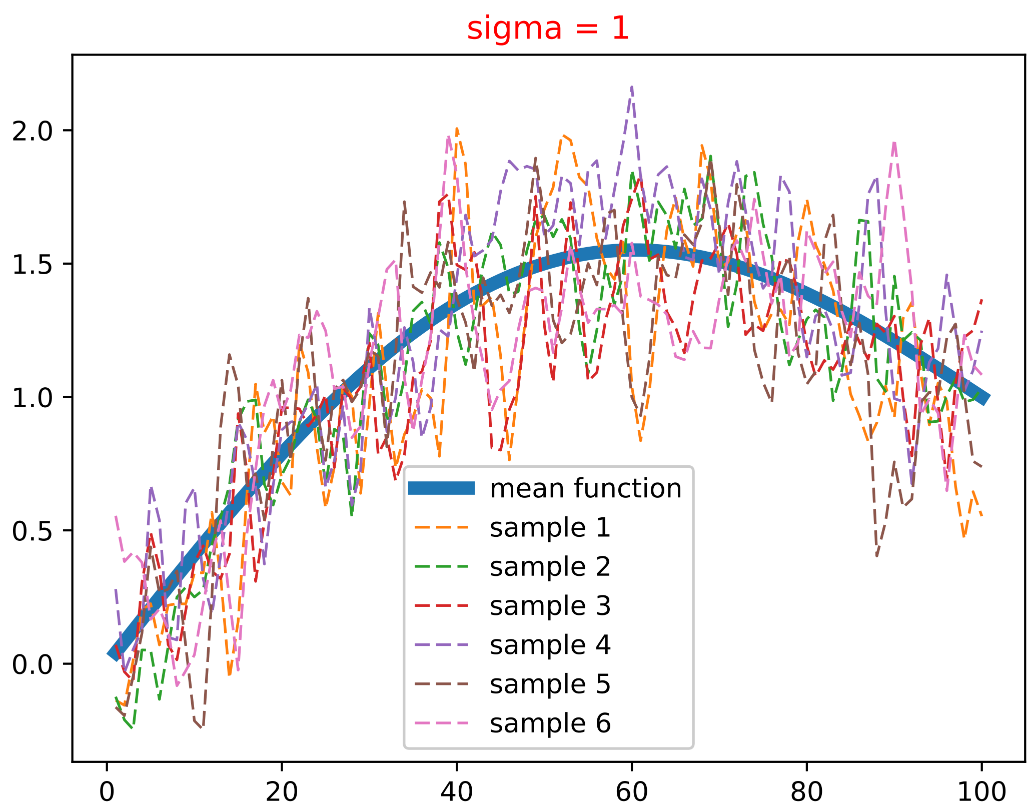
(a) Let $\sigma = 40$ (note: this is the standard deviation, not variance). Approximate the random function $f$ by drawing $f(1), f(2), \ldots, f(100)$ from the appropriate marginal distribution. Plot the curve by connecting the dots. Show six such random functions on the same plot, together with the mean function $m$.

(b) Do the same with $\sigma = 10$.

(c) Do the same with $\sigma = 1$.



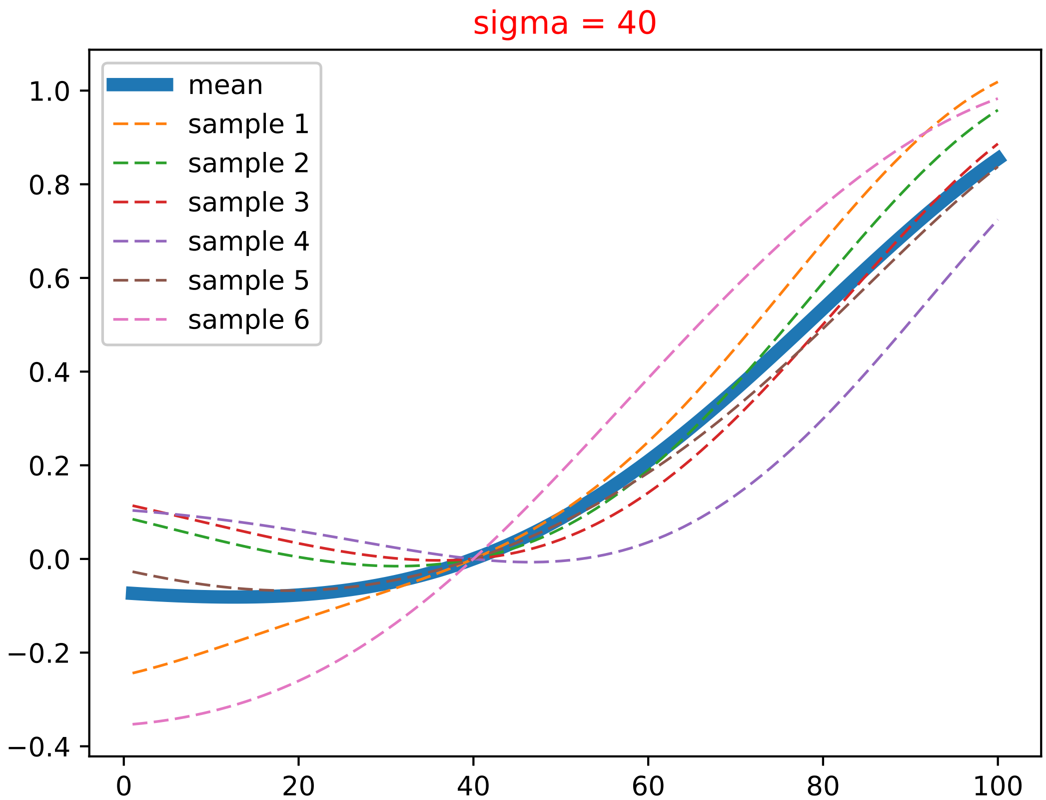

sigma = 10
2.0
1.5
1.0
0.5
0.0
−0.5
0
20
40
60
80
100
mean function
sample 1
sample 2
sample 3
sample 4
sample 5
sample 6

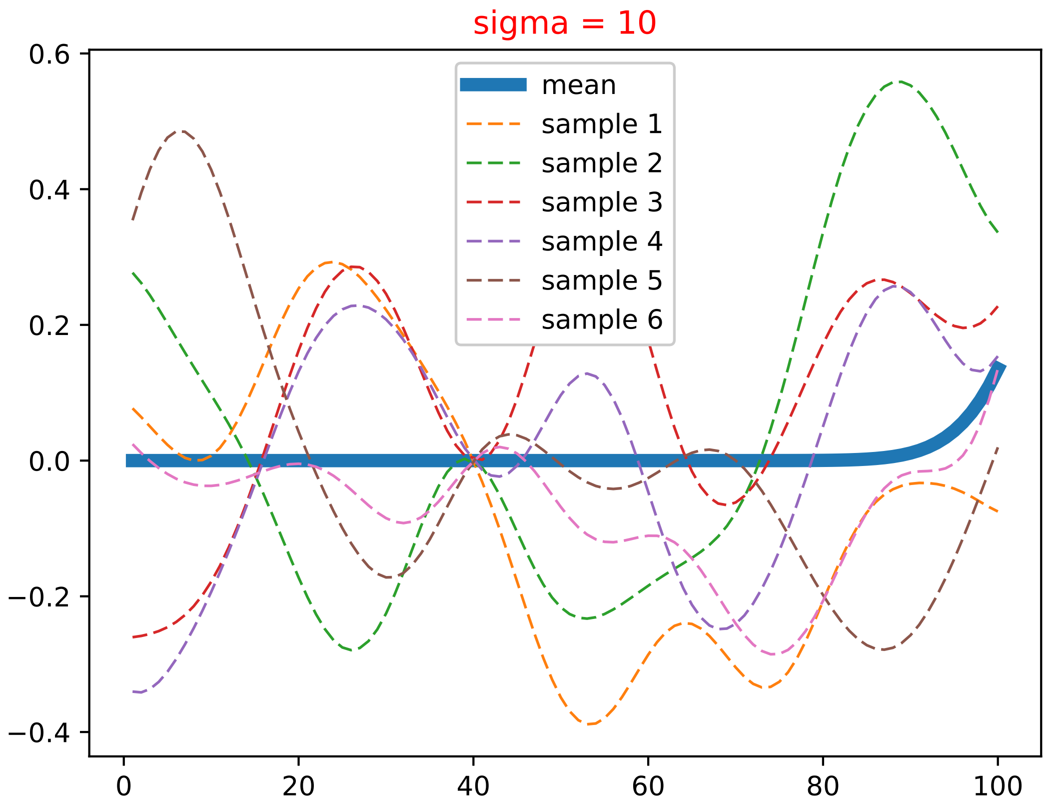



(d) Let $\sigma = 40$. Now let us observe $f(40) = 0$ and $f(120) = 1$. Now draw $f$ from the posterior Gaussian Process conditioned on these two observations. Again, show six such $f$ from the posterior on the same plot.

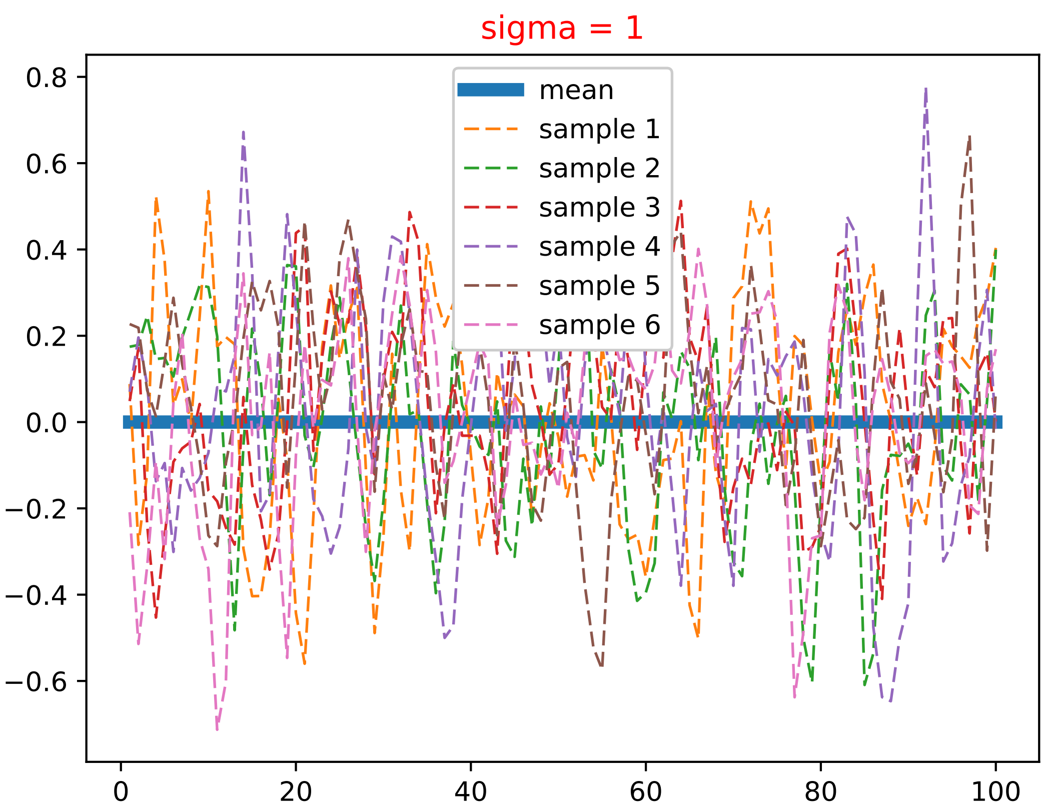
(e) Do the same with $\sigma = 10$.

(f) Do the same with $\sigma = 1$.


sigma = 40
mean
sample 1
sample 2
sample 3
sample 4
sample 5
sample 6
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
20
40
60
80
100



sigma = 10
mean
sample 1
sample 2
sample 3
sample 4
sample 5
sample 6
0.6
0.4
0.2
0.0
−0.2
−0.4
0
20
40
60
80
100




3. Imagine a stick of length $a$. On the ground, draw parallel lines $a$ apart. Randomly throw the stick to the ground. Each time, the stick may or may not intersect with a line.

   (a) What is the probability that the stick intersects with a line? Show your work.
   Let $\theta$ be the angle of the stick from the horizontal, and $\delta$ be the closest distance from the center of the stick to a line. Then if we assume $\theta \perp \delta$ and $\theta \sim Uniform(0, \pi)$ and $\delta \sim Uniform(0, a)$ then the probability that the stick intersects a line is

   $$P(\delta \leq a/2sin(\theta)) = \int_0^{\pi} \int_0^{a/2\sin(\theta)} \frac{2}{a\pi} d\delta d\theta = 2/\pi$$

   (b) Propose a Monte Carlo method for estimating $\pi$ based on this.
   Let $X_1, X_2, ...X_n$ be a random sample of binary variables where 1 indicates dropping a stick resulting in it crossing a line. Then by

the Law of large numbers as $n$ goes to infinity the sample mean, $\bar{X}_n$ converges in probability to $2/\pi$. So a method would be sample for large $n$ to acquire $\bar{X}_n$, then $\pi \approx 2 * \bar{X}^{-1}$

(c) Actually perform the experiment. Tell us about it.

Note due to time constraints I was forced to simulate this experiment on the computer. I calculated $\bar{X}$ for a random sample of 100 sticks and estimated

$$\pi \approx 3.125$$

4. Consider an undirected graphical model on a binary tree with 15 nodes. Each node takes value in $\{-1, 1\}$. All edges share the same potential function $\psi(u, v) = \exp(\alpha uv)$, where $u, v$ are a pair of parent-child nodes.

(a) Write down the joint probability distribution defined by this graphical model.

Let $\text{par}(x_i)$ be the parent of node $x_i$ and let $r$ be the root. Then

$$p(x_1, ..., x_{15}) = \frac{1}{Z} \prod_{i \neq r} \exp(a x_i \text{par}(x_i))$$

Where

$$Z = \sum_{x' \in \{-1,1\}^{15}} \left( \prod_{i \neq r} \exp(a x'_i \text{par}(x'_i)) \right)$$

(b) Let $\alpha = 1$. Let $r$ be the root node and $s$ be the left-most leaf node. Use brute force (enumerating all trees) to compute $p(r \mid s = 1)$. Using brute force,

$$P(r = 1 | s = 1) \approx 0.72087$$

(c) Implement Gibbs sampling to estimate $p(r \mid s = 1)$. Start with the all-minus-1 tree except for $s = 1$. Go over levels in top-down order, left-to-right within each level.

Using Gibbs Sampling,

$$P(r = 1 | s = 1) \approx 0.72472$$

Discard a burn-in of $10^4$ samples. Use the next $10^5$ samples for estimation. Do not perform thinning.

(d) Implement Metropolis-Hastings sampling to estimate $p(r \mid s = 1)$. Clearly define and discuss your proposal distribution (which has to be different than Gibbs). Use the same burn-in and number of samples as above.

For our proposal distribution we will assume all nodes are independent with equal probability of being -1 or 1. So

$$q(x'|x^t) = \mathbf{1}_{x'_s=1} * 2^{-14}$$

Using this proposal and the same burn in and sample size we get,

$$P(r = 1|s = 1) \approx 0.67679$$