the Kolmogorov complexity will be less than $K(x_1) + K(x_2)$, since some information from one of the strings can be used in the generatation of the other string.

16. PROBLEM NOT YET SOLVED

17. The definition "the least number that cannot be defined in less than twenty words" is already a definition of less than twenty words. The definition of the Kolmogorov complexity is the length of the shortest program to describe a string. From the above paradoxical statement, we can see that it is possible that we are not "clever" enough to determine the shortest program length for a string, and thus we will not be able to determine easily the complexity.

Section 9.3

18. The mean-square error is

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})^2] &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D}) - 2g(\mathbf{x};\ \mathcal{D})F(\mathbf{x}) + F^2(\mathbf{x})] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D}] - \mathcal{E}_{\mathcal{D}}[2g(\mathbf{x};\ \mathcal{D}F(\mathbf{x})] + \mathcal{E}_{\mathcal{D}}[F^2(\mathbf{x})] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - 2F(\mathbf{x})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + F^2(\mathbf{x}).\end{aligned}$$

Note, however, that

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D}) - 2g(\mathbf{x};\ \mathcal{D})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + [\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]]^2] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - \mathcal{E}_{\mathcal{D}}[2g(\mathbf{x};\ \mathcal{D})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]] + \mathcal{E}_{\mathcal{D}}[(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - 2\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2 \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2.\end{aligned}$$

We put these two results together and find

$$\mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] = \mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2.$$

We now apply this result to the function $g(\mathbf{x}) - F(\mathbf{x})$ and obtain

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x}))^2] &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - 2F(\mathbf{x})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + F^2(\mathbf{x}) \\ &= \mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D} - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]))^2] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2 \\ &\quad -2F(\mathbf{x})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + F^2(\mathbf{x}) \\ &= \mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})])^2 \\ &= \underbrace{(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})])^2}_{bias^2} + \underbrace{\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2]}_{variance}.\end{aligned}$$

Since the estimate can be more or less than the function $F(\mathbf{x})$, the bias can be negative. The variance cannot be negative, as it is the expected value of a squared number.

19. For a given data set $\mathcal{D}$, if $g(\mathbf{x};\ \mathcal{D})$ agrees with the Bayes classifier, the expected error rate will be $\text{Min}[F(\mathbf{x}, 1 - F(\mathbf{x})]$; otherwise it will be $\text{Max}[F(\mathbf{x}, 1 - F(\mathbf{x})]$. Thus we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y] &= \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})]\Pr[g(\mathbf{x};\ \mathcal{D}) = y_B] \\ &\quad +\text{Max}[F(\mathbf{x}), 1 - F(\mathbf{x})]\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B].\end{aligned}$$

However, under these conditions we can write

$$\text{Max}[F(\mathbf{x}), 1 - F(\mathbf{x})] = \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})] + \underbrace{\text{Max}[F(\mathbf{x}), 1 - F(\mathbf{x})] - \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})]}_{|2F(\mathbf{x})-1}.$$

Thus we conclude

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y &= |2F(\mathbf{x}) - 1|\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] \\ &\quad + \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})](\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] + \Pr[g(\mathbf{x};\ \mathcal{D}) = y_B]) \\ &= |2F(\mathbf{x}) - 1|\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] + \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})].\end{aligned}$$

20. If we make the convenient assumption that $p(g(\mathbf{x};\ \mathcal{D}))$ is a Gaussian, that is,

$$p(g(\mathbf{x};\ \mathcal{D})) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(g-\mu)^2/(2\sigma^2)]$$

where $\mu = \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]$ and $\sigma^2 = \text{Var}[g(\mathbf{x};\ \mathcal{D})]$. From Eq. 19 in the text, then, for $F(\mathbf{x}) < 1/2$ we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] &= \int_{1/2}^{\infty} p(g(\mathbf{x};\ \mathcal{D}))dg \\ &= \int_{1/2}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp[-(g-\mu)^2/(2\sigma^2)]dg \\ &= \frac{1}{\sqrt{2\pi}} \int_{(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du,\end{aligned}$$

where $u = (g-\mu)/\sigma$ and $du = dg/\sigma$. For the other case, that is, $F(\mathbf{x}) \geq 1/2$, we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] &= \int_{-\infty}^{1/2} p(g(\mathbf{x};\ \mathcal{D}))dg \\ &= \int_{-\infty}^{1/2} \frac{1}{\sqrt{2\pi}\sigma}\exp[-(g-\mu)^2/(2\sigma^2)]dg \\ &= \frac{1}{\sqrt{2\pi}} \int_{-(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du,\end{aligned}$$

where $u = -(g-\mu)/\sigma$ and $du = -dg/\sigma$. Therefore, we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D})] &= \begin{cases} \frac{1}{\sqrt{2\pi}} \int_{(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du & \text{if } F(\mathbf{x}) < 1/2 \\ \frac{1}{\sqrt{2\pi}} \int_{-(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du & \text{if } F(\mathbf{x}) \geq 1/2 \end{cases} \\ &= \frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} \exp[-u^2/2]du = \frac{1}{2}\left[1 - \text{erf}[t/\sqrt{2}]\right] = \Phi(t),\end{aligned}$$

where

$$t = \begin{cases} \frac{1/2-\mu}{\sigma} & \text{if } F(\mathbf{x}) < 1/2 \\ -\frac{1/2-\mu}{\sigma} & \text{if } F(\mathbf{x}) \geq 1/2. \end{cases}$$

Thus, we can write

$$\begin{aligned}
\Pr[g(\mathbf{x};\,\mathcal{D})] &= \text{sgn}[F(\mathbf{x})-1/2]\frac{\mu-1/2}{\sigma} \\
&= \text{sgn}[F(\mathbf{x})-1/2]\frac{\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\,\mathcal{D})-1/2]}{\sqrt{\text{Var}[g(\mathbf{x};\,\mathcal{D})]}} \\
&= \underbrace{\text{sgn}[F(\mathbf{x})-1/2][\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\,\mathcal{D})]-1/2]}_{\textit{boundary bias}}\underbrace{\text{Var}[g(\mathbf{x};\,\mathcal{D})]^{-1/2}}_{\textit{variance}}.
\end{aligned}$$

21. PROBLEM NOT YET SOLVED
22. PROBLEM NOT YET SOLVED

Section 9.4

23. The jackknife estimate of the mean is given by Eq. 25 in the text:

$$\begin{aligned}
\mu_{(\cdot)} &= \frac{1}{n}\sum_{i=1}^{n}\mu_{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j\neq i}x_j\right] \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}x_j - x_i\right] \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}[n\hat{\mu}-x_i] \\
&= \frac{n}{n-1}\hat{\mu}-\frac{1}{n(n-1)}\sum_{i=1}^{n}x_i \\
&= \frac{n}{n-1}\hat{\mu}-\frac{1}{n-1}\hat{\mu} \\
&= \hat{\mu}.
\end{aligned}$$

24. PROBLEM NOT YET SOLVED
25. PROBLEM NOT YET SOLVED
26. We must verify that Eq. 26 in the text for the jackknife estimate of the variance of the mean is formally equivalent to the variance estimate given by Eq. 23 in the text. From Eq. 26 we have

$$\begin{aligned}
\text{Var}[\hat{\mu}] &= \frac{n-1}{n}\sum_{i=1}^{n}(\mu_{(i)}-\mu_{(\cdot)})^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\left(\frac{n\bar{x}-x_i}{n-1}\right)-\mu_{(\cdot)}\right)^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n\bar{x}-x_i}{n-1}-\frac{n-1}{n-1}\bar{x}\right)^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n\bar{x}-x_i-(n-1)\bar{x}}{n-1}\right)^2
\end{aligned}$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{\bar{x}-x_i}{n-1}\right)^2$$
$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

which is Eq. 23 in the text.

27. Consider the computational complexity of different statistics based on resampling.

(a) The jackknife estimate of the mean is

$$\theta_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)},$$

which requires n summations, and thus has a complexity $O(n)$.

(b) The jackknife estimate of the median has complexity just that required for the sorting operation, which is $O(n\log n)$.

(c) The jackknife estimate of the standard deviation is

$$\sqrt{\mathrm{Var}[\hat{\theta}]} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\theta_{(i)}-\theta_{(\cdot)}\right)^2}$$
$$= \sqrt{\frac{n-1}{n}\left[\sum_{i=1}^{n}\theta_{(i)}^2-\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}\right)^2\right]},$$

which requires $2n$ summations, and thus has a complexity $O(n)$.

(d) The bootstrap estimate of the mean is

$$\hat{\theta}^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)},$$

which requires B summations, and thus has a complexity $O(B)$.

(e) The bootstrap estimate of the medium has complexity the same as that as the sorting operation, which is $O(B\log B)$.

(f) The bootstrap estimate of the standard deviation is

$$\sqrt{\mathrm{Var}_{Boot}[\theta]} = \sqrt{\frac{1}{B}\sum_{i=1}^{n}\left(\hat{\theta}^{*(b)}-\hat{\theta}^{*(\cdot)}\right)^2}$$
$$= \sqrt{\frac{1}{B}\left[\sum_{b=1}^{B}(\hat{\theta}^{*(b)})^2-\left(\frac{1}{B}\sum_{b'=1}^{B}\hat{\theta}^{*(b')}\right)^2\right]},$$

which requires $2B$ summations, and thus has a complexity $O(B)$.