# Appendix C Analysis of SpamBayes for Chapter 5

In this appendix, we analyze the effect of attack messages on SpamBayes. This analysis serves as the motivation for the attacks presented in Section 5.3.

## C.1 SpamBayes' $I(\cdot)$ message score

As mentioned in Section 5.1.1, the SpamBayes $I(\cdot)$ function used to estimate spaminess of a message, is the average between its score $S(\cdot)$ and one minus its score $H(\cdot)$. Both of these scores are expressed in terms of the *chi-squared* cumulative distribution function (CDF): $\chi^2_{2n}(\cdot)$. In both these score functions, the argument to the CDF is an inner product between the logarithm of a scores vector and the indicator vector $\boldsymbol{\delta}(\hat{\mathbf{x}})$ as in Equation (5.3). These terms can be re-arranged to rewrite these functions as $S(\hat{\mathbf{x}}) = 1 - \chi^2_{2n}\left(-2\log s_{\mathbf{q}}(\hat{\mathbf{x}})\right)$ and $H(\hat{\mathbf{x}}) = 1 - \chi^2_{2n}\left(-2\log h_{\mathbf{q}}(\hat{\mathbf{x}})\right)$ where $s_{\mathbf{q}}(\cdot)$ and $h_{\mathbf{q}}(\cdot)$ are scalar functions that map $\hat{\mathbf{x}}$ onto $[0, 1]$ defined as

$$s_{\mathbf{q}}(\hat{\mathbf{x}}) \triangleq \prod_i q_i^{\delta(\hat{\mathbf{x}})_i} \tag{C.1}$$

$$h_{\mathbf{q}}(\hat{\mathbf{x}}) \triangleq \prod_i (1 - q_i)^{\delta(\hat{\mathbf{x}})_i}. \tag{C.2}$$

We further explore these functions in the next section, but first we expound on the properties of $\chi^2_k(\cdot)$.

The $\chi^2_k(\cdot)$ CDF can be written out exactly using gamma functions. For $k \in \mathfrak{N}$ and $x \in \Re_{0+}$ it is simply

$$\chi^2_k(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)}$$

where the *lower-incomplete gamma function* is $\gamma(k, y) = \int_0^y t^{k-1}e^{-t}dt$, the *upper-incomplete gamma function* is $\Gamma(k, y) = \int_y^\infty t^{k-1}e^{-t}dt$, and the *gamma function* is $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$. By these definitions, it follows that for any $k$ and $y$, the gamma functions are related by $\Gamma(k) = \gamma(k, x) + \Gamma(k, x)$. Also note that for $k \in \mathfrak{N}$

$$\Gamma(k, y) = (k-1)!\, e^{-y} \sum_{j=0}^{k-1} \frac{y^j}{j!} \qquad \Gamma(k) = (k-1)!.$$

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

Based on these properties, the $S(\cdot)$ score can be rewritten as

$$S(\hat{\mathbf{x}}) = \frac{\Gamma\left(n, -\log s_{\mathbf{q}}(\hat{\mathbf{x}})\right)}{\Gamma(n)} = s_{\mathbf{q}}(\hat{\mathbf{x}}) \sum_{j=0}^{n-1} \frac{\left(-\log s_{\mathbf{q}}(\hat{\mathbf{x}})\right)^j}{j!}$$

$$H(\hat{\mathbf{x}}) = \frac{\Gamma\left(n, -\log h_{\mathbf{q}}(\hat{\mathbf{x}})\right)}{\Gamma(n)} = h_{\mathbf{q}}(\hat{\mathbf{x}}) \sum_{j=0}^{n-1} \frac{\left(-\log h_{\mathbf{q}}(\hat{\mathbf{x}})\right)^j}{j!}.$$

It is easy shown that both these functions are monotonically non-decreasing in $s_{\mathbf{q}}(\hat{\mathbf{x}})$ and $h_{\mathbf{q}}(\hat{\mathbf{x}})$ respectively. For either of these functions, the following derivative can be taken (with respect to $s_{\mathbf{q}}(\hat{\mathbf{x}})$ or $h_{\mathbf{q}}(\hat{\mathbf{x}})$):

$$\frac{d}{dz}\left[z \sum_{j=0}^{n-1} \frac{(-\log z)^j}{j!}\right] = \frac{1}{(n-1)!}(-\log z)^{n-1},$$

which is non-negative for $0 \leq z \leq 1$.

## C.2 Constructing Optimal Attacks on SpamBayes

As indicated by Equation (5.7) in Section 5.3.1, an attacker with objectives described in Section 5.2.1 would like to have the maximal (deleterious) impact on the performance of SpamBayes. In this section, we analyze SpamBayes' decision function $I(\cdot)$ to optimize the attacks' impact. Here we show that the attacks proposed in Section 5.3.1 are (nearly) optimal strategies for designing a single attack message that maximally increases $I(\cdot)$.

In the attack scenario described in Section 5.3.1.1, the attacker will send a series of attack messages which will increase $N^{(s)}$ and $n_j^{(s)}$ for the tokens that are included in the attacks. We will show how $I(\cdot)$ changes as the token counts $n_j^{(s)}$ are increased to understand which tokens the attacker should choose to maximize the impact per message. This analysis separates into two parts based on the following observation.

*Remark* C.1 Given a fixed number of attack spam messages, $q_j$ is independent of the number of those messages containing the $k^{\text{th}}$ token for all $k \neq j$.

This remark follows from the fact that the inclusion of the $j^{\text{th}}$ token in attack spams affects $n_j^{(s)}$ and $n_j$ but not $n_k^{(h)}$, $N^{(s)}$, $N^{(h)}$, $n_k^{(s)}$, $n_k^{(h)}$, or $n_k$ for all $k \neq j$ (see Equations (5.1) and (5.2) in Section 5.1.1).

After an attack consisting of a fixed number of attack spam messages, the score $I(\hat{\mathbf{x}})$ of an incoming test message $\hat{\mathbf{x}}$ can be maximized by maximizing each $q_j$ separately. This motivates dictionary attacks and focused attacks—intuitively, the attacker would like to maximally increase the $q_j$ of tokens appearing (or most likely to appear) in $\hat{\mathbf{x}}$ depending on the information the attacker has about future messages.

Thus, we first analyze the effect of increasing $n_j^{(s)}$ on its score $q_i$ in Section C.2.1. Based on this, we subsequently analyze the change in $I(\hat{\mathbf{x}})$ that is caused altering the token score $q_i$ in Section C.2.2. As one might expect, since increasing the number of

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

occurrences of the $j^{\text{th}}$ token in spam should increase the posterior probability that a message with the $j^{\text{th}}$ token is spam, we show that including the $j^{\text{th}}$ token in an attack message generally increases the corresponding score $q_j$ more than not including that token (except in unusual situations which we identify below). Similarly, we show that increasing $q_j$ generally increases the overall spam score $I(\cdot)$ of a message containing the $j^{\text{th}}$ token. Based on these results, we motivate the attack strategies presented in Section 5.3.1.

### C.2.1 Effect of poisoning on token scores

In this section, we establish how token spam scores change as the result of attack messages in the training set. Intuitively, one might expect that the $j^{\text{th}}$ score $q_j$ should increase when the $j^{\text{th}}$ token is added to the attack email. This would be the case, in fact, if the token score in Equation (5.1) were computed according to Bayes' Rule. However, as noted Section 5.1, the score in Equation (5.1) is derived by applying Bayes' Rule with an additional assumption that the prior distribution of spam and ham is equal. As a result, there are circumstances in which the spam score $q_j$ can decrease when the $j^{\text{th}}$ token is included in the attack email—specifically when the assumption is violated. We show that this occurs when there is an extraordinary imbalance between the number of ham and spam in the training set.

As in Section 5.3, we consider an attacker whose attack messages are composed a single set of attack tokens; i.e., each token is either included in *all* attack messages or *none*. In this fashion, the attacker creates a set of $k$ attack messages used in the retraining of the filter, after which the counts become

$$\begin{aligned} N^{(s)} &\mapsto N^{(s)} + k \\ N^{(h)} &\mapsto N^{(h)} \\ n_j^{(s)} &\mapsto \begin{cases} n_j^{(s)} + k, & \text{if } a_j = 1 \\ n_j^{(s)}, & \text{otherwise} \end{cases} \\ n_j^{(h)} &\mapsto n_j^{(h)}. \end{aligned}$$

Using these count transformations, we compute the difference in the smoothed SpamBayes score $q_j$ between training on an attack spam message $\mathbf{a}$ that contains the $j^{\text{th}}$ token and an attack spam that does not contain it. If the $j^{\text{th}}$ token is included in the attack (i.e., $a_j = 1$), then the new score for the $j^{\text{th}}$ token (from Equation 5.1) is

$$P_j^{(s,k)} \triangleq \frac{N^{(h)}\left(n_j^{(s)} + k\right)}{N^{(h)}\left(n_j^{(s)} + k\right) + \left(N^{(s)} + k\right) n_j^{(h)}}.$$

If the token is not included in the attack (i.e., $a_j = 0$), then the new token score is

$$P_j^{(s,0)} \triangleq \frac{N^{(h)} n_j^{(s)}}{N^{(h)} n_j^{(s)} + \left(N^{(s)} + k\right) n_j^{(h)}}.$$

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

Similarly, we use $q_j^{(k)}$ and $q_j^{(0)}$ to denote the smoothed spam score after the attack depending on whether or not the $j^{\text{th}}$ token was used in the attack message. We will analyze the quantity

$$\Delta^{(k)} q_j \triangleq q_j^{(k)} - q_j^{(0)}.$$

One might reasonably expect this difference to always be non-negative, but here we show that there are some scenarios in which $\Delta^{(k)} q_j < 0$. This unusual behavior is a direct result of the assumption made by SpamBayes that $N^{(h)} = N^{(s)}$ rather than using a proper prior distribution. In fact, it can be shown that the usual spam model depicted in Figure 5.1(b) does not exhibit these irregularities. Below, we will show how SpamBayes' assumption can lead to situations where $\Delta^{(k)} q_j < 0$ but also that these irregularities only occur when there is *many* more spam messages than ham messages in the training dataset. By expanding $\Delta^{(k)} q_j$ and rearranging terms, the difference can be expressed as:

$$\begin{aligned}\Delta^{(k)} q_j = &\frac{s \cdot k}{\left(s + n_j + k\right)\left(s + n_j\right)}\left(P_j^{(s,k)} - x\right) \\ &+ \frac{k \cdot N^{(h)} \cdot n_j}{\left(s + n_j\right)\left(N^{(h)} \cdot n_j^{(s)} + \left(N^{(s)} + k\right) n_j^{(h)}\right)} P_j^{(h,k)},\end{aligned}$$

where $P_j^{(h,k)} = 1 - P_j^{(s,k)}$ is the altered ham score of the $j^{\text{th}}$ token. The difference can be rewritten as

$$\Delta^{(k)} q_j = \frac{k}{\left(s + n_j + k\right)\left(s + n_j\right)} \cdot \alpha_j$$

$$\begin{aligned}\alpha_j \triangleq\ & s\,(1 - x) \\ &+ P_j^{(h,k)} \cdot \frac{N^{(h)} \cdot n_j\left(n_j + k\right) + s \cdot N^{(h)} \cdot n_j^{(h)} - s\left(N^{(s)} + k\right) n_j^{(h)}}{N^{(h)} \cdot n_j^{(s)} + \left(N^{(s)} + k\right) n_j^{(h)}}.\end{aligned}$$

The first factor $\frac{k}{(s+n_j+k)(s+n_j)}$ in the above expression is non-negative so only $\alpha_j$ can make $\Delta^{(k)} q_j$ negative. From this, it is easy to show that $N^{(s)} + k$ must be greater that $N^{(h)}$ for $\Delta^{(k)} q_j$ to be negative, but we demonstrate stronger conditions. Generally, we demonstrate that for $\Delta^{(k)} q_j$ to be negative there must be a large disparity between the number of spams after the attack, $N^{(s)} + k$, and the number of ham messages, $N^{(h)}$. This reflects the effect of violating the implicit assumption made by SpamBayes that $N^{(h)} = N^{(s)}$.

Expanding the expression for $\alpha_j$, the following condition is necessary for $\Delta^{(k)} q_j$ to be negative:

$$\begin{aligned}\frac{s\left(N^{(s)} + k\right) n_j^{(h)} x}{N^{(h)}} > &\frac{s(1-x)\left(n_j^{(s)} + k\right)}{n_j^{(h)}\left(N^{(s)} + k\right)}\left[\left(N^{(s)} + k\right) n_j^{(h)} + N^{(h)} \cdot n_j^{(s)}\right] \\ &+ \quad n_j\left(n_j + k\right) + s n_j^{(s)}(1 - x) + s \cdot n_j^{(h)}\end{aligned}.$$

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

Because $1 - x \geq 0$ (since $x \leq 1$) and $n_j = n_j^{(s)} + n_j^{(h)}$, the right-hand side of the above expression is strictly increasing in $n_j^{(s)}$ while the left-hand side is constant in $n_j^{(s)}$. Thus, the weakest condition to make $\Delta^{(k)} q_j$ negative occurs when $n_j^{(s)} = 0$; i.e., tokens that were not observed in any spam prior to the attack are most susceptible to having $\Delta^{(k)} q_j < 0$ while tokens that were observed more frequently in spam prior distribution to the attack require an increasingly larger disparity between $N^{(h)}$ and $N^{(s)}$ for $\Delta^{(k)} q_j < 0$ to occur. Here we analyze the case when $n_j^{(s)} = 0$ and, using the previous constraints that $s > 0$ and $n_j^{(h)} > 0$, we arrive at the weakest condition for which $\Delta^{(k)} q_j$ can be negative. This condition can be expressed succinctly as the following condition on $x$ for the attack to cause a token's score to decrease:[1]

$$x > \frac{N^{(h)} \left(n_j^{(h)} + s\right) \left(n_j^{(h)} + k\right)}{s \left(n_j^{(h)} \left(N^{(s)} + k\right) + k N^{(h)}\right)}.$$

First, notice that the right-hand side is always positive; i.e., there will always be some non-trivial threshold on the value of $x$ to allow for $\Delta^{(k)} q_j$ to be negative. Further, when the right-hand side of this bound is at least one, there are no tokens that have a negative $\Delta^{(k)} q_j$ since the parameter $x \in [0, 1]$. For instance, this occurs when $n_j^{(h)} = 0$ or when $N^{(h)} \geq N^{(s)} + k$ (as previously noted).

Reorganizing the terms, the bound on the number of spams can be expressed as,

$$N^{(s)} + k > N^{(h)} \cdot \frac{\left(n_j^{(h)}\right)^2 + (s + k)\, n_j^{(h)} + s\,(1 - x)\, k}{s n_j^{(h)} x}.$$

This bound shows that the number of spam after the attack, $N^{(s)} + k$, must be larger than a multiple of total number of ham, $N^{(h)}$, to have any token with $\Delta^{(k)} q_j < 0$. The factor in this multiple is always greater than one, but depends on the $n_j^{(h)}$ of the $j^{\text{th}}$ token. In fact, the factor is strictly increasing in $n_j^{(h)}$; thus, the weakest bound occurs when $n_j^{(h)} = 1$ (recall that when $n_j^{(h)} = 0$, $\Delta^{(k)} q_j$ is always non-negative). When we examine SpamBayes' default values of $s = 1$ and $x = \frac{1}{2}$, the weakest bound (for tokens with $n_j^{(h)} = 1$ and $n_j^{(s)} = 0$) is

$$N^{(s)} + k > N^{(h)} \cdot (4 + 3k)\,.$$

Thus, when the number of spam after the attack, $N^{(s)} + k$, is sufficiently larger than the number of ham, $N^{(h)}$, it is possible that the score of a token will be lower if it is *included* in the attack message than if it were excluded. This is a direct result of the assumption made by SpamBayes that $N^{(s)} = N^{(h)}$. We have shown that such aberrations will occur most readily in tokens with low initial values of $n_j^{(h)}$ and $n_j^{(s)}$; i.e., those seen infrequently in the dataset. However, for any significant number of attacks, $k$, the disparity between $N^{(s)} + k$ and $N^{(s)}$ must be tremendous for such aberrations to occur. Under the default SpamBayes settings, there would have to be at least 7 times as many spam as ham with

[1] In the case that $n_j^{(s)} > 0$, the condition is stronger but the expression is more complicated.

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

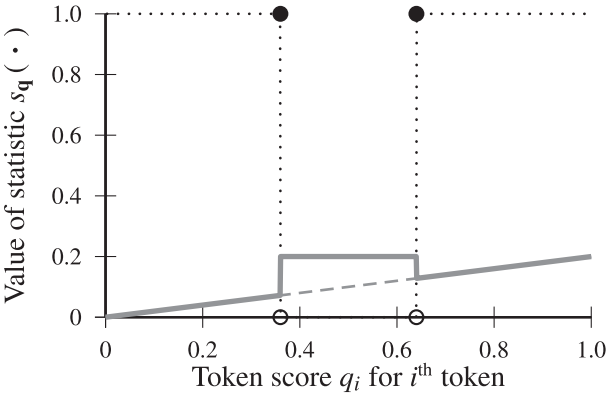


**Figure C.1** Plot of the aggregation statistic $s_{\mathbf{q}}(\cdot)$ relative to a single token score $q_i$; on the $x$-axis is $q_i$ and on the $y$-axis is $s_{\mathbf{q}}(\cdot)$. Here we consider a scenario where $\tau_{\hat{\mathbf{x}}} = 0.14$ and without the $i^{\text{th}}$ token $s_{\mathbf{q}}(\hat{\mathbf{x}} \setminus \{i\}) = 0.2$. The black dotted line is the value of $\delta(\hat{\mathbf{x}})_i$, the gray dotted line is the value of $q_i \prod_{j \neq i} q_j$ (i.e., $s_{\mathbf{q}}(\hat{\mathbf{x}})$ without including $\boldsymbol{\delta}(\hat{\mathbf{x}})$), and the gray solid line is the value of $s_{\mathbf{q}}(\hat{\mathbf{x}})$ as $q_i$ varies.

only a single attack message. For more attack messages ($k > 1$), this bound is even greater. Thus, in designing attacks against SpamBayes, we ignore the extreme cases outlined here and we assume that $\Delta^{(k)} q_j$ always increases if the $j^{\text{th}}$ token is included in the attack. Further, none of the experiments presented in Section 5.5 meet the criteria required to have $\Delta^{(k)} q_j < 0$.

### C.2.2 Effect of poisoning on $I(\cdot)$

The key to understanding effect of attacks and constructing optimal attacks against SpamBayes is characterizing conditions under which SpamBayes' score $I(\hat{\mathbf{x}})$ increases when the training corpus is injected with attack spam messages. To do this, we dissect the method used by SpamBayes to aggregate token scores.

The statistics $s_{\mathbf{q}}(\hat{\mathbf{x}})$ and $h_{\mathbf{q}}(\hat{\mathbf{x}})$ from Equation (C.1) and (C.2) are measures of the *spaminess* and *haminess* of the message represented by $\hat{\mathbf{x}}$, respectively. Both assume that each token in the message presents an assessment of the *spaminess* of the message—the score $q_i$ is the evidence for spam given by observing the $i^{\text{th}}$ token. Further, by assuming independence, $s_{\mathbf{q}}(\hat{\mathbf{x}})$ and $h_{\mathbf{q}}(\hat{\mathbf{x}})$ aggregate this evidence into a measure of the overall message's *spaminess*. For instance, if all tokens have $q_i = 1$, $s_{\mathbf{q}}(\hat{\mathbf{x}}) = 1$ indicates that the message is very *spammy* and $1 - h_{\mathbf{q}}(\hat{\mathbf{x}}) = 1$ concurs. Similarly, when all tokens have $q_i = 0$, both scores indicate that the message is ham.

These statistics also are (almost) nicely behaved. If we instead consider the ordinary product of the scores of all tokens in the message $\hat{\mathbf{x}}$, $\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}}) \triangleq \prod_{i:\hat{x}_i=1} q_i$, it is a linear function with respect to each $q_i$, and is monotonically non-decreasing. Similarly, the product $\tilde{h}_{\mathbf{q}}(\hat{\mathbf{x}}) \triangleq \prod_{i:\hat{x}_i=1} (1 - q_i)$ is linear with respect to each $q_i$ and is monotonically non-increasing. Thus, if we increase any score $q_i$, the first product will not decrease

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

and the second will not increase, as expected.[2] In fact, by redefining the scores $I(\cdot)$, $S(\cdot)$, and $H(\cdot)$ in terms of the simple products $\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})$ and $\tilde{h}_{\mathbf{q}}(\hat{\mathbf{x}})$ (which we refer to as $\tilde{I}(\cdot)$, $\tilde{S}(\cdot)$, and $\tilde{H}(\cdot)$, respectively), the following lemma shows that $\tilde{I}(\cdot)$ is non decreasing in $q_i$.

LEMMA C.2 *The modified $\tilde{I}(\hat{\mathbf{x}})$ score is non-decreasing in $q_i$ for all tokens (indexed by i).*

*Proof* We show that the derivative of $\tilde{I}(\hat{\mathbf{x}})$ with respect to $q_k$ is non-negative for all $k$. By rewriting, Equation (5.3) in terms of $\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})$ as $\tilde{S}(\hat{\mathbf{x}}) = 1 - \chi^2_{2n}\left(-2\log\left(\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})\right)\right)$, the chain rule can be applied as follows:

$$\frac{\partial}{\partial q_k}\tilde{S}(\hat{\mathbf{x}}) = \frac{d}{d\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})}\left[1 - \chi^2_{2n}\left(-2\log\left(\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})\right)\right)\right] \cdot \frac{\partial}{\partial q_k}\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})$$

$$\frac{d}{d\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})}\left[1 - \chi^2_{2n}\left(-2\log\left(\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})\right)\right)\right] = \frac{1}{(n-1)!}\left(-\log\left(\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})\right)\right)^{n-1}.$$

The second derivative is non-negative since $0 \leq \tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}}) \leq 1$. Further, the partial derivative of $\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}})$ with respect to $q_k$ is simply $\frac{\partial}{\partial q_k}\tilde{s}_{\mathbf{q}}(\hat{\mathbf{x}}) = \prod_{i \neq k : \hat{x}_i = 1} q_i \geq 0$. Thus, for all $k$,

$$\frac{\partial}{\partial q_k}\tilde{S}(\hat{\mathbf{x}}) \geq 0.$$

By an analogous derivation, replacing $q_i$ by $1 - q_i$,

$$\frac{\partial}{\partial q_k}\tilde{H}(\hat{\mathbf{x}}) \leq 0.$$

The final result is then give by

$$\frac{\partial}{\partial q_k}\tilde{I}(\hat{\mathbf{x}}) = \frac{1}{2}\frac{\partial}{\partial q_k}\tilde{S}(\hat{\mathbf{x}}) - \frac{1}{2}\frac{\partial}{\partial q_k}\tilde{H}(\hat{\mathbf{x}}) \geq 0.$$

□

However, unlike the simple products, the statistics $s_{\mathbf{q}}(\cdot)$ and $h_{\mathbf{q}}(\cdot)$ have unusual behavior because the function $\boldsymbol{\delta}(\cdot)$ sanitizes the token scores. Namely, $\boldsymbol{\delta}(\cdot)$ is the indicator function of the set $\mathbb{T}_{\hat{\mathbf{x}}}$. Membership in this set is determined by absolute distance of a token's score from the agnostic score of $\frac{1}{2}$; i.e., by the value $g_i \triangleq \left|q_i - \frac{1}{2}\right|$. The $i^{\text{th}}$ token belongs to $\mathbb{T}_{\hat{\mathbf{x}}}$ if *i)* $\hat{x}_i = 1$ *ii)* $g_i \geq Q$ (by default $Q = 0.1$ so all tokens in (0.4, 0.6) are excluded) and *iii)* of the remaining tokens, the token has among the largest $T$ values of $g_i$ (by default $T = 150$).

For our purposes, for every message $\hat{\mathbf{x}}$, there is some value $\tau_{\hat{\mathbf{x}}} < \frac{1}{2}$ that defines an interval $\left(\frac{1}{2} - \tau_{\hat{\mathbf{x}}}, \frac{1}{2} + \tau_{\hat{\mathbf{x}}}\right)$ to exclude tokens. That is

$$\delta(\hat{\mathbf{x}})_i = \hat{x}_i \cdot \begin{cases} 0 & \text{if } q_i \in \left(\frac{1}{2} - \tau_{\hat{\mathbf{x}}}, \frac{1}{2} + \tau_{\hat{\mathbf{x}}}\right) \\ 1 & \text{otherwise} \end{cases}.$$

[2] These statistics also behave oddly in another sense. Namely, adding an additional token will always decrease both products and removing a token will always increase both products. Applying the chi-squared distribution rectifies this effect.

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

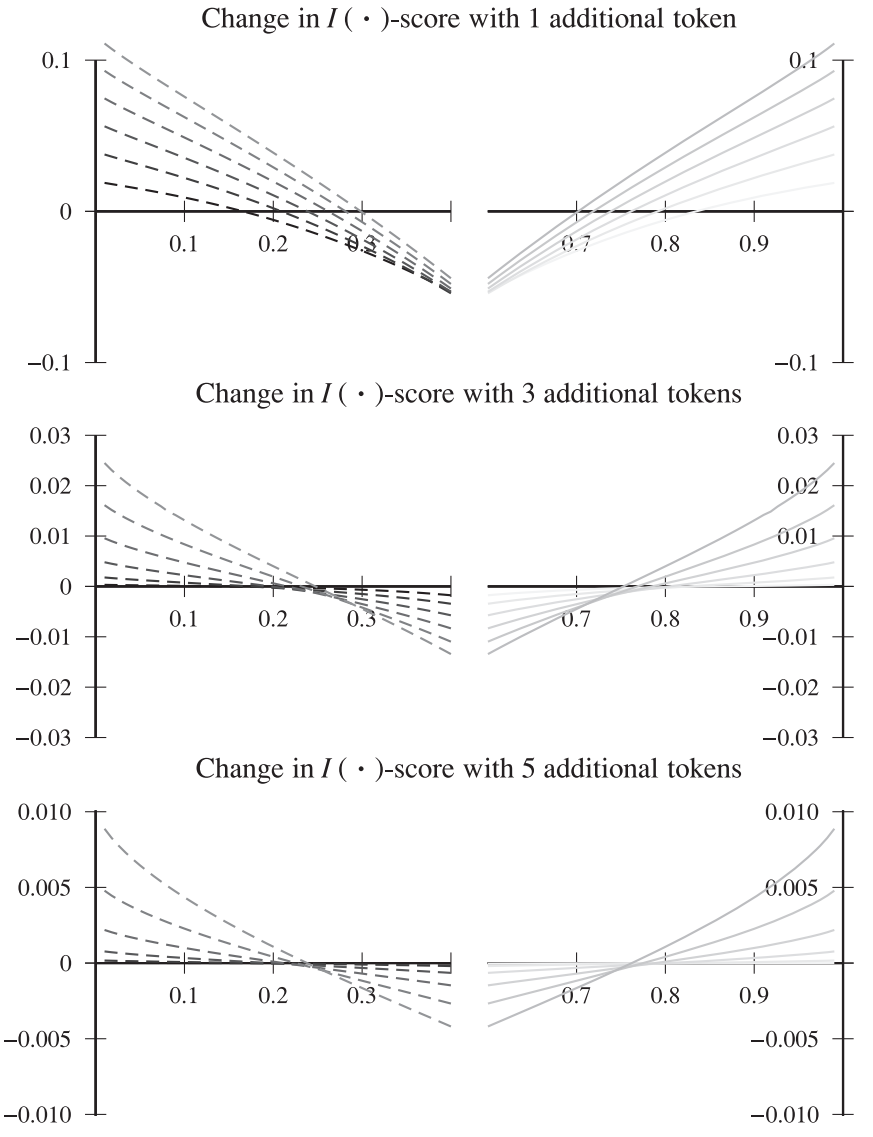


**Figure C.2** The effect of the $\delta(\cdot)$ function on $I(\cdot)$ as the score of the $i^{\text{th}}$ token, $q_i$, increases causing $q_i$ to move into or out of the region (0.4, 0.6) where all tokens are ignored. In each plot, the $x$-axis is the value of $q_i$ before it's removal and the $y$-axis is the change in $I(\cdot)$ due to the removal; note that the scale on the $y$-axis decreases from top to bottom. For the top-most row of plots there is 1 unchanged token scores in addition to the changing one, for the middle row there are 3 additional unchanged token scores, and for the bottom row there are 5 additional unchanged token scores. The plots in the left-most column demonstrate the effect of removing the $i^{\text{th}}$ token when initially $q_i \in (0, 0.4)$; the scores of the additional unchanging tokens are all fixed to the same value of 0.02 (dark dashed black), 0.04, 0.06, 0.08, 0.10, or 0.12 (light dashed black). The plots in the right-most column demonstrate the effect of adding the $i^{\text{th}}$ token when initially $q_i \in (0.4, 0.6)$; the scores of the additional unchanging tokens are all fixed to the same value of 0.88 (dark gray), 0.90, 0.92, 0.94, 0.96, or 0.98 (light gray).

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press

Clearly, for tokens in $\hat{\mathbf{x}}$, $\delta(\hat{\mathbf{x}})_i$ steps from 1 to 0 and back to 1 as $q_i$ increases. This causes $s_{\mathbf{q}}(\hat{\mathbf{x}})$ to have two discontinuities with respect to $q_i$: it increases linearly on the intervals $\left[0, \frac{1}{2} - \tau_{\hat{\mathbf{x}}}\right]$ and $\left[\frac{1}{2} + \tau_{\hat{\mathbf{x}}}, 1\right]$, but on the middle interval $\left(\frac{1}{2} - \tau_{\hat{\mathbf{x}}}, \frac{1}{2} + \tau_{\hat{\mathbf{x}}}\right)$ it jumps discontinuously to its maximum value. This behavior of is depicted in Figure C.1. Similarly, $h_{\mathbf{q}}(\hat{\mathbf{x}})$ decreases linearly except on the middle interval $\left(\frac{1}{2} - \tau_{\hat{\mathbf{x}}}, \frac{1}{2} + \tau_{\hat{\mathbf{x}}}\right)$ where it also jumps to its maximum value. Thus, neither $s_{\mathbf{q}}(\hat{\mathbf{x}})$ or $h_{\mathbf{q}}(\hat{\mathbf{x}})$ have monotonic behavior on the interval [0, 1].

To better understand how $I(\hat{\mathbf{x}})$ behaves when $q_i$ increases given that neither $s_{\mathbf{q}}(\hat{\mathbf{x}})$ or $h_{\mathbf{q}}(\hat{\mathbf{x}})$ are monotonic, we analyze its behavior on a case by case basis. For this purpose, we refer to the three intervals $\left[0, \frac{1}{2} - \tau_{\hat{\mathbf{x}}}\right]$, $\left(\frac{1}{2} - \tau_{\hat{\mathbf{x}}}, \frac{1}{2} + \tau_{\hat{\mathbf{x}}}\right)$, and $\left[\frac{1}{2} + \tau_{\hat{\mathbf{x}}}, 1\right]$ as $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$, respectively. Clearly, if $q_i$ increases but stays within the same interval, $I(\hat{\mathbf{x}})$ also increases. This follows from Lemma C.2 and the fact that $I(\hat{\mathbf{x}})$ will not change if $q_i$ remains within interval $\mathbb{B}$. Similarly, $I(\hat{\mathbf{x}})$ also increases if $q_i$ increases from interval $\mathbb{A}$ to interval $\mathbb{C}$; this too follows from Lemma C.2. The only cases when $I(\hat{\mathbf{x}})$ may decrease when $q_i$ increases occur when either $q_i$ transitions from interval $\mathbb{A}$ to $\mathbb{B}$ or $q_i$ transitions from interval $\mathbb{B}$ to $\mathbb{C}$, but in these cases, the behavior of $I(\hat{\mathbf{x}})$ depends heavily on the scores for the other tokens in $\hat{\mathbf{x}}$ and the value of $q_i$ before it increases as depicted by Figure C.2. It is also worth noting that $I(\hat{\mathbf{x}})$ in fact will *never* decrease if $\hat{\mathbf{x}}$ has more than 150 tokens outside the interval (0.4, 0.6), since in this case increasing $q_i$ either into or out of $\mathbb{B}$ also corresponds to either adding or removing a second token score $q_j$. The effect in this case is that $I(\hat{\mathbf{x}})$ always increases.

The attacks against SpamBayes that we introduced in Section 5.3 ignore the fact that $I(\hat{\mathbf{x}})$ may decrease when increasing some token scores. In this sense, these attacks are not truly optimal. However, determining which set of tokens will optimally increase the overall $I(\cdot)$ of a set of future messages $\{\hat{\mathbf{x}}\}$ is a combinatorial problem that seems infeasible for a real-world adversary. Instead, we consider attacks that are optimal for the relaxed version of the problem that incorporates *all* tokens from $\hat{\mathbf{x}}$ in computing $I(\hat{\mathbf{x}})$. Further, in Section 5.5, we show that these approximate techniques are extraordinarily effective against SpamBayes in spite of the fact some non-optimal tokens are included in the attack messages.

https://doi.org/10.1017/9781107338548.012 Published online by Cambridge University Press