A Crowdsourcing data collection

In this section, we provide details regarding our the design of our annotation interfaces and the quality control measures we took.

A.1 Language quality evaluation.

Each human annotator was shown a short summary that was generated by a system from an article in the CNN/Daily Mail dataset or provided as a reference for that article. The annotators were then asked to (a) provide Likert scale ratings of the summary on multiple facets (fluency, redundancy and overall quality) and (b) perform post-edits to correct any errors (Figure 7a).

Interface design choices. We found that using a five-level Likert scale increased annotator variance as annotators relative to a three-level Likert scale. Annotators were provided specific cues to calibrate their Likert ratings through a tutorial and were reminded of these cues through tooltips on the rating buttons (see Figure 7b for an example). If the annotators rated a summary as lacking along any facet, they were then forced to perform post-edits to "improve [its] quality as much as possible". We found that forcing annotators to provide post-edits on examples significantly decreased the annotator variance even on the Likert ratings.

Following the recommendations of Liu et al. (2016a), we forced annotators to complete an interactive tutorial containing 10 questions each before beginning the task (Figure 7b). The tutorial provided guidelines and examples on how to rate each facet (fluency, redundancy and overall quality) and tested whether they were able to identify and correct language errors using the post-editing interface. The tutorial took about 5–6 minutes to complete and annotators were paid a one-time bonus of $0.75 on completion.

We initially included additional questions to assess focus, coherency and referential clarity adapted from the DUC evaluation guidelines (Dang, 2006), but found that annotators were unable to reliably identify these errors in the short summaries. We also experimented with asking annotators to highlight language errors in the text to justify their ratings, but again found that annotators were unable to localize these errors reliably.

Quality control measures. We initially attempted to use attention-check examples for the Likert rating questions, but found that the ratings on these examples were themselves quite subjective and hence were not a reliable signal to reject work. Instead, we found that requiring post-edits to summaries significantly reduced spam. Additionally, we rejected annotators who took too little time to complete the task, had very low agreement rates on the Likert questions or had edits that were consistently shorter than 5 characters to prevent spam.

A.2 Answer correctness evaluation.

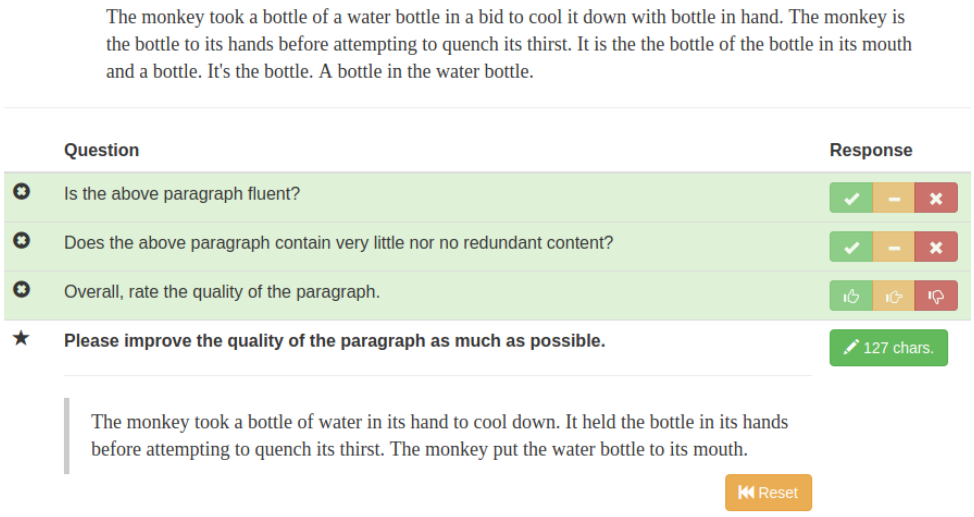
Each annotator was shown a question from the MS MARCO dataset and an answer that was generated by a system or provided as a reference answer from the dataset. The annotators were then asked to (a) rate if the question made sense and the answer was plausibly correct and (b) asked to identify which paragraphs provided in the dataset justified the answer (Figure 8a).

Interface design choices. We found that some of the questions in the MS MARCO dataset were extremely ambiguous (e.g. "metatarsal what causes") and some system responses were implausible (e.g "monogenic bone diseases", for the question "what genes cause osteoporosis"). In these cases, annotators expressed confusion if they were forced to judge if the response was correct or incorrect. We resolved this confusion by first asking annotators if the question made sense and if system response was even plausible.

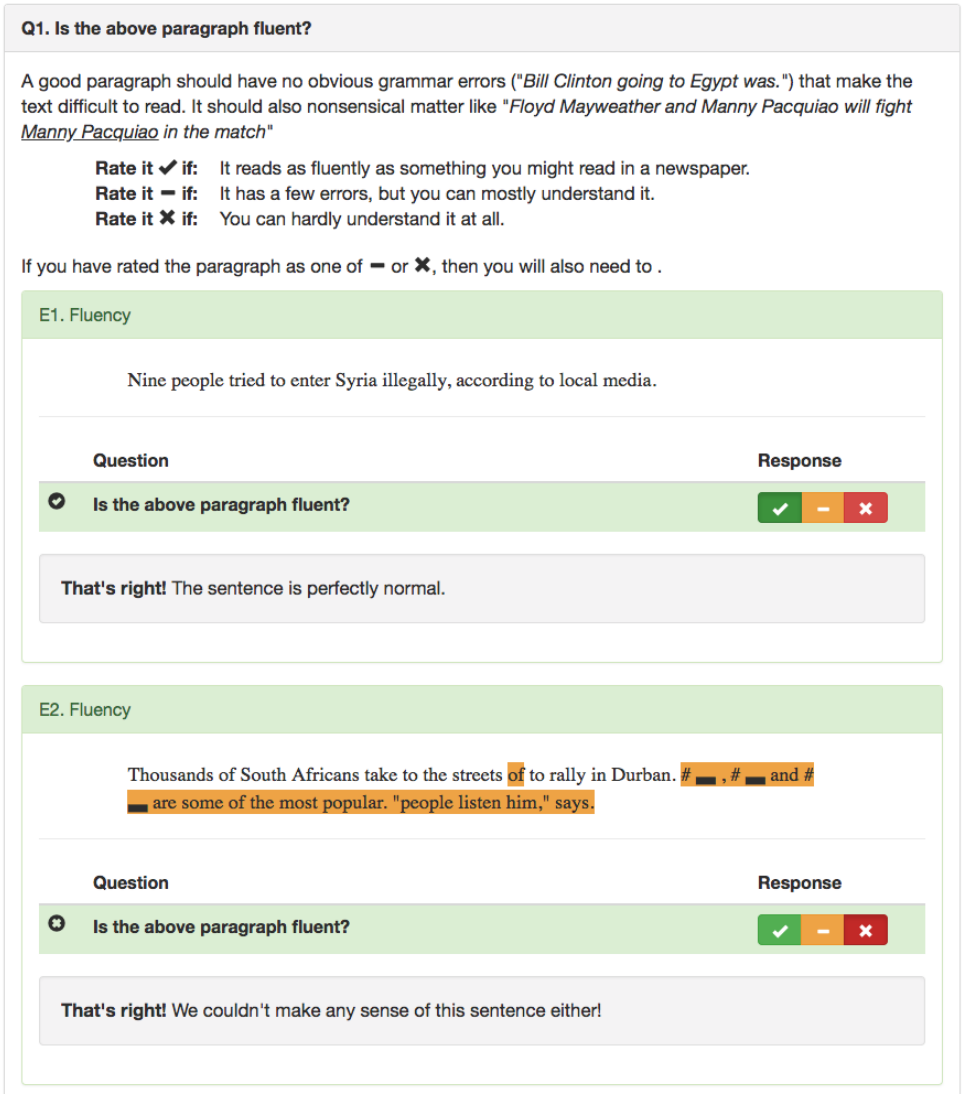
In early pilots, we found that annotators often rated a paragraph that correctly answered the question but was unrelated to the system response to be "correct". We were able to resolve this problem by asking annotators to double-check their work (see the last question in Figure 8a for an example).

Once again, we forced annotators to complete an interactive tutorial containing eight questions each before beginning the task (Figure 8b). The tutorial also took about 5–6 minutes to complete and annotators were paid a one-time bonus of $0.75 on completion.

Quality control measures. We found that requiring annotators to provide justification spans significantly spam. Additionally, we rejected annotators who took too little time to complete the task or had very low agreement rates on the answer correctness.

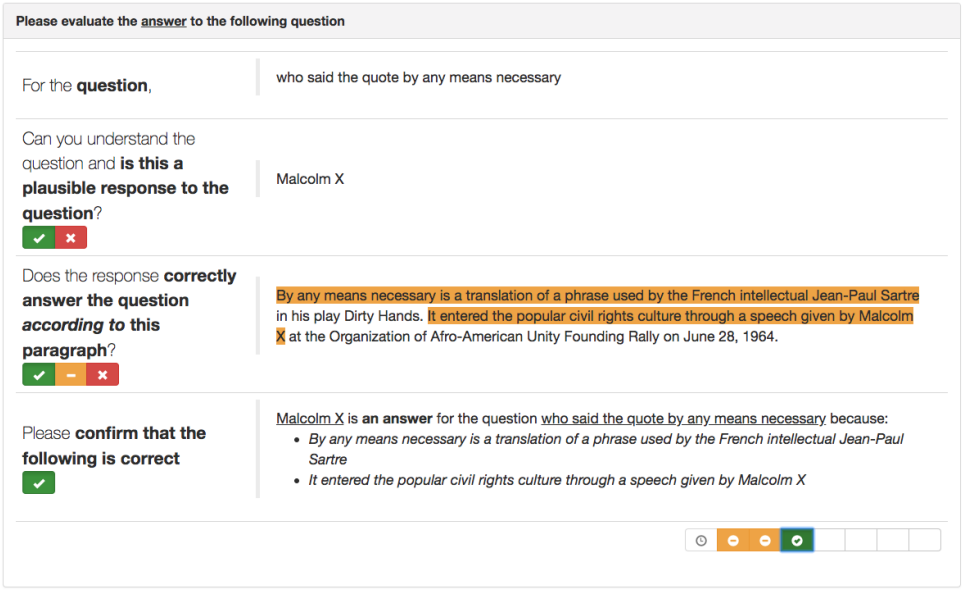


(a)



(b)

Figure 7: Screenshot of the (a) interface and (b) instructions used by crowdworkers for the language quality evaluation task on the CNN/Daily Mail dataset.



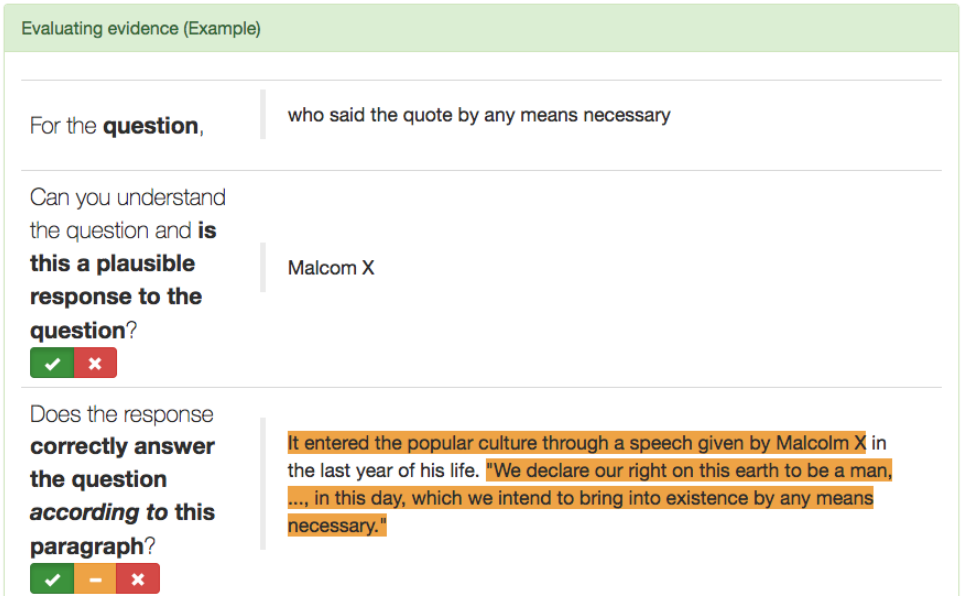
(a)

Evaluating evidence for the response IMPORTANT: PLEASE READ!

If the response is a plausible answer, we would like you to check whether or not it is a *correct answer* according to a few excerpted paragraphs.

1. For each paragraph presented, first **read the paragraph** and indicate if the paragraph provides evidence that the response is correct (✔), incorrect (✖), or that the paragraph simply isn't sufficient to tell us either which way (━). **You only need to use commonsense knowledge and information contained within the question, answer or paragraph. You do not need to search online for further inormation.**
2. If the paragraph provides evidence that the response is either correct (✔) or incorrect (✖), **highlight the regions of the text that you think justifies your decision**.*You can but do not have to highlight regions if the response is neutral (━).* The highlighted regions don't need to be exact, but should help us understand why you are making your decision.
3. **To remove a highlight, simply click on it.**
4. If you judge the response to be correct (or incorrect), you will have to **confirm that the response is an answer (or not an answer) for the question according to your selected evidence**
5. **Use the buttons on the lower right to move through the paragraphs.** You will need to make a decision on each paragraph to complete the task.

Review the different paragraphs below by clicking on the icons in the lower right corner.



(b)

Figure 8: Screenshot of the (a) interface and (b) instructions used by crowdworkers for the answer correctness evaluation task on the MS MARCO dataset.

B Proofs

In this section, we provide proofs for the theorems stated in the main paper.

B.1 Main Theorem

In this section, we prove the main theorem (Theorem 3.1) in the paper about the minimax optimal variance for an unbiased estimator. Theorem 3.1 will follow from the two following lemmas (Lemmas B.1 and B.2). First, we show in Lemma B.1 that for all distributions with fixed σ_f^2, σ_a^2 and ρ, the variance of $\hat{\mu}_{\text{cv}}$ is constant and equal to: $\frac{1}{n}(\sigma_f^2(1-\rho^2)+\sigma_a^2)$. Then we give an explicit distribution, a Gaussian distribution, where *any* estimator yields at least this variance using the theory of sufficient statistics. Together, these show that the max variance of any estimator is at least the max variance of $\hat{\mu}_{\text{cv}}$.

As a reminder, the estimator is

$$\hat{\mu}_{\text{cv}} = \frac{1}{n}\sum_i y^{(i)} - \alpha g(z^{(i)}) \tag{8}$$

where $\alpha = \text{Cov}(f(z), g(z))$.

Lemma B.1. *The variance of $\hat{\mu}_{cv}$ is always*

$$\frac{1}{n}(\sigma_f^2(1-\rho^2)+\sigma_a^2) \tag{9}$$

Proof. By the law of total variance, with respect to the draws of $z^{(i)}$,

$$\text{Var}(\hat{\mu}_{\text{cv}}) = \mathbb{E}_{z^{(i)}}[\text{Var}(\hat{\mu}_{\text{cv}}|z^{(i)})] + \text{Var}_{z^{(i)}}(\mathbb{E}[\hat{\mu}_{\text{cv}}|z^{(i)}]) \tag{10}$$

We will evaluate each of the two terms on the right hand side.

For the first term,

$$\mathbb{E}_{z^{(i)}}[\text{Var}(\hat{\mu}_{\text{cv}}|z^{(i)})] = \mathbb{E}_{z^{(i)}}\left[\text{Var}\left(\frac{1}{n}\sum_i y^{(i)}|z^{(i)}\right)\right] \tag{11}$$

Because the human responses $Y(z^{(i)})$ are uncorrelated,

$$\begin{aligned}\mathbb{E}_{z^{(i)}}[\text{Var}(\hat{\mu}_{\text{cv}}|z^{(i)})] &= \mathbb{E}_{z^{(i)}}\left[\frac{1}{n^2}\sum_i \text{Var}(Y(z^{(i)}))|z^{(i)}\right] &(12)\\ &= \frac{1}{n}\mathbb{E}_z[\text{Var}(Y(z))] &(13)\\ &= \frac{1}{n}\sigma_a^2 &(14)\end{aligned}$$

For the second term,

$$\text{Var}_{z^{(i)}}(\mathbb{E}[\hat{\mu}_{\text{cv}}|z^{(i)}]) = \text{Var}_{z^{(i)}}\left(\frac{1}{n}\sum_i f(z^{(i)}) - \alpha g(z^{(i)})\right) \tag{15}$$

Because the $z^{(i)}$ are sampled independently,

$$\mathrm{Var}_{z^{(i)}}(\mathbb{E}[\hat{\mu}_{\text{cv}}|z^{(i)}]) = \frac{1}{n}\mathrm{Var}(f(z) - \alpha g(z)) \tag{16}$$

$$= \frac{1}{n}[\mathrm{Var}(f(z)) - 2\alpha\,\mathrm{Cov}(f(z), g(z)) + \alpha^2\,\mathrm{Var}(g(z))] \tag{17}$$

Note that $\mathrm{Var}(f(z)) = \sigma_f^2$, $\mathrm{Cov}(f(z), g(z)) = \alpha$, and $\mathrm{Var}(g(z)) = 1$ (since it is normalized). Thus,

$$\mathrm{Var}_{z^{(i)}}(\mathbb{E}[\hat{\mu}_{\text{cv}}|z^{(i)}]) = \frac{1}{n}[\sigma_f^2 - 2\alpha^2 + \alpha^2] \tag{18}$$

$$= \frac{1}{n}[\sigma_f^2 - \alpha^2] \tag{19}$$

Since the correlation $\rho = \frac{\alpha}{\sigma_f \sigma_g} = \frac{\alpha}{\sigma_f}$,

$$\mathrm{Var}_{z^{(i)}}(\mathbb{E}[\hat{\mu}_{\text{cv}}|z^{(i)}]) = \frac{1}{n}[\sigma_f^2 - \sigma_f^2\rho^2] \tag{20}$$

$$= \frac{1}{n}\sigma_f^2(1 - \rho^2) \tag{21}$$

Putting these two terms together, we find that,

$$\mathrm{Var}(\hat{\mu}_{\text{cv}}) = \frac{1}{n}\sigma_a^2 + \frac{1}{n}\sigma_f^2(1 - \rho^2) \tag{22}$$

$$= \frac{1}{n}(\sigma_f^2(1 - \rho^2) + \sigma_a^2) \tag{23}$$

□

For the next lemma, we show that the worst-case variance for any estimator is at least that of $\hat{\mu}_{\text{cv}}$. For this, we will define a simple Gaussian distribution and use the theory of sufficient statistics. We explicitly define a distribution over $f(z)$, $g(z)$, and $Y(Z) - f(z)$. In particular, we assume these are all Gaussian distributions with respective means, $\mu, 0, 0$, and variances, $\sigma_f^2, 1, \sigma_a^2$. Additionally, we assume that $f(z)$ and $g(z)$ have covariance α but $Y(z) - f(z)$ is independent.

Lemma B.2. *$\hat{\mu}_{cv}$ is the minimal variance unbiased estimate (MVUE) for the Gaussian distribution above.*

Proof. The proof is straightforward: we first show that $\hat{\mu}_{\text{cv}}$ is a sufficient statistic using the Fisher-Neyman factorization theorem, and then we apply the Lehman-Scheffe theorem.

For ease of notation, define $g_i = g(z^{(i)})$ and $y_i = y^{(i)}$. For the purposes of statistics, only μ is a parameter; the other "parameters" are known constants. Note that the pdf of the observed variables g_i and y_i is,

$$\prod_i c_1 \exp(-\frac{1}{2}\begin{bmatrix}(y_i - \mu) \\ g_i\end{bmatrix}^T \begin{bmatrix}\sigma_f^2 + \sigma_a^2 & \alpha \\ \alpha & 1\end{bmatrix}^{-1} \begin{bmatrix}(y_i - \mu) \\ g_i\end{bmatrix}) \tag{24}$$

$$= c_2 \exp(-\frac{1}{2}\sum_i \begin{bmatrix}(y_i - \mu) \\ g_i\end{bmatrix}^T \begin{bmatrix}\sigma_f^2 + \sigma_a^2 & \alpha \\ \alpha & 1\end{bmatrix}^{-1} \begin{bmatrix}(y_i - \mu) \\ g_i\end{bmatrix}) \tag{25}$$

Thus, with the Fisher-Neyman factorization theorem, it suffices to show that the exponetiated term T decomposes as a sum of a function that only depends on the data and a function that only depends on $\hat{\mu}_{\text{cv}}$ and μ.

$$T = \sum_i \begin{bmatrix} (y_i - \mu) \\ g_i \end{bmatrix}^T \begin{bmatrix} \sigma_f^2 + \sigma_a^2 & \alpha \\ \alpha & 1 \end{bmatrix}^{-1} \begin{bmatrix} (y_i - \mu) \\ g_i \end{bmatrix} \tag{26}$$

Letting c_3 be the inverse determinant (which is constant),

$$T = c_3 \sum_i \begin{bmatrix} (y_i - \mu) \\ g_i \end{bmatrix}^T \begin{bmatrix} 1 & -\alpha \\ -\alpha & \sigma_f^2 + \sigma_a^2 \end{bmatrix} \begin{bmatrix} (y_i - \mu) \\ g_i \end{bmatrix} \tag{27}$$

$$= c_3 \left[\sum_i (y_i - \mu)^2 - 2\alpha \sum_i (y_i - \mu) g_i + (\sigma_f^2 + \sigma_a^2) \sum_i g_i^2 \right] \tag{28}$$

$$= c_3 \left[\sum_i y_i^2 - 2\mu \sum_i y_i + n\mu^2 - 2\alpha \sum_i y_i g_i + 2\alpha\mu \sum_i g_i + (\sigma_f^2 + \sigma_a^2) \sum_i g_i^2 \right] \tag{29}$$

$$= -2c_3\mu \left[\sum_i y_i - \alpha \sum_i g_i \right] + c_3 n\mu^2 + c_3 \left[\sum_i y_i^2 - 2\alpha \sum_i y_i g_i + (\sigma_f^2 + \sigma_a^2) \sum_i g_i^2 \right] \tag{30}$$

$$= -2nc_3\mu\hat{\mu}_{\text{cv}} + c_3 n\mu^2 + c_3 \left[\sum_i y_i^2 - 2\alpha \sum_i y_i g_i + (\sigma_f^2 + \sigma_a^2) \sum_i g_i^2 \right] \tag{31}$$

Thus, we see the decomposition into the function of only the data on the right and only μ and $\hat{\mu}_{\text{cv}}$ on the left. Thus, $\hat{\mu}_{\text{cv}}$ is a sufficient statistic.

Further, $\hat{\mu}_{\text{cv}}$ is an unbiased estimate of μ since $\mathbb{E}[g_i] = 0$ and $\mathbb{E}[y_i] = \mu$.

Further, since $\hat{\mu}_{\text{cv}}$ is normally distributed with mean dependent on μ, it is complete.

Thus, by the Lehmann-Scheffe theorem, $\hat{\mu}_{\text{cv}}$ is the minimal variance unbiased estimate (MVUE).

□

Theorem 3.1. *Among all unbiased estimators that are functions of $y^{(i)}$ and $g(z^{(i)})$, and for all distributions with a given σ_f^2, σ_a^2, and α,*

$$\text{Var}(\hat{\mu}_{cv}) = \frac{1}{n}(\sigma_f^2(1 - \rho^2) + \sigma_a^2), \tag{32}$$

and no other estimator has a lower worst-case variance.

Proof. From Lemma B.1 we have that the max variance of $\hat{\mu}_{\text{cv}}$ over all distributions with fixed variances, is exactly,

$$\frac{1}{n}(\sigma_f^2(1 - \rho^2) + \sigma_a^2) \tag{33}$$

Further, from Lemma B.2, we know that $\hat{\mu}_{\text{cv}}$ is the MVUE for a particular class of distributions, thus, any estimator has a larger max variance over all distributions.

Combining these two facts, we get that the minimax variance is the variance of $\hat{\mu}_{\text{cv}}$. □

B.2 Added Bias

Proposition 3.1. *The estimator in Algorithm 1 has $O(1/n)$ bias.*

Proof. The bias B is

$$B = |\mathbb{E}[\widehat{\mu}] - \mu| \tag{34}$$

$$= \left| \mathbb{E}[\frac{1}{n} \sum_i y^{(i)} - \hat{\alpha} g(z^{(i)})] - \mu \right| \tag{35}$$

Since $\mathbb{E}[y^{(i)}] = \mu$,

$$B = \left| \mu - \frac{1}{n} \sum_i \mathbb{E}[\hat{\alpha} g(z^{(i)})] - \mu \right| \tag{36}$$

$$= \left| \frac{1}{n} \sum_i \mathbb{E}[\hat{\alpha} g(z^{(i)})] \right| \tag{37}$$

$$= \left| \frac{1}{n^2} \sum_{i,j} \mathbb{E}[(y^{(j)} - \overline{y}) g(z^{(j)}) g(z^{(i)})] \right| \tag{38}$$

$$= \left| \frac{1}{n^2} \sum_{i,j} \mathbb{E}[y^{(j)} g(z^{(j)}) g(z^{(i)})] - \frac{1}{n^3} \sum_{i,j,k} \mathbb{E}[y^{(k)} g(z^{(j)}) g(z^{(i)})] \right| \tag{39}$$

Because $Y(z)$ is independent and has mean $f(z)$,

$$B = \left| \frac{1}{n^2} \sum_{i,j} \mathbb{E}[f(z^{(j)}) g(z^{(j)}) g(z^{(i)})] - \frac{1}{n^3} \sum_{i,j,k} \mathbb{E}[f(z^{(k)}) g(z^{(j)}) g(z^{(i)})] \right| \tag{40}$$

Because $g(z)$ is mean zero and the $z^{(i)}$ are drawn independently,

$$B = \left| \frac{1}{n^2} \sum_i \mathbb{E}[f(z^{(i)}) g(z^{(i)})^2] - \frac{1}{n^3} \sum_{i,k} \mathbb{E}[f(z^{(k)}) g(z^{(i)})^2] \right| \tag{41}$$

$$= \left| \frac{1}{n^2} \sum_i O(1) - \frac{1}{n^3} \sum_{i,k} O(1) \right| \tag{42}$$

$$= \left| \frac{1}{n^2} O(n) - \frac{1}{n^3} O(n^2) \right| \tag{43}$$

$$= \left| O\left(\frac{1}{n}\right) - O\left(\frac{1}{n}\right) \right| \tag{44}$$

$$= O\left(\frac{1}{n}\right) \tag{45}$$

□