# Appendix B Full Proofs for Hypersphere Attacks

In this appendix, we give proofs for the theorems from Chapter 4. For this purpose, we introduce the concept of $(\tau, k)$-*differing sequences*, which are a pair of sequences $\mathbf{a}, \mathbf{b} \in \mathcal{A}^{(M,\infty)}$ that are everywhere identical except in the $\tau^{\text{th}}$, $\tau+1^{\text{th}}$, …, $\tau+k^{\text{th}}$ consecutive elements and have the following mass-balance property:

$$\sum_{t=\tau}^{\tau+k} a_t = \sum_{t=\tau}^{\tau+k} b_t. \tag{B.1}$$

The following lemma for $(\tau, 1)$-differing sequences simplifies several of the subsequent proofs.

LEMMA B.1 *For any* $(\tau, 1)$*-differing sequences* $\mathbf{a}, \mathbf{b} \in \mathcal{A}^{(M,\infty)}$ *that are identical except in their* $\tau^{th}$ *and* $(\tau+1)^{th}$ *elements (with* $a_\tau + a_{\tau+1} = b_\tau + b_{\tau+1}$ *from Equation* (B.1)*), the difference between the distances of these sequences,* $\Delta_{\mathbf{a},\mathbf{b}} \triangleq D(\mathbf{a}) - D(\mathbf{b})$*, can be expressed as*

$$\Delta_{\mathbf{a},\mathbf{b}} = \frac{\left(\mu^{(\mathbf{a})}_{\tau-1} + b_\tau\right) \cdot a_\tau \cdot a_{\tau+1} - \left(\mu^{(\mathbf{a})}_{\tau-1} + a_\tau\right) \cdot b_\tau \cdot b_{\tau+1}}{\left(\mu^{(\mathbf{a})}_{\tau-1} + a_\tau\right)\left(\mu^{(\mathbf{a})}_{\tau-1} + b_\tau\right)\left(\mu^{(\mathbf{a})}_{\tau-1} + a_\tau + a_{\tau+1}\right)} \tag{B.2}$$

*where* $\mu^{(\mathbf{a})}_t = N + \sum_{\ell=1}^{t} a_\ell$ *is the cumulative sum of the first* $t$ *elements of the sequence* $\mathbf{a}$ *as in Equation* (4.7)*. This holds so long as either* $\mu^{(\mathbf{a})}_{\tau-1} > 0$ *or both* $a_\tau > 0$ *and* $b_\tau > 0$*.*

*Proof* First, for $t < \tau$, $\mu^{(\mathbf{a})}_t = \mu^{(\mathbf{b})}_t$ since the two sequences are identical until the $\tau^{\text{th}}$ element. Similarly, for $t > \tau + 1$ we again have $\mu^{(\mathbf{a})}_t = \mu^{(\mathbf{b})}_t$ since the sequences only differ in their $\tau^{\text{th}}$ and $(\tau+1)^{\text{th}}$ elements and they are mass-balanced according to Equation (B.1) for which we define $\gamma \triangleq a_\tau + a_{\tau+1} = b_\tau + b_{\tau+1}$ as the balance constant for these sequences. Using these two facts and that, from Equation (4.12), $\delta_t(\mathbf{a}) = \frac{a_t}{\mu^{(\mathbf{a})}_t}$, we have that $\delta_t(\mathbf{a}) = \delta_t(\mathbf{b})$ if $t < \tau$ or $t > \tau + 1$. Thus difference in the distances achieved

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

by these two sequences is given from Equations (4.11) can be expressed as

$$\begin{aligned}
\Delta_{\mathbf{a},\mathbf{b}} &= \sum_{t=1} \left[\delta_t(\mathbf{a}) - \delta_t(\mathbf{b})\right] \\
&= \underbrace{\sum_{t=1}^{\tau-1} \left[\delta_t(\mathbf{a}) - \delta_t(\mathbf{b})\right]}_{=0} + \delta_\tau(\mathbf{a}) - \delta_\tau(\mathbf{b}) + \delta_{\tau+1}(\mathbf{a}) - \delta_{\tau+1}(\mathbf{b}) \\
&\quad + \underbrace{\sum_{t=\tau+2} \left[\delta_t(\mathbf{a}) - \delta_t(\mathbf{b})\right]}_{=0} \\
&= \frac{a_\tau}{\mu_\tau^{(\mathbf{a})}} - \frac{b_\tau}{\mu_\tau^{(\mathbf{b})}} + \frac{a_{\tau+1}}{\mu_{\tau+1}^{(\mathbf{a})}} - \frac{b_{\tau+1}}{\mu_{\tau+1}^{(\mathbf{b})}} \\
&= \frac{a_\tau}{\mu_{\tau-1}^{(\mathbf{a})} + a_\tau} - \frac{b_\tau}{\mu_{\tau-1}^{(\mathbf{a})} + b_\tau} + \frac{a_{\tau+1}}{\mu_{\tau-1}^{(\mathbf{a})} + \gamma} - \frac{b_{\tau+1}}{\mu_{\tau-1}^{(\mathbf{a})} + \gamma}.
\end{aligned}$$

To combine these four terms, we can obtain a common denominator of $\Gamma = \left(\mu_{\tau-1}^{(\mathbf{a})} + a_\tau\right)\left(\mu_{\tau-1}^{(\mathbf{a})} + b_\tau\right)\left(\mu_{\tau-1}^{(\mathbf{a})} + \gamma\right)$ for which the combined numerator is given by

$$\begin{aligned}
&\left[a_\tau\left(\mu_{\tau-1}^{(\mathbf{a})} + b_\tau\right) - b_\tau\left(\mu_{\tau-1}^{(\mathbf{a})} + a_\tau\right)\right]\left(\mu_{\tau-1}^{(\mathbf{a})} + \gamma\right) \\
&\quad + (a_{\tau+1} - b_{\tau+1})\left(\mu_{\tau-1}^{(\mathbf{a})} + a_\tau\right)\left(\mu_{\tau-1}^{(\mathbf{a})} + b_\tau\right) \\
&= \mu_{\tau-1}^{(\mathbf{a})} a_\tau a_{\tau+1} - \mu_{\tau-1}^{(\mathbf{a})} b_\tau b_{\tau+1} + b_\tau a_\tau a_{\tau+1} - a_\tau b_\tau b_{\tau+1} \\
&= \left(\mu_{\tau-1}^{(\mathbf{a})} + b_\tau\right) \cdot a_\tau \cdot a_{\tau+1} - \left(\mu_{\tau-1}^{(\mathbf{a})} + a_\tau\right) \cdot b_\tau \cdot b_{\tau+1},
\end{aligned}$$

in which we used the definition of $\gamma$ to cancel terms. Combining this numerator with the denominator $\Gamma$ yields Equation (B.2). Finally the condition that either $\mu_{\tau-1}^{(\mathbf{a})} > 0$ or both $a_\tau > 0$ and $b_\tau > 0$ is necessary to prevent the denominator from becoming zero. □

## B.1 Proof of Theorem 4.7

Here we show that the optimal attack according to Equation (4.5) can be optimized in a greedy fashion. Further, we show that the optimal attack points are all placed at the intersection of the hypersphere's boundary and the desired attack direction.

*Proof* Consider the $t^{\text{th}}$ iteration of the attack for any $t \in \mathfrak{N}$. The attacker's goal in the $t^{\text{th}}$ iteration is to maximize the displacement alignment given in Equation (4.5). The attacker accomplishes this by crafting a set of $\alpha_t \in \mathfrak{N}$ attack points: $\mathbb{A}^{(t)} = \left\{\mathbf{a}^{(t,\ell)}\right\}_{\ell=1}^{\alpha_t}$. These points are designed to maximize $\mathbf{D}_t^\top \frac{\mathbf{x}^A - \mathbf{c}^{(0)}}{\left\|\mathbf{x}^A - \mathbf{c}^{(0)}\right\|}$ where $\mathbf{D}_t = \frac{\mathbf{c}^{(t)} - \mathbf{c}^{(0)}}{R}$ by Equation (4.4), $\mathbf{c}^{(t)}$ is defined recursively by Equation (4.8), and each attack vector is constrained to lie within the $(t-1)^{\text{th}}$ hypersphere; i.e., $\left\|\mathbf{a}^{(t,\ell)} - \mathbf{c}^{(t-1)}\right\| \leq R$ for all $\ell = 1, \ldots, \alpha_t$. The attacker's objective can be modified without loss of

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

generality by first transforming the space so that $\mathbf{c}^{(t-1)} = \mathbf{0}$ (via the transform $\hat{\mathbf{x}} \mapsto \mathbf{x} - \mathbf{c}^{(t-1)}$). This yields the following equivalent program that the attack optimizes:

$$\max_{\mathbb{A}^{(t)}} \rho\left(\hat{\mathbf{D}}_t\right) = \hat{\mathbf{D}}_t^\top \frac{\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}}{\left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|}$$
$$\text{s.t.} \qquad \forall\, \ell \in 1, \ldots, \alpha_t \quad \left\|\hat{\mathbf{a}}^{(t,\ell)}\right\|^2 \leq R^2,$$

where $\hat{\mathbf{D}}_t = \frac{\hat{\mathbf{c}}^{(t)} - \hat{\mathbf{c}}^{(0)}}{R}$ and $\hat{\mathbf{c}}^{(t)}$ takes the simplified form $\hat{\mathbf{c}}^{(t)} = \frac{1}{\mu_t} \sum_{\ell=1}^{\alpha_t} \hat{\mathbf{a}}^{(t,\ell)}$. The Lagrangian for this program at the $t^{\text{th}}$ attack iteration is

$$\begin{aligned}
\mathcal{L}_t\left(\left\{\hat{\mathbf{a}}^{(t,\ell)}\right\}, \boldsymbol{\lambda}\right) &= \hat{\mathbf{D}}_t^\top \frac{\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}}{\left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|} - \sum_{\ell=1}^{\alpha_t} \lambda_\ell \left(\left\|\hat{\mathbf{a}}^{(t,\ell)}\right\|^2 - R^2\right) \\
&= \frac{1}{R\mu_t \left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|} \sum_{\ell=1}^{\alpha_t} \left(\hat{\mathbf{a}}^{(t,\ell)}\right)^\top \left(\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right) - \sum_{\ell=1}^{\alpha_t} \lambda_\ell (\hat{\mathbf{a}}^{(t,\ell)})^\top \hat{\mathbf{a}}^{(t,\ell)} \\
&\quad - \frac{\left(\hat{\mathbf{c}}^{(0)}\right)^\top \left(\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right)}{R\left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|} + R^2 \sum_{\ell=1}^{\alpha_t} \lambda_\ell
\end{aligned}$$

where the variables $\lambda_\ell \geq 0$ are the Lagrangian multipliers and the second equality follows from expanding the above form of $\hat{\mathbf{D}}_t$.

We compute the partial derivatives of $\mathcal{L}_t\left(\left\{\hat{\mathbf{a}}^{(t,\ell)}\right\}, \boldsymbol{\lambda}\right)$ with respect to the Lagrangian multipliers $\boldsymbol{\lambda}$ and set them to zero to reveal that, at a solution, $\left\|\hat{\mathbf{a}}^{(t,\ell)}\right\| = R$. Further, by the complementary slackness conditions that arise from the dual of the above program, it follows that the Lagrangian multipliers are non-zero; i.e., $\forall\, i\ \lambda_i \geq 0$. Then computing the partial derivatives of $\mathcal{L}_t\left(\left\{\hat{\mathbf{a}}^{(t,\ell)}\right\}, \boldsymbol{\lambda}\right)$ with respect to each $\hat{\mathbf{a}}^{(t,\ell)}$ and setting them to zero reveals that, at a solution, we must have for all $\ell$,

$$\hat{\mathbf{a}}^{(t,\ell)} = \frac{1}{2\lambda_\ell R\mu_t} \frac{\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}}{\left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|},$$

which demonstrates that all optimal attack vectors must be a scaled version of the vector $\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}$. Thus, by the fact that $\left\|\hat{\mathbf{a}}^{(t,\ell)}\right\| = R$, we must have

$$\hat{\mathbf{a}}^{(t,\ell)} = R \cdot \frac{\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}}{\left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|} \qquad \text{and} \qquad \hat{\mathbf{c}}^{(t)} = R \cdot \frac{\alpha_t}{\mu_t} \cdot \frac{\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}}{\left\|\hat{\mathbf{x}}^A - \hat{\mathbf{c}}^{(0)}\right\|}.$$

By reversing the transform making $\mathbf{c}^{(t-1)} = \mathbf{0}$, the attack vectors can be expressed as

$$\mathbf{a}^{(t,\ell)} = \mathbf{c}^{(t-1)} + R \cdot \frac{\mathbf{x}^A - \mathbf{c}^{(0)}}{\left\|\mathbf{x}^A - \mathbf{c}^{(0)}\right\|},$$

which gives the first part of the theorem. Similarly by reversing this transform for the centroids and solving the resulting simple recursion we arrive at

$$\mathbf{c}^{(t)} = \mathbf{c}^{(t-1)} + R \cdot \frac{\alpha_t}{\mu_t} \cdot \frac{\mathbf{x}^A - \mathbf{c}^{(0)}}{\left\|\mathbf{x}^A - \mathbf{c}^{(0)}\right\|} \qquad = \mathbf{c}^{(0)} + R \cdot \frac{\mathbf{x}^A - \mathbf{c}^{(0)}}{\left\|\mathbf{x}^A - \mathbf{c}^{(0)}\right\|} \cdot \sum_{\ell=1}^{t} \frac{\alpha_\ell}{\mu_\ell},$$

as was to be shown. □

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

## B.2 Proof of Theorem 4.14

*Proof* We show that any optimal sequence with $M \in \mathfrak{N}_0$ attack points (in the sense of Definition 4.10) must have a monotonically increasing sequence of non-zero elements. For $M = 0$, the trivial sequence $\boldsymbol{\alpha}^\star = \mathbf{0}$ is the only sequence in $\mathcal{A}^{(M,\infty)}$ and thus is optimal (and trivially satisfies the theorem).

For $M > 0$, the proof proceeds *by contradiction* by assuming that there exists an such that there is an optimal sequence, $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M,\infty)}$ with a sub-sequence of non-zero elements that is not monotonically non-decreasing. To simplify the proof, we instead consider an equivalent sequence (with respect to the distance function) with all interleaving zero elements removed from $\boldsymbol{\alpha}^\star$. As shown in Theorem 4.12, the placement of zero elements in the sequence *does not affect* the distance function $D(\,\cdot\,)$. Thus, the sequence $\boldsymbol{\alpha}^\star$ achieves the same distance as the sequence $\boldsymbol{\alpha}^{opt}$ created by removing the zero elements of $\boldsymbol{\alpha}^\star$. Moreover, for $\boldsymbol{\alpha}^\star$ to not be non-decreasing in some non-zero sub-sequence, the sequence $\boldsymbol{\alpha}^{opt}$ *must* have at least one pair of adjacent decreasing elements. That is, there exists an index $\tau$ such that the $\tau^{\text{th}}$ and $(\tau+1)^{\text{th}}$ elements decrease: $\alpha_\tau^{opt} > \alpha_{\tau+1}^{opt}$.

We show that, by switching these elements, the distance achieved by the resulting sequence exceeds that of $\boldsymbol{\alpha}^{opt}$; i.e., $\boldsymbol{\alpha}^{opt}$ is not optimal. Formally, we assume that

$$\exists \boldsymbol{\alpha}^{opt} \in \mathcal{A}^{(M,\infty)} \text{ s.t. } \quad \forall \boldsymbol{\alpha} \in \mathcal{A}^{(M,\infty)} \quad D\left(\boldsymbol{\alpha}^{opt}\right) \geq D\left(\boldsymbol{\alpha}\right) \tag{B.3}$$

$$\text{and } \quad \exists \tau \in \mathfrak{N} \quad \text{s.t.} \quad \alpha_\tau^{opt} > \alpha_{\tau+1}^{opt} > 0. \tag{B.4}$$

Now we consider an alternative sequence $\boldsymbol{\alpha}' \in \mathcal{A}^{(M,\infty)}$ that switches the $\tau^{\text{th}}$ and $(\tau+1)^{\text{th}}$ element of $\boldsymbol{\alpha}^{opt}$:

$$\alpha_t' = \begin{cases} \alpha_t^{opt}, & \text{if } t < \tau \\ \alpha_{t+1}^{opt}, & \text{if } t = \tau \\ \alpha_{t-1}^{opt}, & \text{if } t = \tau+1 \\ \alpha_t^{opt}, & \text{if } t > \tau+1 \end{cases}.$$

By design, $\boldsymbol{\alpha}^{opt}$ and $\boldsymbol{\alpha}'$ are $(\tau, 1)$-differing sequences and thus we can compute the difference in their distances by applying Lemma B.1 to yield[1]

$$\begin{aligned} \Delta_{\boldsymbol{\alpha}^{opt},\boldsymbol{\alpha}'} &= \frac{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau'\right) \cdot \alpha_\tau^{opt} \cdot \alpha_{\tau+1}^{opt} - \left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right) \cdot \alpha_\tau' \cdot \alpha_{\tau+1}'}{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right)\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau'\right)\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt} + \alpha_{\tau+1}^{opt}\right)} \\ &= \frac{\left(\alpha_{\tau+1}^{opt} - \alpha_\tau^{opt}\right) \cdot \alpha_\tau^{opt} \cdot \alpha_{\tau+1}^{opt}}{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right)\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_{\tau+1}^{opt}\right)\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt} + \alpha_{\tau+1}^{opt}\right)}, \end{aligned}$$

in which $\mu_t^{(\boldsymbol{\alpha}^{opt})} = N + \sum_{\ell=1}^{t} \alpha_\ell^{opt}$ from Lemma B.1.

[1] Although $\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})}$ may be zero (e.g., if $\tau = 0$), the lemma is applicable since we assumed $\alpha_\tau^{opt} > \alpha_{\tau+1}^{opt} > 0$.

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

The denominator in the above expression is strictly positive since $\alpha_\tau^{opt} > 0$, $\alpha_{\tau+1}^{opt} > 0$ and $\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} \geq 0$. Further, from assumption (B.4), we have that $\alpha_\tau^{opt} > \alpha_{\tau+1}^{opt} > 0$, and hence, the above numerator is strictly less than zero.[2] Thus, we have $\Delta_{\boldsymbol{\alpha}^{opt},\boldsymbol{\alpha}'} = D(\boldsymbol{\alpha}^{opt}) - D(\boldsymbol{\alpha}') < 0$, from which we conclude that $D(\boldsymbol{\alpha}^\star) = D(\boldsymbol{\alpha}^{opt}) < D(\boldsymbol{\alpha}')$. This contradicts assumption (B.3) that $\boldsymbol{\alpha}^\star$ is optimal, thus showing that any sequence with a sub-sequence of its non-zero elements that is not monotonically non-decreasing is non-optimal. Hence, every sub-sequence of the non-zero elements of any optimal attack sequence must be monotonically non-decreasing. □

## B.3 Proof of Theorem 4.15

We show that the optimal distances achieved by attack sequences are strictly monotonically increasing in the attack capacity available to the attacker and the attack duration during which the attack is executed. To do so, we first demonstrate that it is non-optimal to use all attack points during a single retraining iteration unless there is only a single attack point or retraining iteration.

LEMMA B.2 *For $M > 1$ and $T > 1$, any attack $\boldsymbol{\alpha}$ with only a single non-zero element $\tau$ (i.e., such that $\alpha_\tau > 0$ and $\alpha_t = 0$ for all $t \neq \tau$) is a non-optimal sequence.*

*Proof* Any such sequence $\boldsymbol{\alpha}$ described above achieves distance 1 by Equations (4.11) and (4.12). For $\alpha_\tau = 1$, we construct the alternative sequence $\alpha'_1 = 1$ and $\alpha'_2 = 1$, which is in $\mathcal{A}^{(M,T)}$ for $M > 1$ and $T > 1$ and achieves a distance of $\frac{3}{2} > 1$. Thus, this alternative sequence achieves a higher distance than any sequence with a single element that is one and so these are not optimal sequences.

Similarly, for $\alpha_\tau > 1$, we again construct an alternative sequence $\boldsymbol{\alpha}'$ with $\alpha'_1 = \alpha_\tau - 1$ and $\alpha'_2 = 1$, which has the same attack size as $\boldsymbol{\alpha}$ and also has a duration of 2, which places it in $\mathcal{A}^{(M,T)}$ for $M > 1$ and $T > 1$. Further, the alternative sequence achieves a distance of $1 + \frac{1}{\alpha_\tau} > 1$. Thus, we have demonstrated an alternative sequence in this space that achieves a higher distance and so any such sequence with a single non-zero element is not optimal. □

This lemma is one of several results needed for the proof of Theorem 4.15 below. Additionally, it assumes that there is a greatest non-zero element within any sequence of finite attack size, $M$. This is true for all integral sequences, but does not hold for continuously valued sequences as discussed below. We now present the main proof of this section.

*Proof of Theorem 4.15* First we show that, for any fixed $N > 0$, $D_N^\star(M, \infty)$ and $D_N^\star(M, T)$ are strictly monotonically increasing with respect to $M \in \mathfrak{N}_0$; i.e., $\forall\, M^{(1)} < M^{(2)} \in \mathfrak{N}_0$, we claim $D_N^\star\left(M^{(1)}, \infty\right) < D_N^\star\left(M^{(2)}, \infty\right)$ and that, for any fixed $T \in \mathfrak{N}$, $D_N^\star\left(M^{(1)}, T\right) < D_N^\star\left(M^{(2)}, T\right)$. By Definition 4.10 of $D_N^\star(\cdot, \infty)$, there exists a

[2] Notice that, if either $\alpha_\tau^{opt} = 0$ or $\alpha_{\tau+1}^{opt} = 0$, the numerator would be zero, thus giving the two sequences equal distances and making this result consistent with Theorem 4.12.

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

sequence $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M^{(1)},\infty)}$ such that $D(\boldsymbol{\alpha}^\star) = D_N^\star\left(M^{(1)},\infty\right)$. However, we also have $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M^{(2)},\infty)}$; that is, any optimal sequence from $\mathcal{A}^{(M^{(1)},\infty)}$ is also in the space $\mathcal{A}^{(M^{(2)},\infty)}$ but uses at most $M^{(1)} < M^{(2)}$ of its total attack capacity. Moreover, since $\sum_t \alpha_t^\star \le M^{(1)}$ and all sequences consist of elements $\alpha_t^\star \in \mathfrak{N}_0$, there must exist a last non-zero index $\tau \in \mathfrak{N}$ of $\alpha_t^\star$; i.e., for all $t > \tau$, $\alpha_t^\star = 0$. From this, we construct an alternate sequence $\boldsymbol{\alpha}'$, which is identical to $\boldsymbol{\alpha}^\star$ except that we add the excess attack capacity to its last non-zero element: $\alpha'_\tau = \alpha_\tau^\star + m$ where $m = M^{(2)} - M^{(1)} > 0$. The difference in the distances of these two sequences is simply the difference in their final non-zero contributions:

$$\begin{aligned} D(\boldsymbol{\alpha}^\star) - D\left(\boldsymbol{\alpha}'\right) &= \delta_\tau\left(\boldsymbol{\alpha}^\star\right) - \delta_\tau\left(\boldsymbol{\alpha}'\right) \\ &= \frac{\alpha_\tau^\star}{M^{(1)}+N} - \frac{\alpha_\tau^\star + m}{M^{(2)}+N} \\ &= \frac{\alpha_\tau^\star M^{(2)} + \alpha_\tau^\star N - \alpha_\tau^\star M^{(1)} - \alpha_\tau^\star N - mM^{(1)} - mN}{\left(M^{(1)}+N\right)\left(M^{(2)}+N\right)} \\ &= \frac{\left(\alpha_\tau^\star - M^{(1)} - N\right)m}{\left(M^{(1)}+N\right)\left(M^{(2)}+N\right)}, \end{aligned}$$

where $m > 0$ and $M^{(2)} > M^{(1)} \ge 0$. All terms in this ratio are *positive* except the term $\left(\alpha_\tau^\star - M^{(1)} - N\right)$, which is *negative* since $\alpha_\tau^\star \le M^{(1)}$ and $N \ge 1$. Thus, the above difference is negative and $D_N^\star\left(M^{(2)},\infty\right) \ge D(\boldsymbol{\alpha}') > D(\boldsymbol{\alpha}^\star) = D_N^\star\left(M^{(1)},\infty\right)$. This proof also holds for any fixed $T \ge 1$, thus also showing that $D_N^\star\left(M^{(2)},T\right) > D_N^\star\left(M^{(1)},T\right)$.

Second, for $N = 0$, we demonstrate strict monotonicity of $D_0^\star(\cdot,\infty)$ and $D_0^\star(\cdot,T)$ for any $T > 1$. Since $D_0^\star(M,\infty) \ge 0$ for any $M \in \mathfrak{N}_0$, and $D_0^\star(M,\infty) = 0$ if and only if $M = 0$, we have that $D_0^\star(M,\infty) > D_0^\star(0,\infty)$ for any $M > 0$, as required. In the case $M^{(2)} > M^{(1)} = 1$, every sequence in $\mathcal{A}^{(1,\infty)}$ (or $\mathcal{A}^{(1,T)}$) achieves $D_0^\star(1,\infty) = D_0^\star(1,T) = 1$. Further the sequence $(1,1)$ is in $\mathcal{A}^{(M^{(2)},\infty)}$ for all $M^{(2)} > 1$ (and also in $\mathcal{A}^{(M^{(2)},T)}$ for any $T > 1$), and itachieves a distance of $1 + \frac{1}{2}$, thus exceeding $D_0^\star(1,\infty)$. Thus, it again follows that $D_0^\star\left(M^{(2)},\infty\right) > D_0^\star(1,\infty)$ and for any fixed $T > 1$, $D_0^\star\left(M^{(2)},T\right) > D_0^\star(1,T)$. Finally, for $M^{(2)} > M^{(1)} > 1$, we use a similar proof as was used above for $N > 0$. Again, there is a sequence $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M^{(1)},\infty)}$ such that $D(\boldsymbol{\alpha}^\star) = D_0^\star\left(M^{(1)},\infty\right)$ and we take $\tau$ to be index of the last non-zero element of $\boldsymbol{\alpha}^\star$. We again contruct an alternate sequence $\boldsymbol{\alpha}'$ that is identical to $\boldsymbol{\alpha}^\star$ except the excess attack capacity is added to its last non-zero element: $\alpha'_\tau = \alpha_\tau^\star + m$ where $m = M^{(2)} - M^{(1)} > 0$. Then, as above, in examining the difference in the distances of these two sequences, it can be shown that

$$D(\boldsymbol{\alpha}^\star) - D\left(\boldsymbol{\alpha}'\right) = \frac{\left(\alpha_\tau^\star - M^{(1)}\right)m}{M^{(1)} \cdot M^{(2)}},$$

where $m > 0$ and $M^{(2)} > M^{(1)} > 1$. Again, all terms in the fraction are *positive* except the term $\left(\alpha_\tau^\star - M^{(1)}\right)$. But, by Lemma B.2, $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M^{(1)},\infty)}$ for $M^{(1)} > 1$ cannot be optimal unless $\alpha_\tau^\star < M^{(1)}$. Thus the above difference is negative and $D_0^\star\left(M^{(2)},\infty\right) \ge D(\boldsymbol{\alpha}') > D(\boldsymbol{\alpha}^\star) = D_0^\star\left(M^{(1)},\infty\right)$. This proof construction also holds for any fixed duration $T > 1$, thus also showing that $D_0^\star\left(M^{(2)},T\right) > D_0^\star\left(M^{(1)},T\right)$.

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

Thirdly, we show that $D_N^\star(M, T)$ is strictly monotonically increasing with respect to $T \in \{1, \ldots, M\}$; that is, for any fixed $N \in \mathfrak{N}_0$ and $M \in \mathfrak{N}$ and $\forall\, T_1 < T_2 \in \{1, \ldots, M\}$, we claim $D_N^\star(M, T_1) < D_N^\star(M, T_2)$.

For $T_1, T_2 \leq M$, by Definition 4.10 of $D_N^\star(M, T)$, there exists a sequence $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M,T_1)}$ such that $D(\boldsymbol{\alpha}^\star) = D_N^\star(M, T_1)$. However, since $T_2 > T_1$, we also have $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M,T_2)}$; that is, any optimal sequence from $\mathcal{A}^{(M,T_1)}$ is also in the space $\mathcal{A}^{(M,T_2)}$ but has a trailing sequence of zeros: $\alpha^\star_{T_1+1} = \ldots = \alpha^\star_{T_2} = 0$. Alternatively, there is some last index $\tau < T_2$ such that $\alpha^\star_\tau > 0$ and $\alpha^\star_t = 0$ for all $t > \tau$.

In fact, this $\tau^{\text{th}}$ element must be greater than 1 since, by Theorem 4.14, the non-zero elements of $\boldsymbol{\alpha}^\star$ must be non-decreasing. Thus, either $\alpha^\star_\tau > 1$ or all previous elements must be in $\{0, 1\}$. However, since $\tau < T_2 \leq M$, such a sequence can have at most $M - 1$ elements, but the first part of this theorem already showed that such a sequence is not optimal. Hence, $\alpha^\star_\tau > 1$.

Using this fact, we can construct an alternative sequence $\boldsymbol{\alpha}' \in \mathcal{A}^{(M,T_2)}$ that moves one attack point from the $\tau^{\text{th}}$ element of $\boldsymbol{\alpha}^\star$ to its $(\tau+1)^{\text{th}}$ element:

$$\alpha'_t = \begin{cases} \alpha^\star_t, & \text{if } t < \tau \\ \alpha^\star_t - 1, & \text{if } t = \tau \\ 1, & \text{if } t = \tau + 1 \\ \alpha^\star_{t-1}, & \text{if } t > \tau + 1 \end{cases}.$$

By design, $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\alpha}'$ are $(\tau, 1)$-differing sequences and Lemma B.1 yields

$$\begin{aligned} \Delta_{\boldsymbol{\alpha}^\star, \boldsymbol{\alpha}'} &= \frac{\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha'_\tau\right) \cdot \alpha^\star_\tau \cdot \alpha^\star_{\tau+1} - \left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha^\star_\tau\right) \cdot \alpha'_\tau \cdot \alpha'_{\tau+1}}{\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha^\star_\tau\right)\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha'_\tau\right)\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha^\star_\tau + \alpha^\star_{\tau+1}\right)} \\ &= \frac{-1 \cdot \left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha^\star_\tau\right) \cdot \left(\alpha^\star_\tau - 1\right)}{\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha^\star_\tau\right)\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha'_\tau\right)\left(\mu^{(\boldsymbol{\alpha}^\star)}_{\tau-1} + \alpha^\star_\tau + \alpha^\star_{\tau+1}\right)} \end{aligned}$$

in which $\mu_t^{\left(\boldsymbol{\alpha}^{opt}\right)} = N + \sum_{\ell=1}^t \alpha_\ell^{opt}$. This difference is negative, from which we conclude that $D(\boldsymbol{\alpha}^\star) < D(\boldsymbol{\alpha}')$. This contradicts assumption (B.3) that $\boldsymbol{\alpha}^\star$ is optimal in $\mathcal{A}^{(M,T_2)}$; i.e., we have shown there is a sequence in $\mathcal{A}^{(M,T_2)}$ whose distance exceeds $D_N^\star(M, T_1)$. Thus, $D_N^\star(M, T)$ is strictly monotonically increasing for $T \leq M$.

Finally, to see that $D_N^\star(M, T) = D_N^\star(M, \infty)$ for $T \geq M$, any sequence in $\mathcal{A}^{(M,T)}$ must have at least $M - T$ zero elements. As we showed in Theorem 4.12, the distance of a sequence is invariant to the placement of these zero elements so, without loss of generality, we can place them at the end. Thus, any sequence in $\mathcal{A}^{(M,T)}$ achieves the same distance as a sequence in $\mathcal{A}^{(M,M)}$, and the optimal distances achieved within these two spaces is equal. □

Notice that the above argument does not hold for the space $\mathcal{B}^{(M,\infty)}$ of all positive-real-valued sequences with total mass of $M$ since there need not be a greatest non-zero element in such a sequence. However, optimality is not well-defined on such a space. This proof does, however, extend directly to sequences in $\mathcal{B}^{(M,T)}$ since the finite attack duration $T$ implies the existence of a greatest non-zero element.

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

## B.4 Proof of Theorem 4.16

The proof of this theorem is again similar to the proofs in the previous sections.

*Proof* We show that any optimal sequence $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M,\infty)}$ only has elements in the set $\{0, 1\}$. This is trivially satisfied for $\mathbf{0}$, the only sequence in $\mathcal{A}^{(0,\infty)}$, and thus is optimal.

For $M > 0$, the proof proceeds *by contradiction* by assuming that there is an optimal sequence, $\boldsymbol{\alpha}^\star \in \mathcal{A}^{(M,\infty)}$, for which there exists a $\tau$ such that $\alpha_\tau^\star > 1$. We will arrive at a contradiction to the claim that this $\boldsymbol{\alpha}^\star$ achieves an optimal displacement within $\mathcal{A}^{(M,\infty)}$ by instead considering the equivalent (with respect to the distance function) sequence $\boldsymbol{\alpha}^{opt}$, which is identical to $\boldsymbol{\alpha}^\star$ except that a zero is inserted after the $\tau^{\text{th}}$ element and all subsequent elements are shifted to the next index; i.e., $\alpha_\tau^{opt} > 1$ and $\alpha_{\tau+1}^{opt} = 0$. As shown in Lemma 4.11, removing (or inserting) a zero elements in the sequence *does not affect* the distance function $D(\,\cdot\,)$. Thus, the sequence $\boldsymbol{\alpha}^\star$ achieves the same distance as the sequence $\boldsymbol{\alpha}^{opt}$.

We show that there is an alternative sequence, whose distance exceeds that of $\boldsymbol{\alpha}^{opt}$; i.e., $\boldsymbol{\alpha}^{opt}$ is not optimal. Formally, we first assume that

$$\exists \boldsymbol{\alpha}^{opt} \in \mathcal{A}^{(M,\infty)} \text{ s.t. } \quad \forall \boldsymbol{\alpha} \in \mathcal{A}^{(M,\infty)} \quad D\left(\boldsymbol{\alpha}^{opt}\right) \geq D\left(\boldsymbol{\alpha}\right) \tag{B.5}$$

$$\text{and } \quad \exists \tau \in \mathfrak{N} \quad \text{s.t. } \quad \alpha_\tau^{opt} > 1 \ \wedge \ \alpha_{\tau+1}^{opt} = 0. \tag{B.6}$$

Now we consider an alternative sequence $\boldsymbol{\alpha}' \in \mathcal{A}^{(M,\infty)}$ that shifts 1 unit from $\alpha_\tau^{opt}$ to $\alpha_{\tau+1}^{opt}$

$$\alpha_t' = \begin{cases} \alpha_t^{opt}, & \text{if } t < \tau \\ \alpha_t^{opt} - 1, & \text{if } t = \tau \\ 1, & \text{if } t = \tau + 1 \\ \alpha_t^{opt}, & \text{if } t > \tau + 1 \end{cases}.$$

By design, $\boldsymbol{\alpha}^{opt}$ and $\boldsymbol{\alpha}'$ are $(\tau, 1)$-differing sequences and thus we can compute the difference in their distances by applying Lemma B.1 to yield[3]

$$\begin{aligned} \Delta_{\boldsymbol{\alpha}^{opt},\boldsymbol{\alpha}'} &= \frac{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau'\right) \cdot \alpha_\tau^{opt} \cdot \alpha_{\tau+1}^{opt} - \left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right) \cdot \alpha_\tau' \cdot \alpha_{\tau+1}'}{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right)\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau'\right)\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt} + \alpha_{\tau+1}^{opt}\right)} \\ &= \frac{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right) \cdot \left(1 - \alpha_\tau^{opt}\right)}{\left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt}\right)^2 \left(\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} + \alpha_\tau^{opt} - 1\right)}, \end{aligned}$$

in which $\mu_t^{(\boldsymbol{\alpha}^{opt})} = N + \sum_{\ell=1}^{t} \alpha_\ell^{opt}$ from Lemma B.1.

The denominator in the above expression is strictly positive since $\alpha_\tau^{opt} > 1$ and $\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})} \geq 0$. Further, from assumption (B.6), we have that $\alpha_\tau^{opt} > 1$, and hence, the

[3] Although $\mu_{\tau-1}^{(\boldsymbol{\alpha}^{opt})}$ may be zero (e.g., if $\tau = 0$), the lemma is applicable since we assumed $\alpha_\tau^{opt} > 1$.

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

above numerator is negative.[4] Thus, we have $\Delta_{\boldsymbol{\alpha}^{opt},\boldsymbol{\alpha}'} = D(\boldsymbol{\alpha}^{opt}) - D(\boldsymbol{\alpha}') < 0$ , from which we conclude that $D(\boldsymbol{\alpha}^{\star}) = D(\boldsymbol{\alpha}^{opt}) < D(\boldsymbol{\alpha}')$. This contradicts assumption (B.5) that $\boldsymbol{\alpha}^{\star}$ is optimal, thus showing that any sequence with an element not in $\{0, 1\}$ is non-optimal. Hence, any optimal sequence must have $\alpha_t^{\star} \in \{0, 1\}$ for all $t$. Further, an optimal sequence $\boldsymbol{\alpha}^{\star} \in \mathcal{A}^{(M,\infty)}$ must have exactly $M$ ones since, by Theorem 4.15, $D_0^{\star}(M, \infty)$ strictly increases in $M$.

Finally, the optimal displacement can be derived by supposing first that the adversary can control all $M + N$ points, including the initial $N$ points. As we showed above, one optimal sequence is $\mathbf{1}_{M+N}$. It then follows from substituting $\alpha_t = 1$ and $\mu_t = t$ into Equation (4.11) that $D(\mathbf{1}_{M+N}) = \sum_t^{M+N} \frac{1}{t} = h_{M+N}$. Now we subtract the contribution from the first $N$ points, which is $h_N$. This yields the result for $D_N^{\star}(M, \infty)$. □

## B.5 Proof of Theorem 4.18

The final proof we present here is for optimal attacks of a limited duration $T$ using the relaxation to continuous attack sequences introduced in Section 4.4.1. In this continuous domain, we optimize Program (4.16) using optimization techniques.

*Proof* To optimize the objection function of Equation (4.15) in terms of the continuous sequences $\boldsymbol{\mu}$, we first verify that this function is well-behaved for feasible sequences. For all $t$, $\mu_t > 0$ since $\mu_t$ is monotonically non-decreasing in $t$ and $\mu_0 = N > 0$. Any such sequence thus can be characterized as a positive-valued vector; i.e., $\boldsymbol{\mu} \in (0, \infty)^T$. On this domain, the objective function given in Equation (4.15) is continuous as are its first derivatives. Thus, by Theorem 1.2.3 of Peressini, Sullivan & Jerry J. Uhl (1988), the extrema of this function are either its stationary points or lie on the boundary.

First, we eliminate the possibility that an optimum exists at the boundary. Any sequence on the boundary of this domain must have two or more consecutive elements in the total mass sequence that are equal, or rather, in the original formulation, there must be an element $\beta_j = 0$. By Theorem 4.12, such a boundary sequence is equivalent to a sequence of length $T - 1$. However, by Theorem 4.15, the function $D_0^{\star}(M, T)$ is increasing in $T$, unless $T \geq M$. Thus, no boundary point of the total mass formulation is an optimal sequence and so every optimal sequence must be a critical point of the objective.

Second, to find the stationary points of this objective function, we solve for sequences that make its partial derivatives equal zero. For each $\tau \in 1 \ldots T - 1$, the partial

[4] An astute reader may wonder what this analysis implies when $\alpha_\tau^{opt} \in \{0, 1\}$. For $\alpha_\tau^{opt} = 0$, the sequence $\boldsymbol{\alpha}'$ would have have a negative $\tau^{\text{th}}$ element and thus is not in $\mathcal{A}^{(M,\infty)}$. For $\alpha_\tau^{opt} = 1$, the above numerator is zero, but the denominator also may be zero so Lemma B.1 does not always apply. Instead, the alternate sequence $\boldsymbol{\alpha}'$ can be viewed as simply swapping the position of a 1 with a 0 in the sequence. Thus, Theorem 4.12 can be applied to show that $\boldsymbol{\alpha}^{opt}$ and $\boldsymbol{\alpha}'$ have equal distances and no contradiction arises.

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press

derivative with respect to the $\tau^{\text{th}}$ element, $\mu_\tau$, yields the following condition:

$$\frac{\partial}{\partial \mu_\tau}\left[T - \sum_{t=1}^{T} \frac{\mu_{t-1}}{\mu_t}\right] = 0 \quad \Rightarrow \quad \frac{\mu_{\tau-1}}{\mu_\tau^2} = \frac{1}{\mu_{\tau+1}}.$$

These conditions do not hold for $\tau = 0$ or $\tau = T$, but we already have $\mu_0 = N$ and $\mu_T = M + N$. Further, since $\mu_t^\star \in \Re_+$, we can instead consider the logarithm of these variables: $\ell\mu_t \triangleq \log(\mu_t)$), for which any stationary point must satisfy the following system of equations:

$$\begin{aligned} \ell\mu_0 &= \log(N) \\ 2 \cdot \ell\mu_t &= \ell\mu_{t-1} + \ell\mu_{t+1} \quad \forall\, t \in \{1 \dots T-1\} \\ \ell\mu_T &= \log(M+N), \end{aligned} \tag{B.7}$$

The second condition is equivalently defined by the recurrence relation $\ell\mu_t = 2\ell\mu_{t-1} - \ell\mu_{t-2}$, for $t \geq 2$. This recurrence has a characteristic polynomial given by $\chi(r) = r^2 - 2r + 1$. Solving $\chi(r) = 0$ yields the single root $r = 1$, for which there must exist $\phi$ and $\psi$ such that $\ell\mu_t = \phi \cdot r^t + \psi \cdot t \cdot r^t = \phi + \psi \cdot t$. Using the boundary conditions $\ell\mu_0 = \log(N)$ and $\ell\mu_T = \log(M+N)$, we find that $\phi = \log(N)$ and $\psi = \frac{1}{T}\log\left(\frac{M+N}{N}\right)$. Thus, the *unique* solution to this linear recurrence relation is given by

$$\ell\mu_t = \log(N) + \tfrac{t}{T}\log\left(\tfrac{M+N}{N}\right).$$

Naturally, this corresponds to the sequence $\mu_t^\star = N\left(\frac{M+N}{N}\right)^{\left(\frac{t}{T}\right)}$, which satisfies $\mu_0^\star = N$ and $\mu_T^\star = N + M$ and is a non-decreasing sequence. Moreover, the logarithmic conditions given in Equation (B.7) must hold for any optimal positive sequence, but specify a system of $T+1$ equalities in terms of $T+1$ variables and thus have a unique solution. Thus, $\boldsymbol{\mu}^\star$ is the unique positive sequence that maximizes the program given in Equation (4.16).

Having established optimality, the optimal distance achieved is

$$D_N^\star(M,T) = T - \sum_{t=1}^{T} \frac{N}{N}\left(\frac{M+N}{N}\right)^{\left(\frac{t-1}{T}\right)}\left(\frac{M+N}{N}\right)^{\left(\frac{-t}{T}\right)} = T\left(1 - \left(\frac{N}{M+N}\right)^{\frac{1}{T}}\right)$$

as was to be shown. Finally, using the definition of total mass in Equation (4.7), for $t \geq 1$, $\beta_t^\star = \mu_t^\star - \mu_{t-1}^\star = N\left(\frac{M+N}{N}\right)^{\frac{t-1}{T}}\left(\left(\frac{M+N}{N}\right)^{\frac{1}{T}} - 1\right)$. This completes the proof. □

https://doi.org/10.1017/9781107338548.011 Published online by Cambridge University Press