Lecture #6: Monte Carlo Integration

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020

IACS

VE RI TAS

Outline

1. Basics of Monte Carlo Simulation
2. Variance Reduction: Control Variates
3. Variance Reduction: Stratified Sampling
4. Variance Reduction: Importance Sampling
5. Application: Monte Carlo Estimation of Arbitrary Integrals

Motivation

Posterior samples allow us to approximately compute posterior point estimates, for example, we can approximate the posterior mean as

$$\mathbb{E}_{\theta|Y}[\theta] = \int_{\Theta} \theta \, p(\theta|Y) \, d\theta \approx \frac{1}{S} \sum_{s=1}^{S} \theta_s, \; \theta_s \sim p(\theta|Y)$$

In fact, for any function f of $\theta|Y$, we can estimate the expected vaue of f by first sampling S samples from the posterior $p(\theta|Y)$ and then compute the average value of f on these samples:

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \int_{\Theta} f(\theta) \, p(\theta|Y) \, d\theta \approx \frac{1}{S} \sum_{s=1}^{S} f(\theta_s), \; \theta_s \sim p(\theta|Y)$$

Question: But is this estimate consistent? Unbiased? Of minimal variance?

Basics of Monte Carlo Integration

Naive Monte Carlo Estimation of Integrals

Let I denote the integral

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \int_{\Theta} f(\theta)\, p(\theta|Y)\, d\theta$$

and let $\widehat{I}$ denote the approximation

$$\frac{1}{S}\sum_{s=1}^{S} f(\theta_s),\ \theta_s \sim p(\theta|Y).$$

We call $\widehat{I}$ the ***Monte Carlo estimate*** of I.

In general, ***Monte Carlo integration*** is the process of estimating a deterministic quantity (an integral) using a procedure involving stochasticity (sampling).

The Consistency and Unbiasedness of Monte Carlo Estimators

Recall that the **Strong Law of Large Numbers** says that the sample mean of S iid random variables converges to the theoretical mean, with probability 1, as $S \to \infty$. This means that

$$\lim_{S\to\infty} \frac{1}{S} \sum_{s=1}^{S} f(\theta_s),\ \theta_s \sim p(\theta|Y) \ = \ \mathbb{E}_{\theta|Y}\left[f(\theta)\right]$$

with probability 1. Hence, the Monte Carlo Estimator $\hat{I}$ is ***consistent***.

The expected value of $\hat{I}$ is

$$\mathbb{E}_{\theta_S}[\hat{I}] \ = \ \mathbb{E}_{\theta_S}\left[\frac{1}{S}\sum_{s=1}^{S} f(\theta_s)\right] \ = \ \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{\theta_s}\left[f(\theta_s)\right] = \frac{1}{S}\sum_{s=1}^{S} I = I,$$

where $\theta_s \sim p(\theta|Y)$. Hence, the Monte Carlo Estimator $\hat{I}$ is ***unbiased***.

The Variance and Error of Monte Carlo Estimators

The variance of $\widehat{I}$ is given by

$$\mathrm{Var}\left[\hat{I}\right] = \mathrm{Var}\left[\frac{1}{S}\sum_{s=1}^{S} f(\theta_s)\right] = \frac{1}{S^2}\sum_{s=1}^{S}\mathrm{Var}\left[f(\theta_s)\right] = \frac{\mathrm{Var}\left[f(\theta)\right]}{S},$$

where $\theta_s, \theta \sim p(\theta|Y)$. Plainly put, **the variance of the estimator is reduced when number of samples $\boldsymbol{S}$ is large and the variance $\boldsymbol{\sigma_f^2 = \mathrm{Var}\left[f(\theta)\right]}$ of $\boldsymbol{f(\theta)}$ is low**.

The **Central Limit Theorem** says that the differnece between sample mean and the theoretical mean of iid random variables will approach a normal distribution as $N \to \infty$. For us, this means that the error of $\widehat{I}$ has a roughly normal distribution:

$$p\left(\hat{I} - I\right) \to \mathcal{N}\left(0, \frac{\sigma_f^2}{S}\right), \mathcal{N} \to \infty$$

Again, this says that, to reduced the error of $\widehat{I}$ we can increase the number of samples S or decrease the variance σ_f^2 of $f(\theta)$.

Variance Reduction: Control Variates

A Baseline for Variance of Monte Carlo Estimates

The variance of the naive Monte Carlo estimate $\hat{I}$ is:

$$\begin{aligned}
\mathrm{Var}\left[\hat{I}\right] &= \frac{1}{S}\mathrm{Var}\left[f(\theta)\right] \\
&= \mathbb{E}_{\theta|Y}\left[\left(f(\theta) - \mathbb{E}_{\theta|Y}\left[f(\theta)\right]\right)^2\right] \\
&= \mathbb{E}_{\theta|Y}\left[f(\theta)^2\right] - \mathbb{E}_{\theta|Y}[f(\theta)]^2 \\
&= \int_\Theta f(\theta)^2 p(\theta|Y)d\theta - I^2
\end{aligned}$$

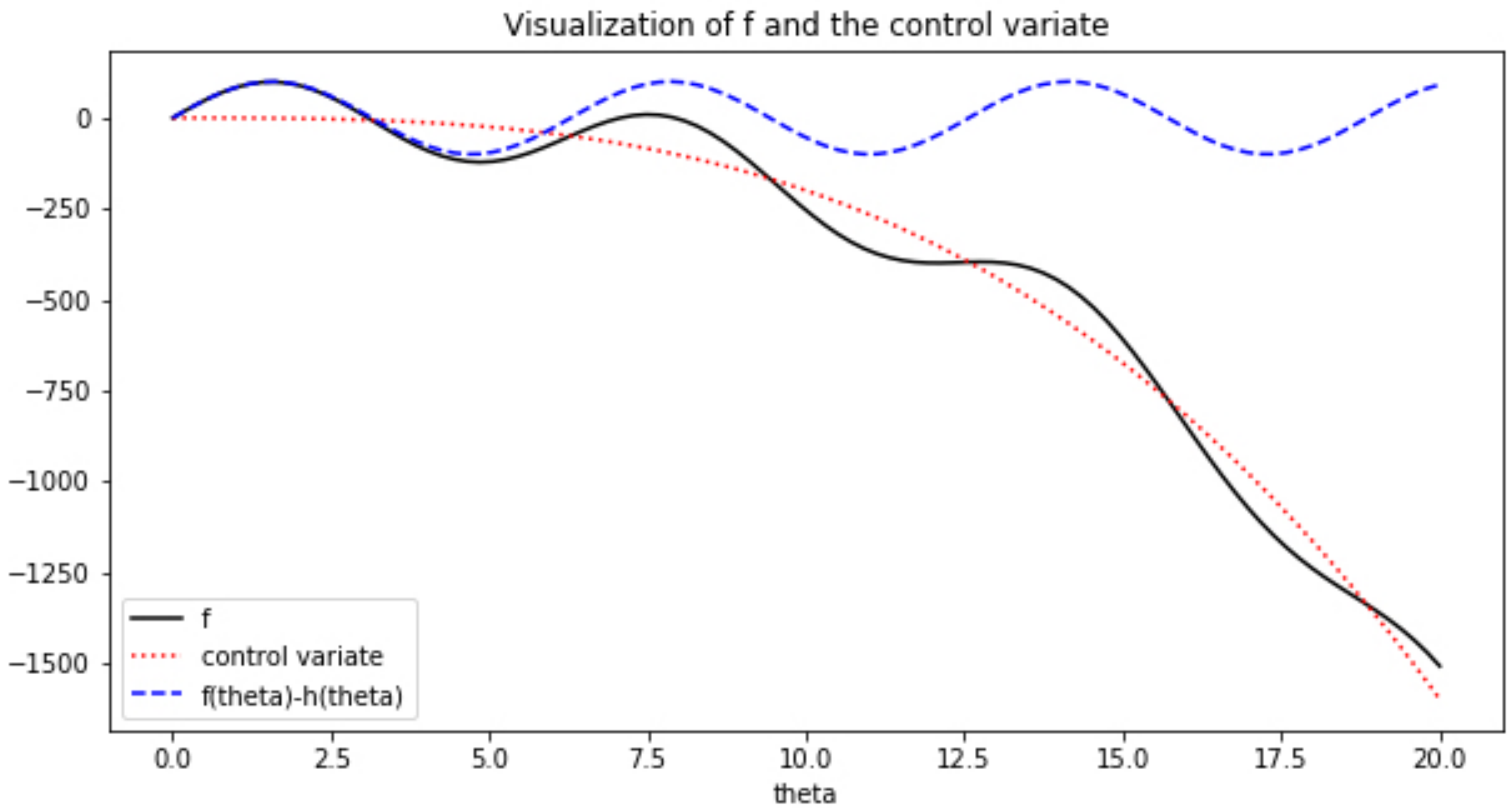
We'll compare the variance of alternative estimator to the above when checking for variance reduction.

Note: The variance of $f(\theta)$ depends on *both* the variance of θ and the amount of variation in the function f. **Functions that change a great amount over the domain will have a high variance and functions that are flat will have a low variance.**

The General Idea of Control Variates

Based on our realization that "flat" functions have lower variance, we will try to engineer "flat" functions that allow use to compute the integral of f.

Observation: For a given function $f(X)$, if h is a function of X that correlates with f, then $f(X) - h(X)$ will have lower variance.



Processing math: 100%

Variance of Control Variates

Fix a function $f(\theta), \theta \sim p(\theta|Y)$. Let $h(\theta)$ be a function with known mean $\mu_h = \mathbb{E}_{\theta|Y}[h(\theta)]$ and such that $h(\theta)$ is correlated with $f(\theta)$. We call h the ***control variate*** for f.

We define the ***control variate Monte Carlo estimate*** of $\mathbb{E}_{\theta|Y}[f(\theta)]$ to be

$$\hat{I}_{\text{control}} = \frac{1}{S}\sum_{s=1}^{S} f(\theta) - c(h(\theta) - \mu_h)$$

where c is our choice of a constant.

The variance of this estimator is:

$$\sigma^2_{\hat{I}_{\text{control}}} = \sigma^2_f + c^2\sigma^2_h - 2c * \mathbf{cov}[f(\theta), h(\theta)] / S$$

Thus, we see that when $\mathbf{cov}[f(\theta), h(\theta)] \neq 0$ there is hope for variance reduction, $\sigma^2_{\hat{I}_{\text{control}}} < \sigma^2_f / S$!

Exercise: check that $\hat{I}_{\text{control}}$ is consistent and unbiased! Derive $\sigma^2_{\hat{I}_{\text{control}}}$!

The Nitty Gritty of Control Variates

Using control variate Monte Carlo requires that we:

1. choose a control variate h with known mean and who is correlated with f.
2. choose a constant c.

Typically, we want to choose an h that follows to the trends of f that is easy to integrate.

The value of c that minimizes $\sigma^2_{\hat{I}_{\text{control}}}$ is

$$c^* = \frac{\mathbf{cov}[f(\theta), h(\theta)]}{\sigma_h^2}.$$

In case $\mathbf{cov}[f(\theta), h(\theta)]$ and σ_h^2 are difficult to compute analytically, they can be empirically estimated from the samples.

When we choose the optimal $c = c^*$, the variance of $\widehat{I}_{\text{control}}$ is

$$\sigma^2_{\hat{I}_{\text{control}}} = \left(1 - \frac{\mathbf{cov}[f(\theta), h(\theta)]}{\sigma_h^2 \sigma_f^2}\right)\sigma_f^2 / S \quad = (1 - \rho_{f,h})\sigma_f^2 / S,$$

where $\rho_{f,h}$ is the correlation between the outputs of f and h. Hence, the more h is correlated with the ouput the greater the variance reduction.

Example: Control Variates

Let $f(\theta) = -0.2\theta^3 + 100\sin(\theta)$ for $\theta|Y \sim U(0, 20)$. Let the control variate be $h(\theta) = -0.2\theta^3$. We aim to estimate:

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \int_\Theta f(\theta)p(\theta|Y)d\theta = \frac{1}{20}\int_\Theta f(\theta)d\theta.$$

We compare the sample variance of $\widehat{I} = \frac{1}{S}\sum_{s=1}^{S} f(\theta_s)$ with that of the Monte Carlo estimate with the control variate:

$$\widehat{I}_{\text{control}} = \frac{1}{S}\sum_{s=1}^{S}\left(f(\theta_s) - 1.5(h(\theta_s) - \mu_h)\right),$$

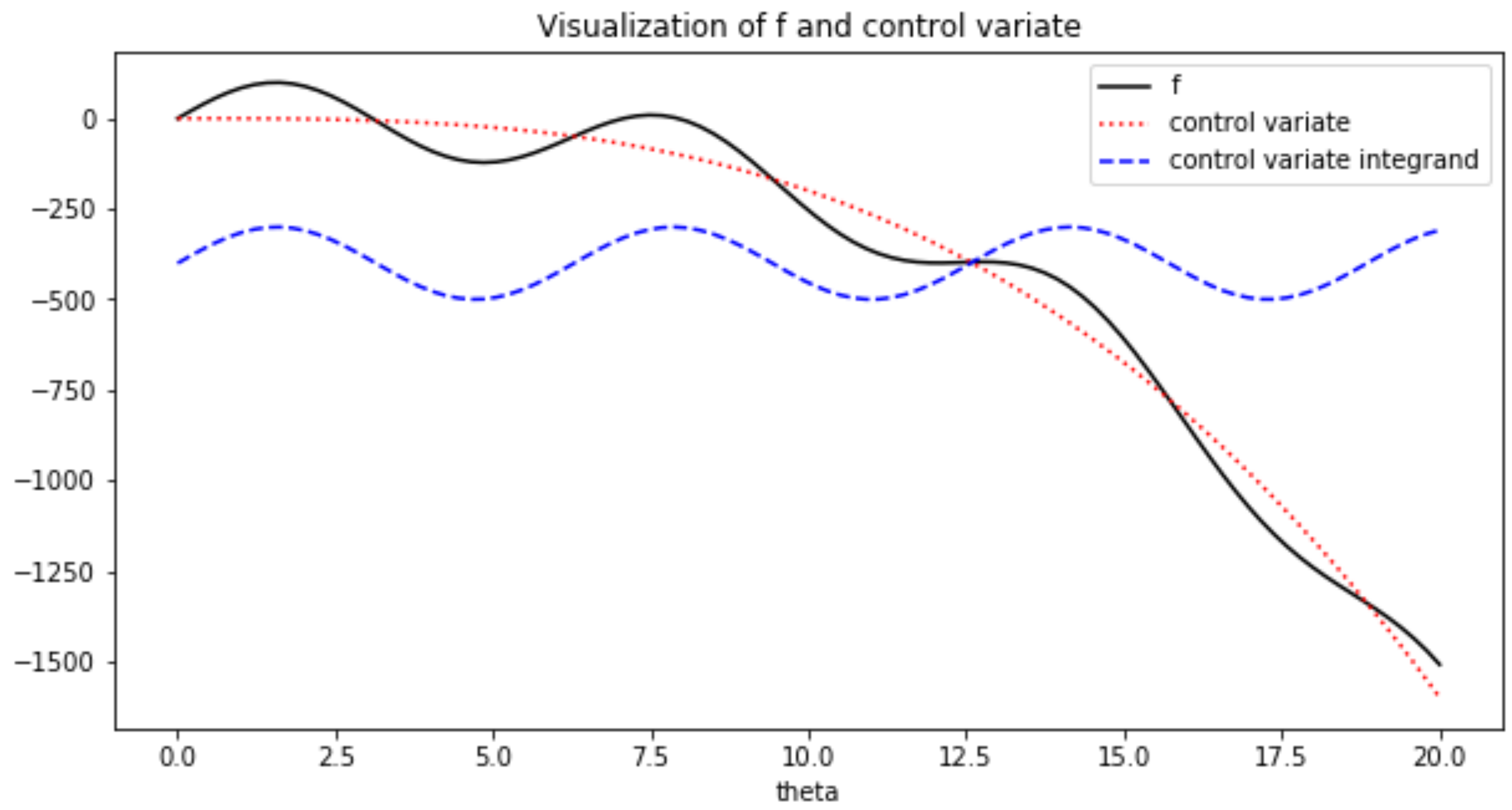
for $S = 100$ and where

$$\mu_h = \mathbb{E}_{\theta|Y}[h(\theta)] = \frac{1}{20}\int_0^{20} h(\theta)d\theta = -400.$$

Example: Control Variates

In [6]:

```
x = np.linspace(0, 20, 100)
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax.plot(x, f(x), color='black', label='f')
ax.plot(x, h(x),linestyle=':', color='red', label='control variate')
ax.plot(x, control_mc(x), linestyle='--', color='blue', label='control variate i
ntegrand')
ax.set_title('Visualization of f and control variate')
ax.set_xlabel('theta')
ax.legend(loc='best')
plt.show()
```



Example: Control Variates

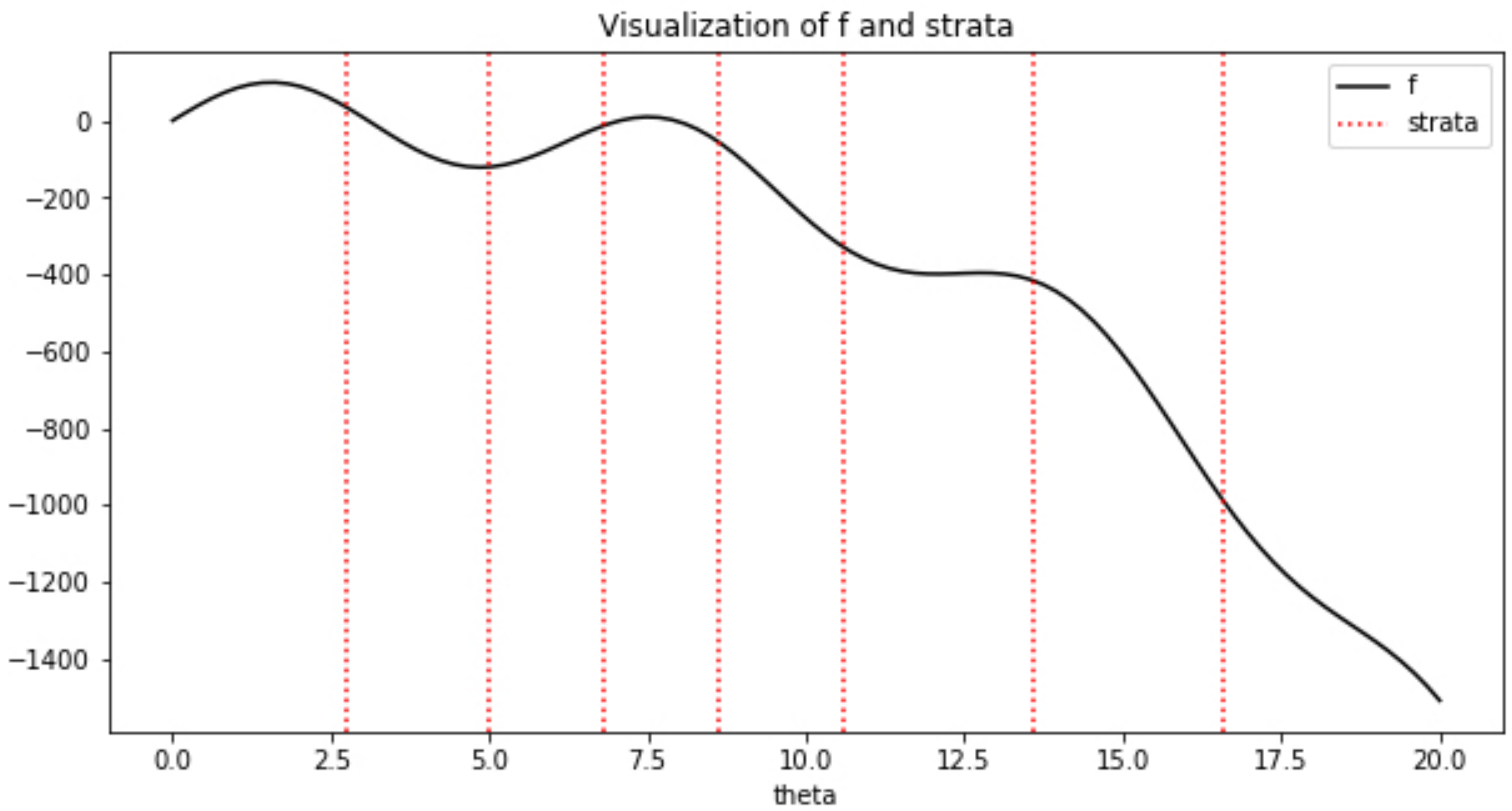
In [7]:
```
print('variance of monte carlo estimate:', mc_variance)
print('variance of control variate monte carlo estimate:', control_mc_variance)
```

```
variance of monte carlo estimate: 2002.2389242017346
variance of control variate monte carlo estimate: 51.052475302251885
```

Variance Reduction: Stratified Sampling

The General Idea of Stratified Sampling

Again, based on our realization that "flat" functions have lower variance, we will try to partition the domain of f into regions where f is relatively "flat". We estimate the integral on each piece of the partition then sum over the pieces.



The Stratified Sampling Monte Carlo Estimator

Fix a function $f(\theta)$, where θ is a random variable over the domain Θ whose pdf is $p(\theta|Y)$. Then the integral for $\mathbb{E}_{\theta|Y}[f(\theta)]$ can be decomposed as

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \int f(x)p(x)dx = \sum_{m=1}^{M} \int_{\Theta_m} f(\theta)p(\theta)d\theta$$

where $\bigcup_{m=1}^{M} \Theta_m = \Theta$ and $\Theta_m \cap \Theta_{m'} = \emptyset$ if $m \neq m'$.

To make $p(\theta)$ a pdf over $\Theta_{m'}$ we need to normalize it, i.e.

$$\frac{p(\theta)}{w_m}, \quad w_m = \int_{\Theta_m} p(\theta)d\theta.$$

We can then rewrite the integral for $\mathbb{E}_{\theta|Y}[f(\theta)]$ again

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \sum_{m=1}^{M} \int_{\Theta_m} w_m f(\theta) \left(\frac{p(\theta)}{w_m}\right) d\theta = \sum_{m=1}^{M} w_m \mathbb{E}_{\theta \sim \frac{p(\theta)}{w_m} \mathbb{1}_{\Theta_m}}[f(\theta)],$$

where $\mathbb{1}_{\Theta_m}$ is the indicator function for the set Θ_m.

Fix a total sample size S. Let $\widehat{I}_m$ be the Monte Carlo estimate of $\mathbb{E}_{\theta \sim \frac{p(\theta)}{w_m} \mathbb{1}_{\Theta_m}}[f(\theta)]$ estimated with $S_m = w_m * S$ samples, the ***stratified sampling Monte Carlo estimator*** is

$$\widehat{I}_{\text{SS}} = \sum_{m=1}^{M} w_m \widehat{I}_m.$$

since $\text{Var}\left[\widehat{I}_m\right]$ is lower than $\text{Var}\left[\widehat{I}\right]$, we hope that $\text{Var}\left[\widehat{I}_{\text{SS}}\right]$ is also lower.

Example: Stratified Sampling

Again, let $f(\theta) = -0.2\theta^3 + 100\sin(\theta)$ for $\theta|Y \sim U(0, 20)$. We aim to estimate:

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \int_\Theta f(\theta)p(\theta|Y)d\theta = \frac{1}{20}\int_\Theta f(\theta)d\theta.$$

We compare the sample variance of $\hat{I} = \frac{1}{S}\sum_{s=1}^{S} f(\theta_s)$ with that of the Monte Carlo estimate with stratified sampling:

$$\hat{I}_{\mathrm{SS}} = \sum_{m=1}^{M} \frac{1}{S_m} \sum_{s_m=1}^{S_m} w_m f(\theta_{s_m}),\ \ \theta_{s_m} \sim U(t_{m-1}, t_m),$$

for sample sizes

$$[13, 20, 9, 10, 15, 32],$$

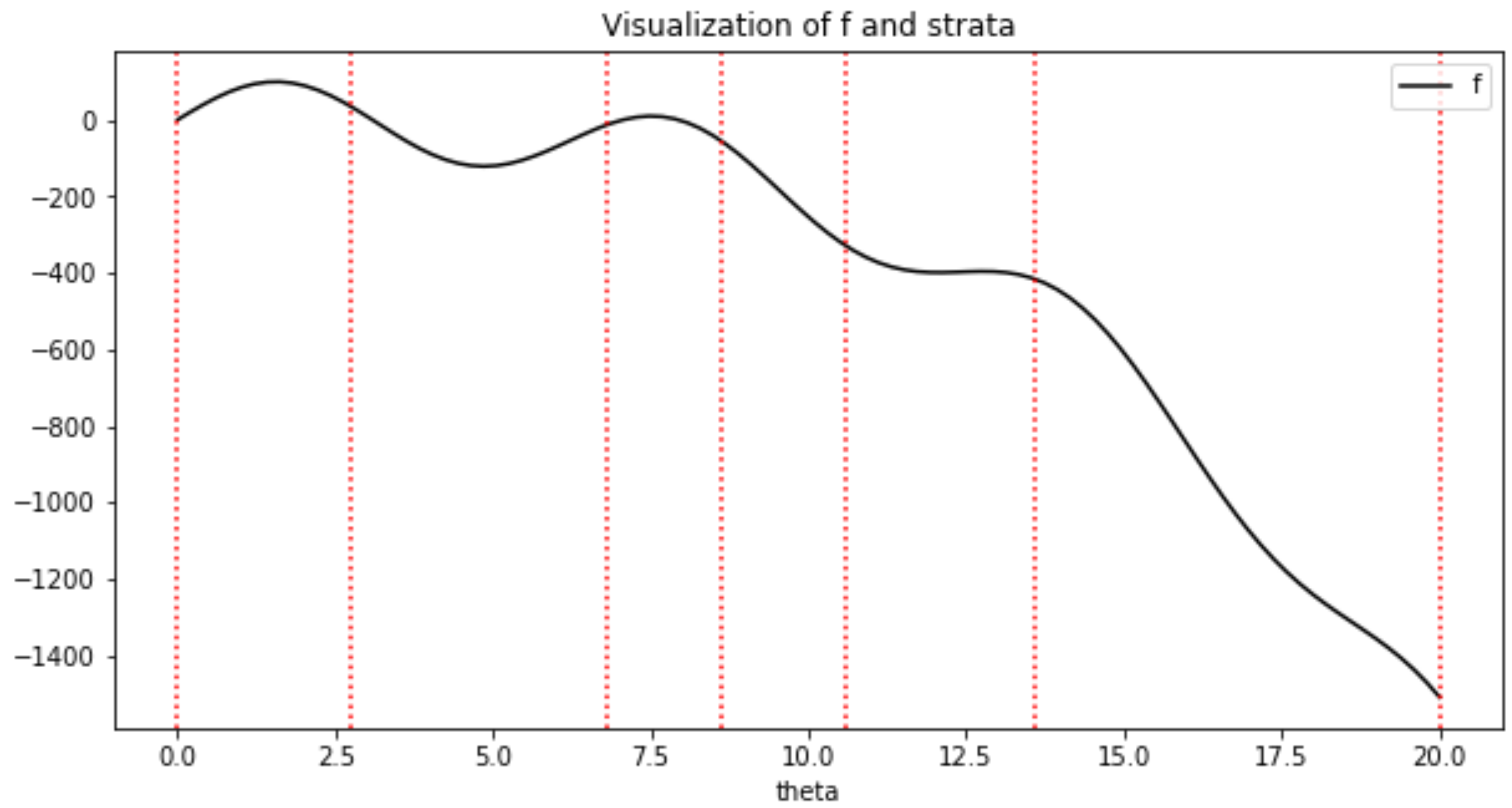
where the number M of strata is 10, $w_m = \frac{t_m - t_{m-1}}{20}$, and where the end points of the strata is given by

$$[0, 2.75, 6.8, 8.6, 10.6, 13.6, 20.].$$

Example: Stratified Sampling

In [8]:

```
#define the end points of the strata
strata = [0, 2.75, 6.8, 8.6, 10.6, 13.6, 20.]
x = np.linspace(0, 20, 100)
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax.plot(x, f(x), color='black', label='f')
for stratus in strata:
    ax.axvline(x=stratus, linestyle=':', color='red')
ax.set_title('Visualization of f and strata')
ax.set_xlabel('theta')
ax.legend(loc='best')
plt.show()
```



Example: Stratified Sampling

In [13]:
```python
print('variance of monte carlo estimate:', mc_variance)
print('variance of stratified monte carlo estimate:', stratified_mc_variance)
```

```
variance of monte carlo estimate: 1993.194946380082
variance of Stratified s. monte carlo estimate: 368.9644656704295
```

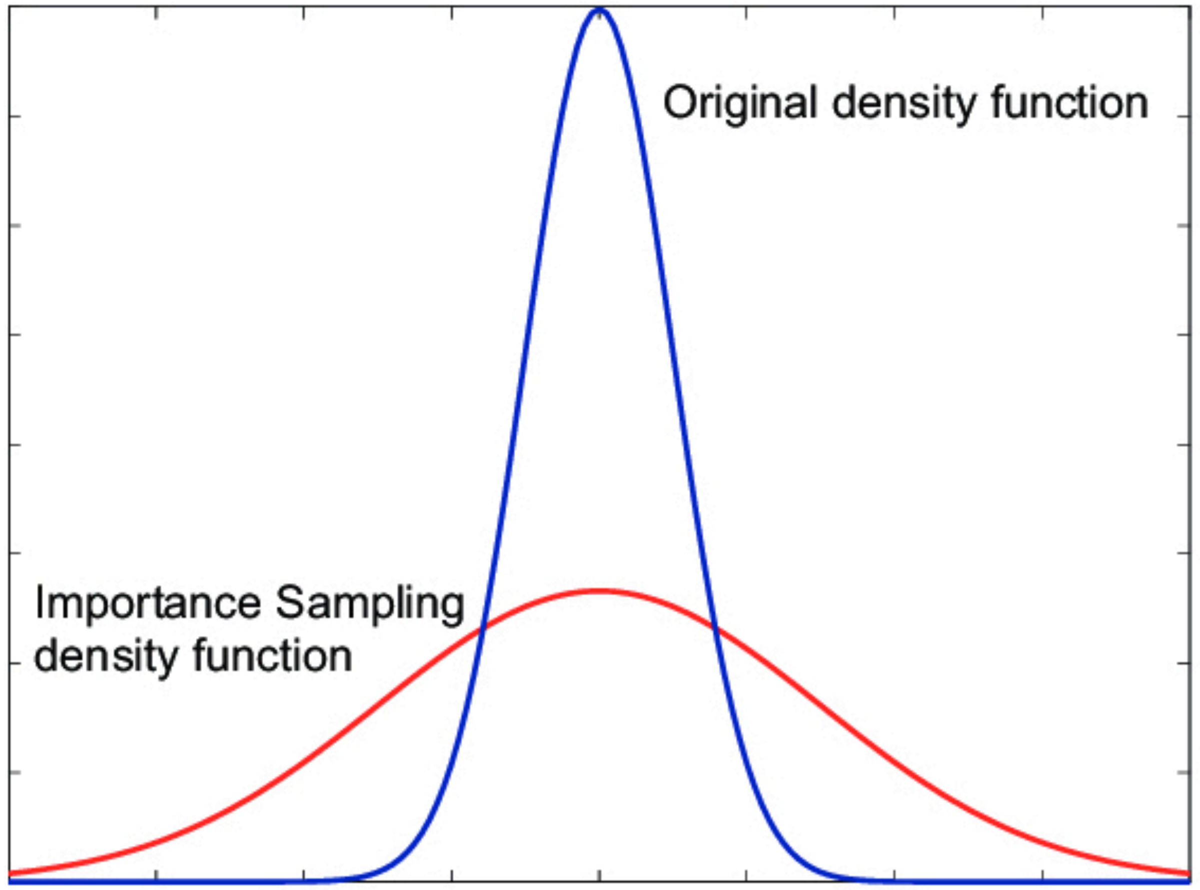
Variance Reduction: Importance Sampling

The General Idea of Importance Sampling

The idea behind importance sampling is very similar to that of rejection sampling: we tackle the problem where the posterior $p(\theta|Y)$ is typically hard to sample from. Instead,

1. we approximate $p(\theta|Y)$ with an easy distribution $q(\theta)$ and sample $\theta_1, \ldots, \theta_S$ from q.

2. rather than rejecting some of the samples, we evaluate f on all the samples, and then weight $f(\theta_s)$ based on how likely is θ_s to be from p vs q. That is, we estimate $\mathbb{E}_{\theta|Y}[f(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S} \frac{p(\theta_s|Y)}{q(\theta_s)} f(\theta_s), \theta_s \sim q(\theta)$. **Why is this a good idea?**



Variance of Importance Sampling

Fix a total sample size S. The ***importance sampling Monte Carlo estimator*** is

$$\hat{I}_{\text{IS}} = \frac{1}{S} \sum_{s=1}^{S} \frac{p(\theta_s|Y)}{q(\theta_s)} f(\theta_s).$$

We call q is the ***importance distribution***, p the ***nominal distribution*** and $\frac{p(\theta_s|Y)}{q(\theta_s)}$ the ***importance weight***.

We can show that $\hat{I}_{\text{IS}}$ is unbiased and consistent using the following fact (assuming q and p have the same support):

$$I = \mathbb{E}_{\theta\sim p(\theta|Y)}\left[f(\theta)\right] = \int_\Theta \frac{p(\theta|Y)}{q(\theta)} f(\theta) q(\theta) d\theta = \mathbb{E}_{\theta\sim q(\theta)}\left[\frac{p(\theta|Y)}{q(\theta)} f(\theta)\right].$$

We can directly calculate the variance of $\hat{I}_{\text{IS}}$ to be $\mathrm{Var}\left[\hat{I}_{\text{IS}}\right] = \frac{\sigma_q^2}{S}$, where σ_q^2 is the variance of $\frac{p(\theta_s|Y)}{q(\theta_s)} f(\theta_s)$:

$$\sigma_q^2 = \int_\Theta f^2(\theta) p(\theta|Y) \frac{p(\theta|Y)}{q(\theta)} d\theta - I^2 = \int_\Theta \frac{(f(\theta)p(\theta|Y) - Iq(\theta))^2}{q(\theta)} d\theta.$$

For comparison, $\mathrm{Var}\left[\hat{I}\right] = \int_\Theta f(\theta)^2 p(\theta|Y) d\theta - I^2$. So, we hope to get a variance reduction if $\frac{p(\theta|Y)}{q(\theta)} < 1$ when $f(\theta)^2 p(\theta|Y)$ is large.

The Nitty Gritty of Importance Sampling: the Design of q

1. **(When $q = 0$)** if $q(\theta_0) = 0$ but $p(\theta_0|Y) \neq 0$ for some θ_0 then $\widehat{I}_{\text{IS}}$ will be biased, since sampling from q can never simulate sampling from p even after reweighting.

2. **(When $q \approx 0$)** related, if $q(\theta_0)$ is very small (esp. when $q(\theta_0) < p(\theta_0|Y)$) then it is unlikely that your samples will include θ_0. Hence, $\widehat{I}_{\text{IS}}$ will be biased. **If you've never see snow, you can't prepare for it!**

3. **(Again, when $q \approx 0$)** recall that $\sigma_q^2 = \int_\Theta \frac{(f(\theta)p(\theta|Y) - Iq(\theta))^2}{q(\theta)}$. For smaller σ_q^2, we want $(f(\theta)p(\theta|Y) - Iq(\theta))^2 \approx 0$, i.e. $q \propto f(\theta)p(\theta|Y)$. But where $q(\theta) \approx 0, \sigma_q^2$ can be large even when $(f(\theta)p(\theta|Y) - Iq(\theta))^2$ is small.

4. **(Heterogenous weights)** when q and p are very dissimlar, the weights will either be near 0 or huge. Say we have one sample with weight 0 and one with weight 999, how many samples are we effectively simulating from p? A way to measure the "quality" of the samples from q in estimating I, is through the ***effective sample size***.

   When we have 100 samples from q but ESS is 10, then essentially your IS estimate of I is as good as as Monte Carlo estimate with 10 samples from p.

Example: Importance Sampling

Again, let $f(\theta) = -0.2\theta^3 + 100\sin(\theta)$ for $\theta|Y \sim U(0, 20)$. We aim to estimate:

$$\mathbb{E}_{\theta|Y}[f(\theta)] = \int_\Theta f(\theta)p(\theta|Y)d\theta = \frac{1}{20}\int_\Theta f(\theta)d\theta.$$

We compare the sample variance of $\widehat{I} = \frac{1}{S}\sum_{s=1}^{S} f(\theta_s)$ with that of the Monte Carlo estimate with importance sampling, with the importance distribution being $\mathcal{N}(12, 3)$:

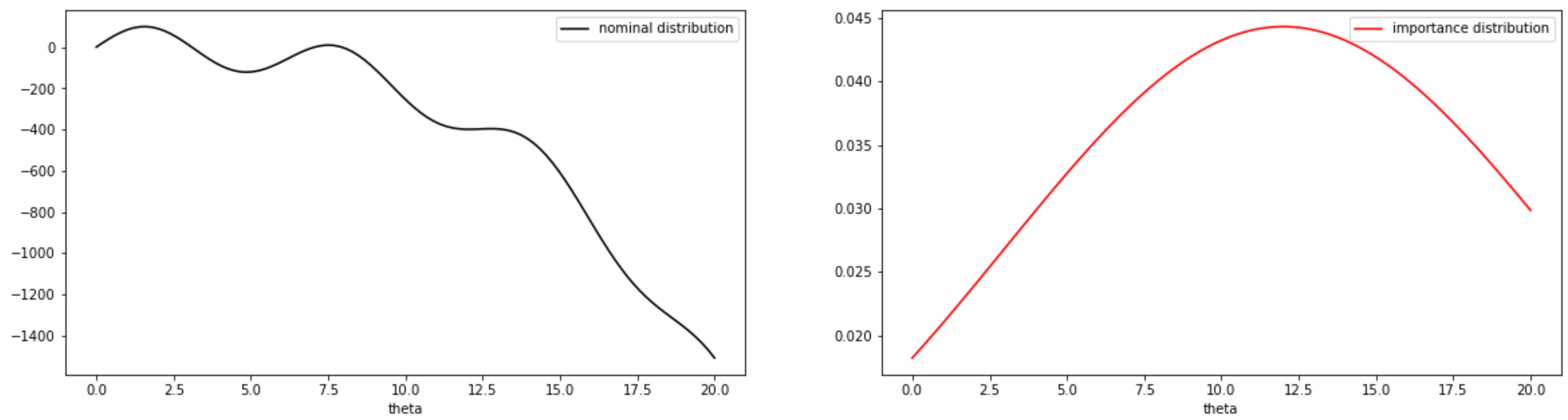
$$\widehat{I}_{\text{IS}} = \sum_{s=1}^{S} \frac{1/20 * \mathbb{1}_{[0,20]}}{\mathcal{N}(\theta_s; 12, 3)} f(\theta_s), \;\; \theta_s \sim \mathcal{N}(12, 3),$$

for sample size $S = 100$.

Example: Importance Sampling

In [8]:

```
#normal pdf
normal = sp.stats.norm(loc=12, scale=3**2).pdf
#visualize nominal and importance distributions
x = np.linspace(0, 20, 100)
fig, ax = plt.subplots(1, 2, figsize=(20, 5))
ax[0].plot(x, f(x), color='black', label='nominal distribution')
ax[0].set_xlabel('theta')
ax[0].legend(loc='best')
ax[1].plot(x, normal(x), color='red', label='importance distribution')
ax[1].set_xlabel('theta')
ax[1].legend(loc='best')
plt.show()
```



Example: Importance Sampling

In [50]:

```
print('variance of monte carlo estimate:', mc_variance)
print('variance of IS monte carlo estimate:', is_mc_variance)
```

```
variance of monte carlo estimate: 2212.2325702371
variance of IS monte carlo estimate: 1364.0624019299828
```

In [51]:

```
print('average numnber of zero importance weight:', np.mean(importance_weights))
```

```
average numnber of zero importance weight: 0.37
```

Summary of Monte Carlo Integration

Monte Carlo Integration

Last week we studied ways of ***sampling*** from an arbitrary distribution $p(\theta)$:

- **Inverse CDF Sampling:** when the inverse of the CDF of p can be computed
- **Rejection Sampling:** when a good substitute distribution for p can be found
- **Gibbs Sampling:** when the conditionals of a multivariate distribution are simple

Samples from $p(\theta)$ can be used to estimate the expected value of any function of θ:

$$\underbrace{\mathbb{E}_{\theta\sim p(\theta)}\left[f(\theta)\right]}_{I} = \int_\Theta f(\theta)p(\theta)d\theta \approx \underbrace{\frac{1}{S}\sum_{s=1}^{S} f(\theta_s),\;\; \theta_s \sim p(\theta)}_{\widehat{I}}$$

We call $\widehat{I}$ the ***Monte Carlo estimate*** of I. We proved that:

- $\widehat{I}$ is consistent and unbiased
- The variance of $\widehat{I}$ is $\frac{\sigma_f}{S}$, where $\sigma_f = \mathrm{Var}\left[f(\theta)\right]$

What do these two properties of $\widehat{I}$ tell us to do if we want better estimates of I?

Processing math: 100%

Variance Reduction for Monte Carlo Integration

We can reduce the variance of $\widehat{I}$ by increasing the sample size S. But this is not always practical. **Why?**

We studied three ways of reducing the variance of $\widehat{I}$ by reducing the variance of the function inside the expectation:

1. **Control Variates:** find a function h that captures the trend of f:
$$\widehat{I}_{\text{control}} = \frac{1}{S}\sum_{s=1}^{S} f(\theta) - c(h(\theta) - \mu_h)$$
2. **Stratified Sampling:** Split the domain Θ into m pieces (strata):
$$\widehat{I}_{\text{SS}} = \sum_{m=1}^{M} w_m \widehat{I}_m.$$
where $w_m = \int_{\Theta_m} p(\theta)d\theta$, and $\widehat{I}_m$ is the MC estimate over the m-th strata
3. **Importance Sampling:** choose a q that places more weight where $f(\theta)$ has high variance and p is more likely to produce samples:
$$\widehat{I}_{\text{IS}} = \frac{1}{S}\sum_{s=1}^{S} \frac{p(\theta_s|Y)}{q(\theta_s)} f(\theta_s),\ \ \theta_s \sim q(\theta).$$

For each method: What are the design choices you need to make? Where is the variance reduction coming from? Why is it hard to achieve?

Application: Monte Carlo Estimation of Arbitrary Integrals

Processing math: 100%

Arbitrary Definite Integrals

We formulated Monte Carlo estimation as a way to approximate the expected value of functions of random variables. However, given a function of deterministic variables, $f(x)$, we can rewrite the integral

$$\int_{\Omega} f(x)dx,$$

where $x \in \mathbb{R}^D$ and $\Omega \subseteq \mathbb{R}^D$, like an expectation

$$\int_{\Omega} f(x)dx = \int_{\Omega} g(x)h(x)dx,$$

if we can factor f as $f(x) = g(x)h(x)$.

The rewritten integral looks a lot like the expectation of $g(X)$ where the pdf of X is $h(X)$. The only problem is that $h(x)$ is not necessarily a pdf!

Rewriting an Arbitrary Definite Integral as an Expectation

We want to rewrite the following expression as an expectation:

$$\int_{\Omega} f(x)dx = \int_{\Omega} g(x)h(x)dx.$$

Let V denote the integral of h over Ω:

$$V = \int_{\Omega} h(x)dx.$$

Note that, then, $\int_{\Omega} \frac{1}{V} h(x)dx = 1$. Hence, $\frac{1}{V} h(x)$ is a valid pdf over Ω.

Then we can write:

$$\int_{\Omega} f(x)dx = \int_{\Omega} (Vg(x)) \left(\frac{1}{V} h(x)\right) dx.$$

We recognize the above integral as an expectation, and hence can estimated using Monte Carlo methods:

$$\mathbb{E}_{X \sim \frac{1}{V} h(X)}[Vg(X)] \approx \frac{V}{S} \sum_{s=1}^{S} g(X_s), \;\; X_s \sim \frac{1}{V} h(X).$$

How Do You Factor the Integrand f?

Once we factor $f(x)$ as $f(x) = g(x)h(x)$, then we can rewrite the integral of f (over Ω) as an expectation

$$\int_{\Omega} f(x)dx = \int_{\Omega} (Vg(x)) \left(\frac{1}{V} h(x) \right) dx = \mathbb{E}_{X \sim \frac{1}{V} h(X)} [Vg(X)].$$

But how does one choose g and h? Two simple things to keep in mind:

1. since we need to sample from h, we need to choose h such that it is easily approximated by distributions from which we know how to sample.
2. since the variance of the Monte Carlo estimator depends on g, we should choose g to be relatively flat (so that the output of g has low variance).