involve

a journal of mathematics

msp

involve

msp.org/involve

Theoretical properties of the length-biased inverse Weibull distribution

Jing Kersey and Broderick O. Oluyede

(Communicated by Kenneth S. Berenhaut)

We investigate the length-biased inverse Weibull (LBIW) distribution, deriving its density function, hazard and reverse hazard functions, and reliability function. The moments, moment-generating function, Fisher information and Shannon entropy are also given. We discuss parameter estimation via the method of moments and maximum likelihood, and hypothesis testing for the LBIW and parent distributions.

1. Introduction

Weighted distributions occur in many areas, including medicine, ecology, reliability, and branching processes. Results and applications in these and other areas can be seen in [Patil and Rao 1978; Gupta and Kirmani 1990; Gupta and Keating 1986; Oluyede 1999]. In a weighted distribution problem, a realization x of X enters into the investigator's record with probability proportional to a weight function $w(x)$. The recorded x is not an observation of X, but rather an observation of a weighted random variable X_w.

In this article we are interested in the case where $w(x) = x$. This is called *length bias*; it approximates situations common in practice (see [Arratia and Goldstein 2009] for an introductory discussion). We will apply length bias to the inverse Weibull distribution (see Section 2 below), which has a wide range of applications in diverse areas such as medicine, reliability and ecology; for example, Keller et al. [1985] found it to be a good fit in their investigation of failures of mechanical components subject to degradation. As a result, the inverse Weibull distribution is well studied; see [Johnson et al. 1994] or [Rinne 2009] for a tabulation of results.

To proceed, we need some standard terminology. If X is a continuous, nonnegative random variable with distribution function F and probability density function (pdf) f (so that $f(u) = dF(u)/du$), we call $\bar{F}(x) = 1 - F(x)$ the associated *reliability function*, from the situation where $\bar{F}(x)$ describes the probability that

some piece of equipment, say, will still be working at time x. The *hazard function* $\lambda_F(x)$ and *mean residual life function* $\delta_F(x)$ are defined by

$$\lambda_F(x) = \frac{f(x)}{\bar{F}(x)} \quad \text{and} \quad \delta_F(x) = \int_x^\infty \frac{\bar{F}(u)}{\bar{F}(x)} \, du. \tag{1}$$

The *reverse hazard function* is $\tau_F(x) = f(x)/F(x)$. When λ_F is monotone increasing, we say that F is an *increasing hazard rate* (IHR) *distribution*. Likewise, a *decreasing mean residual life* (DMRL) *distribution* is one where δ_F is monotone decreasing. It can be shown that IHR implies DMRL. IHR distributions have a number of nice properties, including finiteness of moments of all orders.

Now let $w(x)$, $x \geq 0$, be a positive function, and assume that the expectation of $w(X)$ is positive and finite:

$$0 < E[w(X)] := \int_0^\infty f(x)w(x) \, dx < \infty. \tag{2}$$

We define the weighted random variable X_w by specifying its pdf:

$$f_w(x) = \frac{w(x)f(x)}{E[w(X)]}, \quad x \geq 0. \tag{3}$$

(The denominator ensures that the total mass is 1.)

As mentioned, we will be interested in the case of *length bias*, where $w(x) = x$. In Section 2 we apply this weighting to the inverse Weibull distribution to obtain our main object of study, the LBIW (length-biased inverse Weibull) distribution. We briefly study the shape of the LBIW pdf. In Section 3 we calculate the LBIW moments and moment-generating function, together with the variance, skewness and kurtosis. Section 4 deals with Fisher information and Shannon entropy. In Section 5 we discuss the estimation of the parameters of an LBIW, and describe a test for the detection of length bias. Section 6 showcases a numerical example.

2. The inverse Weibull distribution and its length-biased version

The inverse Weibull distribution function is defined by

$$F(x; x_0, \alpha, \beta) = \exp\left(-(\alpha(x - x_0))^{-\beta}\right), \quad x \geq 0, \ \alpha > 0, \ \beta > 0, \tag{4}$$

where α, x_0 and β are the scale, location and shape parameters, respectively. We will consider only the case $x_0 = 0$, so our distribution function of departure is

$$F(x; \alpha, \beta) = \exp\left(-(\alpha x)^{-\beta}\right), \quad x \geq 0, \ \alpha > 0, \ \beta > 0. \tag{5}$$

(When $\alpha = 1$, this is known as the Fréchet distribution, and its value at $x = 1$ is independent of β; it equals $e^{-1} = 0.3679$, and is known as the characteristic life of

the distribution.) By differentiation we get the corresponding pdf:

$$f(x; \alpha, \beta) = \beta \alpha^{-\beta} x^{-\beta-1} \exp(-(\alpha x)^{-\beta}), \quad x \geq 0, \ \alpha > 0, \ \beta > 0. \qquad (6)$$

To introduce the length bias we first multiply this pdf by the weighting function $w(x) = x$, obtaining

$$\begin{aligned} x f(x; \alpha, \beta) &= \beta \alpha^{-\beta} x^{-\beta} \exp(-(\alpha x)^{-\beta}) \\ &= \beta F(x; \alpha, \beta)(-\log F(x; \alpha, \beta)), \quad x \geq 0, \ \alpha > 0, \ \beta > 0. \qquad (7) \end{aligned}$$

As we saw in (3), we need to divide this function by its integral (2), which is of course the mean of the original distribution, denoted by μ_F. Evaluation yields

$$\mu_F = \frac{\Gamma\left(1 - \frac{1}{\beta}\right)}{\alpha}.$$

Therefore the LBIW (length-biased inverse Weibull) pdf is

$$\begin{aligned} g_w(x; \alpha, \beta) &:= \frac{\alpha}{\Gamma\left(1 - \frac{1}{\beta}\right)} \beta F(x; \alpha, \beta)(-\log F(x; \alpha, \beta)) \\ &= \frac{\beta \alpha^{-\beta+1} x^{-\beta}}{\Gamma\left(1 - \frac{1}{\beta}\right)} \exp(-(\alpha x)^{-\beta}) \quad x \geq 0, \ \alpha > 0, \ \beta > 1. \qquad (8) \end{aligned}$$

(We use the notation g_w instead of f_w as in (3) to make it more distinctive.) The corresponding distribution function is given by

$$G_w(x; \alpha, \beta) = \int_0^x g_w(u; \alpha, \beta) \, du = \frac{1}{\Gamma\left(1 - \frac{1}{\beta}\right)} \int_0^{(\alpha x)^{-\beta}} y^{-1/\beta} \exp(-y) \, dy, \qquad (9)$$

the last equality resulting from rewriting the integral in the variable $y = (\alpha u)^{-\beta}$.

We now turn to the shape of g_w. From (8) we see that $\lim_{x \to 0} g_w(x; \alpha, \beta) = 0$ and $\lim_{x \to \infty} g_w(x; \alpha, \beta) = 0$. Next we look for extrema. It is easier to work with the logarithmic derivative. Since

$$\eta_w(x) := \frac{\partial \log g_w(x; \alpha, \beta)}{\partial x} = \frac{\beta}{x}\left((\alpha x)^{-\beta} - 1\right), \qquad (10)$$

we see that an extremum requires that $(\alpha x)^{-\beta} = 1$. Thus the only extremizer is $x = 1/\alpha$; the pdf increases to a maximum at $1/\alpha$ and then decreases.

For the study of the hazard function it will be useful to consider the *second* derivative of $\log g_w(x; \alpha, \beta)$, namely

$$\eta'_w(x) = -\beta \frac{(\beta + 1)(\alpha x)^{-\beta} - 1}{x^2}. \qquad (11)$$

The numerator on the right has only one zero, at $x = x^* := (\beta + 1)^{1/\beta}/\alpha$, so the

same is true of η'_w. More precisely, we have

$$\begin{aligned}
\eta'_w(x) < 0 &\quad \text{if } x < x^*, \\
\eta'_w(x) = 0 &\quad \text{if } x = x^*, \\
\eta'_w(x) > 0 &\quad \text{if } x > x^*.
\end{aligned} \tag{12}$$

A criterion of Glaser [1980, Theorem on p. 668, case (d)(i), and Lemma on p. 669, case (iii)] then implies that the hazard function is "upside-down bathtub-shaped"; that is, it is initially increasing, reaches a maximum, and decreases thereafter. The conditions of the criterion are that the pdf is twice differentiable and positive for $x > 0$, that it tends to 0 as $x \to 0+$, and that the second derivative of its log satisfies (12) for some x^*. (Note that our η_w differs from Glaser's η by a sign.)

With the qualitative behavior of the hazard function in hand, there remains to write its formula. Recalling the definition in (1), we write

$$\overline{G}_w(x; \alpha, \beta) = \frac{\beta \alpha^{-\beta+1}}{\Gamma\left(1 - \frac{1}{\beta}\right)} \int_x^\infty t^{-\beta} \exp(-(\alpha t)^{-\beta}) \, dt \tag{13}$$

and

$$\lambda_{G_w}(x; \alpha, \beta) = \frac{g_w(x; \alpha, \beta)}{\overline{G}_w(x)} = \frac{x^{-\beta} \exp(-(\alpha x)^{-\beta})}{\int_x^\infty t^{-\beta} \exp(-(\alpha t)^{-\beta}) \, dt}. \tag{14}$$

3. Moments and moment-generating function

In this section we derive the moments, moment-generating function, mean, variance, coefficients of variation, skewness, and kurtosis for the LBIW distribution.

The moments of a length-biased random variable X_w are related to those of the original or parent random variable X by

$$E_{G_w}[X_w^k] = \frac{E_F[X^{k+1}]}{E_F[X]}, \quad k = 1, 2, \ldots, \tag{15}$$

provided $E_F[X^{k+1}]$ exists. Noting that the moments of F are given by

$$E_F[X^k] = \gamma_k := \frac{\Gamma\left(1 - \frac{k}{\beta}\right)}{\alpha^k}, \quad k \geq 1, \ \beta > k, \tag{16}$$

we obtain the moments of X_w as follows:

$$E_{G_w}[X_w^k] = \frac{\Gamma\left(1 - \frac{k+1}{\beta}\right)}{\alpha^k \Gamma\left(1 - \frac{1}{\beta}\right)} = \frac{\gamma_{k+1}}{\gamma_1}, \quad k \geq 1, \ \beta > k. \tag{17}$$

In particular, the mean of X_w is

$$\mu_{G_w} = E_{G_w}[X_w] = \frac{\Gamma\left(1 - \frac{2}{\beta}\right)}{\alpha \Gamma\left(1 - \frac{1}{\beta}\right)} = \frac{\gamma_2}{\gamma_1} \tag{18}$$

and the variance is

$$\sigma_{G_w}^2 = E_{G_w}[X_w^2] - E_{G_w}[X_w]^2 = \frac{\gamma_1\gamma_3 - \gamma_2^2}{\gamma_1^2}, \tag{19}$$

where $\gamma_k = \Gamma(1 - k/\beta)/\alpha^k$. The coefficient of variation (CV) is

$$CV = \frac{\sigma_{G_w}}{\mu_{G_w}} = \sqrt{\frac{\gamma_3\gamma_1}{\gamma_2^2} - 1}. \tag{20}$$

The coefficients of skewness (CS) and kurtosis (CK) are given by

$$CS = \frac{E[(X_w - \mu_{G_w})^3]}{E[(X_w - \mu_{G_w})^2]^{3/2}} = \frac{\gamma_1^2\gamma_4 - 3\gamma_1\gamma_2\gamma_3 + 2\gamma_2^3}{(\gamma_1\gamma_3 - \gamma_2^2)^{3/2}} \tag{21}$$

and

$$CK = \frac{E[(X_w - \mu_{G_w})^4]}{E[(X_w - \mu_{G_w})^2]^2} = \frac{\gamma_1^3\gamma_5 - 4\gamma_1^2\gamma_2\gamma_4 + 6\gamma_1\gamma_2^2\gamma_3 - 3\gamma_2^4}{\gamma_1^2\gamma_3^2 - 2\gamma_1\gamma_2^2\gamma_3 + \gamma_2^4}. \tag{22}$$

The moment-generating function is given by

$$M_{X_w}(t) = \frac{\beta\alpha^{-\beta+1}}{\Gamma(1-\frac{1}{\beta})} \int_0^\infty e^{ty} y^{-\beta} e^{-(\alpha y)^{-\beta}} dy$$

$$= \frac{\beta\alpha^{-\beta+1}}{\Gamma(1-\frac{1}{\beta})} \sum_{j=0}^\infty \frac{t^j}{j!} \int_0^\infty y^{j-\beta} e^{-(\alpha y)^{-\beta}} dy = \frac{\beta\alpha^{-\beta+1}}{\Gamma(1-\frac{1}{\beta})} \sum_{j=0}^\infty \frac{t^j}{j!} \Psi_{j,\alpha,\beta}, \tag{23}$$

where

$$\Psi_{j,\alpha,\beta} = \int_0^\infty y^{j-\beta} e^{-(\alpha y)^{-\beta}} dy.$$

4. Fisher information and Shannon entropy

The information (or Fisher information) that a random variable X contains about the parameter θ is given by

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f(X, \theta)\right)^2\right]. \tag{24}$$

If, in addition, the second derivative with respect to θ of $f(x, \theta)$ exists for all x and θ, and if the second derivative with respect to θ of $\int f(x, \theta) dx = 1$ can be obtained by differentiating twice under the integral sign, then

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(X, \theta)\right]. \tag{25}$$

The Shannon entropy of a random variable X is a measure of the uncertainty and is

given by $E_F[-\log f(X)]$, where $f(x)$ is the pdf of the random variable X.

For the LBIW distribution, the Fisher information that X_w (now renamed X for simplicity) contains about the parameters $\theta = (\alpha, \beta)$ is obtained as follows:

$$
E\left[\left(\frac{\partial \log g_w(X; \alpha, \beta)}{\partial \alpha}\right)^2\right]
$$

$$
= \int_0^\infty \left(\frac{1-\beta}{\alpha} + \beta\alpha^{-\beta-1}x^{-\beta}\right)^2 g_w(x; \alpha, \beta)\, dx
$$

$$
= (1-\beta)^2\alpha^{-2} \int_0^\infty g_w(x; \alpha, \beta)\, dx + \frac{2\beta^2(1-\beta)\alpha^{-2\beta-1}}{\Gamma\left(1-\frac{1}{\beta}\right)} \int_0^\infty x^{-2\beta} e^{-(\alpha x)^{-\beta}}\, dx
$$

$$
+ \frac{\beta^3\alpha^{-3\beta-1}}{\Gamma\left(1-\frac{1}{\beta}\right)} \int_0^\infty x^{-3\beta} e^{-(\alpha x)^{-\beta}}\, dx
$$

$$
= (1-\beta)^2\alpha^{-2} + \frac{2\beta(1-\beta)\alpha^{-2}}{\Gamma\left(1-\frac{1}{\beta}\right)}\Gamma\left(2-\frac{1}{\beta}\right) + \frac{\beta^2\alpha^{-2}}{\Gamma\left(1-\frac{1}{\beta}\right)}\Gamma\left(3-\frac{1}{\beta}\right)
$$

$$
= \beta(\beta-1)\alpha^{-2}, \tag{26}
$$

$$
E\left[\left(\frac{\partial \log g_w(X; \alpha, \beta)}{\partial \beta}\right)^2\right]
$$

$$
= \int_0^\infty \left(\frac{1}{\beta} - \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\beta^2\Gamma\left(1-\frac{1}{\beta}\right)} + \log(\alpha x)\left((\alpha x)^{-\beta} - 1\right)\right)^2 g_w(x; \alpha, \beta)\, dx
$$

$$
= \left(\frac{1}{\beta} - \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\beta^2\Gamma\left(1-\frac{1}{\beta}\right)}\right)^2 - 2\left(\frac{1}{\beta} - \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\beta^2\Gamma\left(1-\frac{1}{\beta}\right)}\right)\frac{\Gamma'\left(2-\frac{1}{\beta}\right) - \Gamma'\left(1-\frac{1}{\beta}\right)}{\beta\Gamma\left(1-\frac{1}{\beta}\right)}
$$

$$
+ \frac{\beta^2\left(\Gamma''\left(3-\frac{1}{\beta}\right) - 2\Gamma''\left(2-\frac{1}{\beta}\right) + \Gamma''\left(1-\frac{1}{\beta}\right)\right)}{\Gamma\left(1-\frac{1}{\beta}\right)}, \tag{27}
$$

$$
E\left[\frac{\partial^2 \log g_w(X; \alpha, \beta)}{\partial\alpha\partial\beta}\right] = E\left[\frac{\partial^2 \log g_w(X; \alpha, \beta)}{\partial\beta\partial\alpha}\right]
$$

$$
= \int_0^\infty \left(\alpha^{-\beta-1}x^{-\beta}(1 - \beta\log\alpha - \beta\log x) - \frac{1}{\alpha}\right) g_w(x; \alpha, \beta)\, dx
$$

$$
= \alpha^{-\beta-1}(1 - \beta\log\alpha) \int_0^\infty x^{-\beta} g_w(x; \alpha, \beta)\, dx
$$

$$
- \frac{\alpha^{-2\beta}\beta^2}{\Gamma\left(1-\frac{1}{\beta}\right)} \int_0^\infty x^{-2\beta}\log x\, e^{-(\alpha x)^{-\beta}}\, dx - \frac{1}{\alpha}\int_0^\infty g_w(x; \alpha, \beta)\, dx
$$

$$
= \alpha^{-1}\beta^{-2}(1-\beta) + \alpha^{-1}\beta^{-3}(\beta-1)\frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\Gamma\left(1-\frac{1}{\beta}\right)} = \frac{\beta-1}{\alpha\beta^3}\left(\frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\Gamma\left(1-\frac{1}{\beta}\right)} - \beta\right). \tag{28}
$$

Thus the information matrix, namely

$$
I(\alpha, \beta) =
\begin{pmatrix}
E\left[\left(\dfrac{\partial \log g_w(X; \alpha, \beta)}{\partial \alpha}\right)^2\right] & E\left[\dfrac{\partial^2 \log g_w(X; \alpha, \beta)}{\partial \alpha\, \partial \beta}\right] \\[2em]
E\left[\dfrac{\partial^2 \log g_w(X; \alpha, \beta)}{\partial \beta\, \partial \alpha}\right] & E\left[\left(\dfrac{\partial \log g_w(X; \alpha, \beta)}{\partial \beta}\right)^2\right]
\end{pmatrix},
\tag{29}
$$

is given by

$$
I(\alpha, \beta) =
\begin{pmatrix}
E\left[\left(\dfrac{\partial \log g_w(X; \alpha, \beta)}{\partial \alpha}\right)^2\right] & \dfrac{\beta-1}{\alpha\beta^3}\left(\dfrac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\Gamma\left(1-\frac{1}{\beta}\right)} - \beta\right) \\[2em]
\dfrac{\beta-1}{\alpha\beta^3}\left(\dfrac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\Gamma\left(1-\frac{1}{\beta}\right)} - \beta\right) & E\left[\left(\dfrac{\partial \log g_w(X; \alpha, \beta)}{\partial \beta}\right)^2\right]
\end{pmatrix},
\tag{30}
$$

where the diagonal entries are stated in (26) and (27).

Note that, for fixed β, the top left entry of this matrix is monotonically decreasing in α, since

$$
\frac{\beta(\beta-1)}{\alpha_1^2} \geq \frac{\beta(\beta-1)}{\alpha_2^2} \iff \alpha_2^2 \geq \alpha_1^2 \iff \alpha_2 \geq \alpha_1.
\tag{31}
$$

On the other hand, for fixed α, the same function is monotonically increasing in β, since

$$
\frac{\beta_1(\beta_1-1)}{\alpha^2} \geq \frac{\beta_2(\beta_2-1)}{\alpha^2} \iff \beta_1(\beta_1-1) \geq \beta_2(\beta_2-1) \iff \beta_1^2 - \beta_2^2 - (\beta_1 - \beta_2) \geq 0
$$

$$
\iff (\beta_1 - \beta_2)(\beta_1 + \beta_2 - 1) \geq 0 \iff \beta_1 \geq \beta_2,
\tag{32}
$$

the last equivalence being a consequence of the inequalities $\beta_1 > 1$, $\beta_2 > 1$.

Under the LBIW distribution, the Shannon entropy is given by

$$
E_G(-\log g_w(X; \alpha; \beta))
$$

$$
= \int_0^\infty \left(-\log \frac{\beta \alpha^{-\beta+1}}{\Gamma\left(1-\frac{1}{\beta}\right)} + \beta \log x + (\alpha x)^{-\beta}\right) g_w(x; \alpha, \beta)\, dx
$$

$$
= -\log \frac{\beta \alpha^{-\beta+1}}{\Gamma\left(1-\frac{1}{\beta}\right)} + \beta \int_0^\infty (\log x) g_w(x; \alpha, \beta)\, dx + \int_0^\infty (\alpha x)^{-\beta} g_w(x; \alpha, \beta)\, dx
$$

$$
= -\log \frac{\beta \alpha^{-\beta+1}}{\Gamma\left(1-\frac{1}{\beta}\right)} + \beta\left(-\log \alpha - \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\beta \Gamma\left(1-\frac{1}{\beta}\right)}\right) + \frac{\beta-1}{\beta}
$$

$$
= \log \frac{\Gamma\left(1-\frac{1}{\beta}\right)}{\alpha\beta} - \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\Gamma\left(1-\frac{1}{\beta}\right)} + \frac{\beta-1}{\beta}.
\tag{33}
$$

5. Estimation of parameters

In this section we derive formulas to estimate the parameters α and β for an unknown
LBIW distribution. We also present a test for the detection of length bias in a sample.

(For the inverse Weibull parent distribution, Calabria and Pulcini [1990; 1994]
derived maximum likelihood, least squares and Bayes estimates for the parameters.
They also obtained confidence limits for reliability and tolerance limits for the same
distribution [Calabria and Pulcini 1989].)

We continue to use X for the LBIW random variable whose parameters α and β
we wish to estimate. We use two standard methods to obtain the estimators: the
method of moments and maximum likelihood.

Method of moments estimators. The method of moments with two parameters
involves setting the first two moments $E[X]$ and $E[X^2]$ equal to the corresponding
moments of an independent sample $X_1, X_2, \ldots, X_n$ of the LBIW random variable.
In view of (18) and (19), this leads to the equations

$$\frac{\Gamma\left(1-\frac{2}{\beta}\right)}{\alpha\Gamma\left(1-\frac{1}{\beta}\right)} = \frac{1}{n}\sum_{j=1}^{n} X_j \quad \text{and} \quad \frac{\Gamma\left(1-\frac{3}{\beta}\right)}{\alpha^2\Gamma\left(1-\frac{1}{\beta}\right)} = \frac{1}{n}\sum_{j=1}^{n} X_j^2. \tag{34}$$

These equations are then solved (numerically, for example) for α and β, leading to
the estimators $\hat{\alpha}$ and $\hat{\beta}$.

If β is known, we only need the first equation in (34). In that case (i.e., for fixed
$\beta > 1$), the method of moments estimate (MME) of α is given by

$$\hat{\alpha} = \frac{n}{\sum_{j=1}^{n} X_j} \frac{\Gamma\left(1-\frac{2}{\beta}\right)}{\Gamma\left(1-\frac{1}{\beta}\right)}. \tag{35}$$

Maximum likelihood estimators. In this method we take the log-likelihood func-
tion of the distribution, take its partial derivatives with respect to the parameters, and
equate their expectations to 0. The log-likelihood function for a single observation
x of X is

$$l(\alpha, \beta) = \log\left(\frac{\beta\alpha^{-\beta+1}}{\Gamma\left(1-\frac{1}{\beta}\right)} x^{-\beta} \exp(-(\alpha x)^{-\beta})\right)$$
$$= \log\beta - (\beta-1)\log\alpha - \beta\log x - (\alpha x)^{-\beta} - \log\Gamma\left(1-\frac{1}{\beta}\right), \tag{36}$$

which leads to

$$\frac{\partial l}{\partial \alpha} = -\frac{\beta-1}{\alpha} + \frac{\beta(\alpha x)^{-\beta}}{\alpha}, \tag{37}$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\beta} - \log\alpha - \log x + (\alpha x)^{-\beta}\log(\alpha x) + \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\beta^2\Gamma\left(1-\frac{1}{\beta}\right)}. \tag{38}$$

From $E[\partial l/\partial \alpha] = 0$, we obtain

$$E[X^{-\beta}] = \frac{\alpha^{\beta}(\beta - 1)}{\beta}, \tag{39}$$

and from $E[\partial l/\partial \beta] = 0$, we have

$$E[-\log X + (\alpha X)^{-\beta}\log(\alpha X)] = \log \alpha - \frac{1}{\beta} - \frac{\Gamma'\left(1-\frac{1}{\beta}\right)}{\beta^2 \Gamma\left(1-\frac{1}{\beta}\right)}. \tag{40}$$

The full log-likelihood function is given by

$$L(\alpha, \beta) = n \log \beta - n(\beta - 1)\log \alpha - \beta \sum_{j=1}^{n} \log x_j - \sum_{j=1}^{n} (\alpha x_j)^{-\beta} - n \log \Gamma\left(1-\tfrac{1}{\beta}\right).$$

The normal equations are

$$\frac{\partial L(\alpha, \beta)}{\partial \alpha} = \frac{-n(\hat{\beta} - 1)}{\hat{\alpha}} + \hat{\beta}\hat{\alpha}^{-\hat{\beta}-1}\sum_{j=1}^{n} x_j^{-\hat{\beta}} = 0, \tag{41}$$

$$\frac{\partial L(\alpha, \beta)}{\partial \beta} = \frac{n}{\hat{\beta}} - n\log\hat{\alpha} - \sum_{j=1}^{n}\log x_j - \sum_{j=1}^{n}\frac{\log(\hat{\alpha}x_j)}{(\hat{\alpha}x_j)^{\hat{\beta}}} - \frac{n}{\hat{\beta}^2}\Psi(1-1/\hat{\beta}) = 0. \tag{42}$$

From (41), the MLE of α is

$$\hat{\alpha} = \left(\frac{n(\hat{\beta} - 1)}{\hat{\beta}\sum_{j=1}^{n} x_j^{-\hat{\beta}}}\right)^{-1/\hat{\beta}}. \tag{43}$$

Now replace $\hat{\alpha}$ in (42) to obtain

$$\frac{\partial L(\alpha, \beta)}{\partial \beta}\bigg|_{\hat{\alpha}, \hat{\beta}} = \frac{n}{\hat{\beta}} - n\log\left(\frac{n(\hat{\beta}-1)}{\hat{\beta}\sum_{j=1}^{n} x_j^{-\hat{\beta}}}\right)^{-1/\hat{\beta}} - \sum_{j=1}^{n}\log x_j$$

$$- \sum_{j=1}^{n}\left(\left(\frac{n(\hat{\beta}-1)}{\hat{\beta}\sum_{j=1}^{n} x_j^{-\hat{\beta}}}\right)^{-1/\hat{\beta}} x_j\right)^{-\hat{\beta}}\log\left(\left(\frac{n(\hat{\beta}-1)}{\hat{\beta}\sum_{j=1}^{n} x_j^{-\hat{\beta}}}\right)^{-1/\hat{\beta}} x_j\right)$$

$$- \frac{1}{\hat{\beta}^2}\sum_{j=1}^{n}\frac{\Gamma'(1-1/\hat{\beta})}{\Gamma(1-1/\hat{\beta})} = 0. \tag{44}$$

This equation does not have a closed form solution and must be solved iteratively to obtain the MLE of the scale parameter β. When α is unknown and β is known, the MLE of α is obtained from (41) with the value of β in place of $\hat{\beta}$. When both α and β are unknown the MLEs of α and β are obtained by solving the normal

equations in (41) and (42). The MLEs of the reliability and hazard functions can be obtained by replacing α and β by their MLEs $\hat{\alpha}$ and $\hat{\beta}$.

The expectations in the Fisher information matrix (FIM) can be obtained numerically. Under the conditions that the parameters are in the interior of the parameter space, but not on the boundary, we have

$$\sqrt{n}\begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I^{-1}(\alpha, \beta)\right) \quad \text{as } n \to \infty,$$

where $I(\alpha, \beta) = \lim_{n \to \infty} n^{-1} I_n(\alpha, \beta)$ and

$$I_n(\alpha, \beta) = n \begin{pmatrix} I(1, 1) & I(1, 2) \\ I(2, 1) & I(2, 2) \end{pmatrix}.$$

The entries $I(i, j)$, $i = 1, 2$ and $j = 1, 2$, are given in (30). The multivariate normal distribution with mean vector $(0, 0)^T$ and covariance matrix $I_n(\alpha, \beta)$ can be used to construct confidence intervals for the model parameters.

Test for generalized length bias. We now seek to discriminate whether a random variable, represented by a random sample of size n, is likely to be the result of length-biased sampling. More precisely, we compare the *null hypothesis H_0*, to the effect that the random variable has the inverse Weibull pdf (6) with given α and β, to the *alternative hypothesis H_c*, which says that the random variable is LBIW ($c = 1$) or perhaps inverse Weibull with some other power weighting $w(x) = x^c$. In this context it's natural to allow this extra generality (and in our particular case this doesn't demand much extra effort). A calculation similar to the one leading to (8) shows that the pdf under the alternative hypothesis is

$$g_w(x; \alpha, \beta, c) = \frac{\beta \alpha^{c-\beta}}{\Gamma(1 - c/\beta)} x^{c-\beta-1} \exp(-(\alpha x)^{-\beta}),$$
$$x \geq 0,\ \alpha > 0,\ \beta > 0,\ c > 0. \quad (45)$$

To decide whether it's plausible that our random sample $x_1, \ldots, x_n$ represents the parent inverse Weibull distribution (null hypothesis H_0) relative to the weighted inverse Weibull distribution (alternative hypothesis H_c), we use the following test statistic, where α and β are assumed known and c is also fixed (several values can be tried, including $c = 1$ for the LBIW):

$$\Lambda = \prod_{i=1}^{n} \frac{g_w(x_i; \alpha, \beta, c)}{f(x_i; \alpha, \beta)} = \prod_{i=1}^{n} \frac{\dfrac{\beta \alpha^{c-\beta}}{\Gamma(1 - c/\beta)} x_i^{c-\beta-1} \exp(-(\alpha x_i)^{-\beta})}{\beta \alpha^{-\beta} x_i^{-\beta-1} \exp(-(\alpha x_i)^{-\beta})}$$

$$= \prod_{i=1}^{n} \frac{\alpha^c x_i^c}{\Gamma\left(1 - \frac{1}{\beta}\right)} = \frac{\alpha^{nc} \prod_{i=1}^{n} x_i^c}{\left(\Gamma\left(1 - \frac{1}{\beta}\right)\right)^n}. \quad (46)$$

We reject H_0 when

$$\Lambda = \frac{\alpha^{nc} \prod_{i=1}^{n} x_i^c}{\left(\Gamma\left(1-\frac{1}{\beta}\right)\right)^n} > K, \tag{47}$$

where $K > 0$ is some threshold chosen beforehand, indicating the level of confidence we want to have in our prediction. Equivalently, we reject the null hypothesis when

$$\Lambda^* = \prod_{i=1}^{n} x_i^c > K^*, \quad \text{where } K^* = \frac{K \Gamma(1 - c/\beta)^n}{\alpha^{nc}} > 0. \tag{48}$$

The choice of K is related to the *p-value*, defined as the probability that, under H_0, the expected value of the test statistic Λ^* is at least as high as the one actually observed. For large n we have $2 \log \Lambda^* \sim \chi^2$, and from the χ^2 one obtains the *p*-value using the χ^2 table (or software). The *p*-value can also be readily computed via Monte Carlo simulation: simulate N samples from the distribution under H_0, for some large value of N, and compute the test statistic Λ_i^* for each sample. Then take

$$p\text{-value} = \frac{\#\{i : \Lambda_i^* > \Lambda^*\}}{N}.$$

Reject the null hypothesis if the *p*-value is less than the desired level of significance (typically 5% or 1%).

6. Examples

In this section we apply the formulas obtained in the previous section to two examples from the literature. The first set of data, given in Table 1, represents the waiting times (in minutes) before service of 100 bank customers [Ghitany et al. 2008]. The second data set, shown in Table 2, represents the number of millions of revolutions before failure of each of 23 ball bearings in a life testing experiment [Lawless 2003].

We modeled these data sets using the weighted inverse Weibull distribution with unknown parameters α and β (we keep the assumption made after (4) that $x_0 = 0$). The normal equations were solved by numerical methods to estimate the model parameters. Specifically, the MLEs of the parameters were computed by maximizing the objective function with the trust-region algorithm in the NLPTR subroutine in SAS. We present in Table 3 the estimated values of the parameters α and β and corresponding gradient objective functions (normal equations) under the length-biased inverse Weibull distribution for both sets of data.

We also conducted, for each set of data, a test for the detection of length bias, to compare the hypothesis that the waiting time distribution follows the LBIW distribution is to be preferred to the null hypothesis that the distribution is unweighted inverse Weibull.

0.8	0.8	4.3	5.0	6.7	8.2	9.7	11.9	14.1	19.9
0.8	0.8	4.3	5.3	6.9	8.6	9.8	12.4	15.4	20.6
1.3	1.3	4.4	5.5	7.1	8.6	10.7	12.5	15.4	21.3
1.5	1.5	4.4	5.7	7.1	8.6	10.9	12.9	17.3	21.4
1.8	1.8	4.6	5.7	7.1	8.8	11.0	13.0	17.3	21.9
1.9	1.9	4.7	6.1	7.1	8.8	11.0	13.1	18.1	23.0
1.9	1.9	4.7	6.2	7.4	8.9	11.1	13.3	18.2	27.0
2.1	2.1	4.8	6.2	7.6	8.9	11.2	13.6	18.4	31.6
2.6	2.6	4.9	6.2	7.7	9.5	11.2	13.7	18.9	33.1
2.7	2.7	4.9	6.3	8.0	9.6	11.5	13.9	19.0	38.5

Table 1. Waiting times of 100 bank customers, from [Ghitany et al. 2008].

17.88	28.92	33.00	41.52	42.12	45.60	48.80	51.84	51.96	54.12
55.56	67.80	68.64	68.64	68.88	84.12	93.12	98.64	105.12	105.84
127.92	128.04	173.40	–	–	–	–	–	–	–

Table 2. Lifetimes of 23 ball bearings, from [Lawless 2003].

Data	α	β	$\partial L/\partial\alpha$	$\partial L/\partial\beta$
I ($n = 100$)	0.400	1.819	7.46×10^{-4}	-5.98×10^{-5}
II ($n = 23$)	0.02795	2.4610	1.990×10^{-9}	1.930×10^{-11}

Table 3. Estimated values of the parameters.

For the set of waiting times given in Table 1, where (as shown in Table 3) the estimated values of the parameters α and β are $\hat{\alpha} = 0.3997$ and $\hat{\beta} = 1.81887$, we obtained for the test statistic the value $2 \log \Lambda = 270.927$, and the p-value for the test was less than 0.000001. Therefore, we have strong statistical evidence that the hypothesis that the waiting time distribution follows the LBIW distribution is to be preferred to the null hypothesis.

For the second set of data, the estimated values of the parameters are $\hat{\alpha} = 0.027952$ and $\hat{\beta} = 2.46097$. The value of the test statistic is $2 \log \Lambda = 170.893$, and the p-value is less than 0.00001. Again, the null hypothesis corresponding to the parent distribution is rejected.

References

[Arratia and Goldstein 2009] R. Arratia and L. Goldstein, "Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent?", 2009, Available at http://bcf.usc.edu/~larry/papers/pdf/csb.pdf.

[Calabria and Pulcini 1989] R. Calabria and G. Pulcini, "Confidence limits for reliability and tolerance limits in the inverse Weibull distribution", *Reliability Engineering and System Safety* **24**:1 (1989), 77–85.

[Calabria and Pulcini 1990] R. Calabria and G. Pulcini, "On the maximum likelihood and least-squares estimation in the inverse Weibull distribution", *Statistica Applicata* **2**:1 (1990), 53–66.

[Calabria and Pulcini 1994] R. Calabria and G. Pulcini, "Bayes 2-sample prediction for the inverse Weibull distribution", *Comm. Statist. Theory Methods* **23**:6 (1994), 1811–1824. MR 1281239 Zbl 0825.62167

[Ghitany et al. 2008] M. E. Ghitany, B. Atieh, and S. Nadarajah, "Lindley distribution and its application", *Math. Comput. Simulation* **78**:4 (2008), 493–506. MR 2009m:62040 Zbl 1140.62012

[Glaser 1980] R. E. Glaser, "Bathtub and related failure rate characterizations", *J. Amer. Statist. Assoc.* **75**:371 (1980), 667–672. MR 83b:62194 Zbl 0497.62017

[Gupta and Keating 1986] R. C. Gupta and J. P. Keating, "Relations for reliability measures under length biased sampling", *Scand. J. Statist.* **13**:1 (1986), 49–56. MR 87h:62173 Zbl 0627.62098

[Gupta and Kirmani 1990] R. C. Gupta and S. N. U. A. Kirmani, "The role of weighted distributions in stochastic modeling", *Comm. Statist. Theory Methods* **19**:9 (1990), 3147–3162. MR 92e:62180 Zbl 0734.62093

[Johnson et al. 1994] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions, 1*, 2nd ed., Wiley, New York, 1994. MR 96j:62028 Zbl 0811.62001

[Keller et al. 1985] A. Z. Keller, M. T. Goblin, and N. R. Farnworth, "Reliability analysis of commercial vehicle engines", *Reliability Engineering* **10**:1 (1985), 15–25.

[Lawless 2003] J. F. Lawless, *Statistical models and methods for lifetime data*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2003. MR 2003i:62001 Zbl 1015.62093

[Oluyede 1999] B. O. Oluyede, "On inequalities and selection of experiments for length biased distributions", *Probab. Engrg. Inform. Sci.* **13**:2 (1999), 169–185. MR 2000a:62266 Zbl 0960.62115

[Patil and Rao 1978] G. P. Patil and C. R. Rao, "Weighted distributions and size-biased sampling with applications to wildlife populations and human families", *Biometrics* **34**:2 (1978), 179–189. MR 80e:62010 Zbl 0384.62014

[Rinne 2009] H. Rinne, *The Weibull distribution: a handbook*, CRC Press, Boca Raton, FL, 2009. MR 2010c:62001 Zbl 05509748

jkersey@ega.edu   Division of Mathematics and Science,
                               East Georgia State College, 10449 US Highway 301 South,
                               Statesboro, GA 30458, United States

boluyede@georgiasouthern.edu   Department of Mathematical Sciences,
                               Georgia Southern University, 65 Georgia Avenue,
                               Statesboro, GA 30460, United States

The firefighter problem for
regular infinite directed grids

Daniel P. Biebighauser, Lise E. Holte and Ryan M. Wagner

(Communicated by Ann Trenk)

We investigate the firefighter problem for regular infinite directed grids. We provide a complete classification of these grids by dividing them into two categories: grids where a single outbreak of fire can be contained with one firefighter per time step and grids that require a second firefighter at some time step. We then investigate infinite directed grids where the degrees of a single vertex are different from the degrees of all other vertices in the grid.

1. Introduction

The firefighter problem was introduced by Bert Hartnell at a conference talk [1995]. A fire breaks out at one or more vertices of a graph G at time zero. At each subsequent time step, one or more defenders are placed on nonburning and undefended vertices, and then the fire spreads from each burning vertex to all of its undefended neighbors. Once a vertex is burning or defended, it remains in that state for the duration of the problem. In particular, firefighters cannot move. The goal is to place firefighters in a way that achieves a desired optimal result, such as containing the fire in as few time steps as possible or minimizing the total number of burned vertices. For a comprehensive introduction to the problem, see [Finbow and MacGillivray 2009].

Question 26 in this last reference suggests investigating the firefighter problem for directed graphs. In this paper, we study infinite directed grids. An *infinite grid* is the graph with vertex set $\mathbb{Z} \times \mathbb{Z}$ where (x_1, y_1) is adjacent to (x_2, y_2) if and only if $|x_1 - x_2| + |y_1 - y_2| = 1$. We consider the firefighter problem on *regular infinite directed grids*, which are infinite grids where a direction is assigned to each edge in such a way that every vertex has in-degree two and out-degree two. We will always consider our grids to be embedded in the plane such that each vertex (x, y) is on the lattice point (x, y).

In this paper, we are concerned with the number of firefighters needed at each time step to eventually contain a fire that starts at a single vertex in a regular infinite directed grid. By "contain," we mean that there is some time step where no new vertices are burned. We are not necessarily interested in containing the fire as soon as possible, but in determining the minimum number of firefighters per time step needed for containment.

Fogarty [2003] proved that two firefighters per time step is necessary and sufficient to contain any finite outbreak of fire in an infinite grid. (Wang and Moeller [2002] had proved earlier that this number was necessary and sufficient for a single vertex initially on fire.) This number is sufficient for infinite directed grids, since directions on the arcs potentially restrict the movement of the fire. If there is an arc joining a burning vertex to an undefended vertex, the fire will spread to the undefended vertex on the next time step (as it would in an undirected graph), but if the arc points in the opposite direction, the fire will not spread along that arc. We will prove that, for regular infinite directed grids with a single vertex initially on fire, we can always contain the fire with fewer defenders.

Our main result is the following theorem, which we prove in Section 3. Without loss of generality, assume that the fire begins at the origin. We will say that an infinite directed grid is a *category A grid* if one firefighter per time step is sufficient to contain the fire. An infinite directed grid is a *category B grid* if one firefighter per time step is not enough to contain the fire, but one firefighter per time step and a second firefighter at any single time step is sufficient to contain the fire.

Theorem 1.1. *Let G be a regular infinite directed grid. Then G is either a category A or a category B grid.*

At the end of this paper, we consider infinite directed grids where at least one vertex has degrees other than in-degree two and out-degree two.

2. A lemma

Fogarty [2003] introduced a theorem with a "Hall-type condition" that is useful for proving that a certain number of defenders per time step is not enough to contain an outbreak of fire in an infinite graph. Her applications of this theorem were mostly to two-dimensional grids. Hartke [2004] extended Fogarty's result using a more general Hall-type condition that allowed him to make stronger statements about infinite grids in higher dimensions. We will use a modified version of Fogarty's theorem that applies to directed graphs. The proof is nearly identical to Fogarty's original proof and will not be included here.

Let G be a directed graph. Assume that one vertex catches on fire at time $t = 0$. Let D_k denote the set of vertices of distance k from the original burned vertex, where the distance from v to w is the length of a shortest directed path from v to w.

Let $B_k \subseteq D_k$ be the set of vertices in D_k that have been burned after time k. Let f_k denote the number of new firefighters available at time step k. Let r_k be the number of firefighters in D_{k+1}, D_{k+2}, ... after time k. We call these *reserve firefighters*. Let $N(S)$ be the neighborhood of a set of vertices S, that is, the set of vertices which are distance 1 from any vertex in S in the underlying undirected graph of G. For any subset $A \subseteq D_k$ let $N^+(A) = N(A) \cap D_{k+1}$.

Theorem 2.1. *Let G be a directed graph. For each k, if every $A \subseteq D_k$ satisfies $|N^+(A)| \geq |A| + f_k$, then $|B_n| \geq 1 + r_n$ for all n.*

We will now apply this theorem to prove the following lemma. An *infinite quarter-plane* is the subgraph of the infinite grid that includes all of the vertices and edges in the first quadrant, including the origin and the positive x- and y-axes.

Lemma 2.2. *Consider an infinite directed quarter-plane where all horizontal arcs point right and all vertical arcs point up. If the fire starts at the origin, one firefighter per time step is not enough to contain the fire. If we are given at least one firefighter per time step, and a second firefighter at any time step, then the fire can be contained.*

Proof. We first prove that one firefighter per time step is not enough to contain the fire. For each k, if $A \subseteq D_k$, we can see that

$$|N^+(A)| \geq |A| + 1,$$

since each vertex in D_k has exactly two neighbors in D_{k+1} and any two vertices in D_k can share at most one neighbor in D_{k+1}. So from Theorem 2.1, since the origin is initially on fire, for every k, we have $|B_k| \geq 1$. Thus one firefighter per time step is not enough to contain the fire.

If we are given at least one firefighter per time step, we can force the fire along an axis of the grid until we get a second firefighter, at which point the fire can be contained by placing this defender on the axis directly ahead of the fire. $\square$

In terms of our categories, the grid in Lemma 2.2 is an example of a category B grid.

3. Regular infinite directed grids

We now prove Theorem 1.1 for regular infinite directed grids.

There are two cases that we must consider. First is the case in which the origin has two consecutive arcs (in cyclic order) facing out. The second case is where the two arcs facing out point in opposite directions. Without loss of generality, in the first case we can assume that the two arcs coming from the origin point along the positive x-axis and the positive y-axis and in the second case they point along the positive and negative y-axes.

The following theorem proves Theorem 1.1 for the first case.

Theorem 3.1. *Let G be a regular infinite directed grid where the two arcs coming from the origin point along the positive x-axis and the positive y-axis. If each arc in the first quadrant (including the axes) is facing either right or up, then G is a category B grid. If at least one arc in the first quadrant (including the axes) faces down or left, then G is a category A grid unless it is the exception shown in Figure 1, which is a category B grid.*

Proof. Suppose first that each arc in the first quadrant (including the axes) faces either right or up. Then the fire cannot leave this quadrant, and by Lemma 2.2, one firefighter per time step is not enough to contain the fire, but a second firefighter at any time step will allow us to contain the fire.

In most of the grids in the rest of this proof, we will show that we can contain the fire with one firefighter per time step. Our strategy will often be to "steer" the fire into a directed cycle. Then we can place defenders on outward neighbors not in the cycle until the fire returns to the first vertex in the cycle. In this way, we can always contain a fire once it reaches a directed cycle.

From now on, suppose that at least one arc in the first quadrant (including the axes) faces either left or down. Consider a closest arc in the first quadrant to the origin (where the distance is measured in the undirected grid from the origin to the head of the arc) that faces either left or down. Call such an arc e. The vertex, v, incident to the head of e must be on an axis, because, if it is not, then since v has in-degree two and out-degree two, at least one of the arcs coming from v must face down or left, and the vertex incident to the head of this arc must be closer to the origin than v was, contrary to our definitions of v and e.
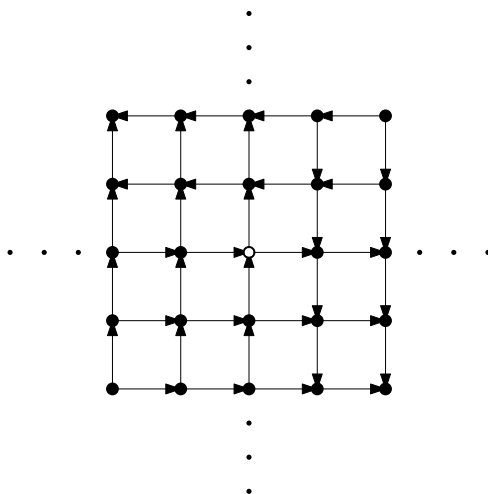
Figure 1. The exception to Theorem 3.1.
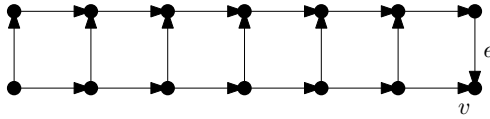
Figure 2. The grid when e is facing down.

Assume that the arc between $(0, 1)$ and $(1, 1)$ points right or the arc between $(1, 0)$ and $(1, 1)$ points up (or both). Thus at least one of these arcs could not be chosen as e. The special case when both of these arcs point toward the axes will be considered at the end of the proof.

Without loss of generality, assume v is on the positive x-axis. If there are multiple edges which could have v as their heads and could be considered to be e, we will break the tie by choosing the vertical arc that is pointing down. There are two cases that we must now consider. First, we will consider the case where e is facing down onto the axis. Since e is the closest arc facing down or to the left, all arcs closer to the origin than e must be facing either up or to the right. A picture of what this grid must look like is shown in Figure 2.

Let v_1 be the vertex directly above v, and label the vertices besides v and v_1 in the four faces containing v in the planar embedding of the grid in a clockwise cycle around v with v_2 through v_8 (so that v_8 is directly to the left of v_1). Let v_9 be the vertex directly above v_8, and let v_{10} be the vertex directly above v_1. See Figure 3.

We will defend along the line $y = 1$, forcing the fire to continue spreading to the right until the fire is at v_7, which is directly to the left of v. (It is possible for v_7 to be the origin.) Then we will defend on v, forcing the fire to spread up to v_8, and then defend on v_9 in order to force the fire to spread to v_1. Now we will defend on either v_{10} or v_2, whichever is incident with an arc coming from v_1. Then the fire could only spread to v, but since v is already defended, the fire is contained.

Next we will consider the case where e is horizontal on the x-axis, facing left toward v. See Figure 4. By our choice of e, the arc between v and v_1 must come from v (if not, we would have chosen this arc as e by our tie-breaking procedure). Also, the arc between v_1 and v_2 must come from v_1, because otherwise there would
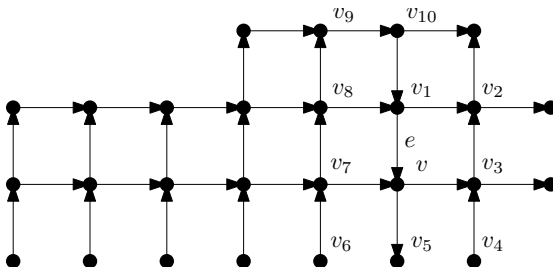
Figure 3. The vertices near v when e is facing down.
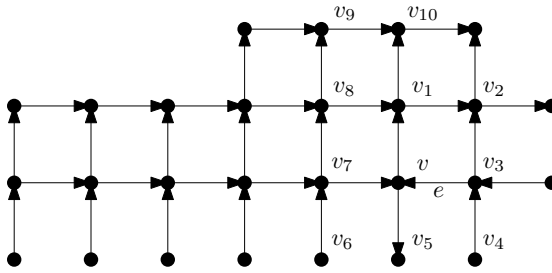
Figure 4. The vertices near v when e is facing left.

be either a downward arc with its head on the positive x-axis to the left of v, or there would be a path of leftward arcs from v_1 that would necessarily follow the line $y = 1$ to the positive y-axis at the point $(1, 0)$. In either case, this contradicts our choices of v and e. (We are still assuming that at least one of the two arcs between $(1, 1)$ and the axes points away from the axis the arc touches.) There are now two subcases that must be considered. The subcases are the arc between v_2 and v_3 facing up or facing down.

If the arc faces down, then there is a directed cycle from v to v_1 to v_2 to v_3 and back to v. We will defend along the line $y = 1$ until the fire spreads to v. Next we defend the outward neighbors of v, v_1, v_2, and v_3 that are not in the cycle until the fire returns to v, at which point we have contained the fire.

If the arc between v_2 and v_3 faces up, then since every vertex has in-degree two and out-degree two, the arc between v_3 and v_4 must face up, the arc between v and v_5 must face down, and the arc between v_6 and v_7 must face up because of our choice of v. The arc between v_4 and v_5 can face either left or right. If it faces right, then there is a directed cycle from v to v_5 to v_4 to v_3 and back to v. If it faces left, then the arc between v_5 and v_6 must also face left because v_5 must have out-degree two. This gives a directed cycle from v_7 to v to v_5 to v_6 and back to v_7. In either case, we can defend along $y = 1$ until the fire reaches v or v_7, respectively, and then contain the fire once it enters the directed cycle.

We are now left with the case where the arc between $(0, 1)$ and $(1, 1)$ points left and the arc between $(1, 0)$ and $(1, 1)$ points down. We will show that any grid in this case can be defended with one firefighter per time step except for the exception in Figure 1. We now have three subcases to consider. The first subcase is when at least one arc on the positive x-axis is pointing left or at least one arc on the positive y-axis is pointing down. The second subcase is when all arcs on the positive x-axis face right, all arcs on the positive y-axis face up, and at least one arc in the first quadrant (not including the axes) faces up or right. The third subcase is when all arcs on the positive x-axis face right, all arcs on the positive y-axis face up, and no arcs in the first quadrant (not including the axes) face up or right.

For the first subcase, consider a nearest arc on a positive axis facing the origin. Without loss of generality, assume it is on the x-axis. Let the vertex at the head of this arc be called u. All arcs to the left of u on the positive x-axis must point right. The arc directly above u must point up. The arc between $(2, 1)$ and $(1, 1)$ must point left. If the arc between $(2, 0)$ and $(2, 1)$ points up, then we have a directed cycle and we can contain the fire. If not, then the arc between $(3, 1)$ and $(2, 1)$ must face left. Again, if the arc between $(3, 0)$ and $(3, 1)$ points up, we have a directed cycle. The only way we might not be able to contain the fire is if all arcs face down from the line $y = 1$ to the positive x-axis. However, as we said, the arc directly above u must face up. Therefore, we will have a directed cycle at or before u, and we can defend the fire so that it spreads along the positive x-axis until it reaches this directed cycle. Therefore, we can contain the fire with one defender per time step for any grid in this subcase.

For the second subcase, when all arcs on the positive x-axis face right, all arcs on the positive y-axis face up, and at least one arc in the first quadrant (not including the axes) faces up or right, choose a closest arc (in terms of the underlying undirected grid) to the origin in the first quadrant (not including the axes) facing up or right and call it e'. We claim that e' has its tail on an axis. Suppose not. Then the arcs directly below the tail and directly to the left of the tail must be facing down and left, respectively, because e' was the closest arc facing up or right. However, then the tail of e' has out-degree at least three, which is not possible. Therefore, e' must have its tail on an axis.

Without loss of generality, assume the tail of e' is on the x-axis and therefore e' is pointing up. All other vertical arcs directly to the left of e' between the positive x-axis and $y = 1$ and to the right of the positive y-axis must point down by our choice of e'. All arcs on $y = 1$ to the left of e' and to the right of the positive y-axis must point left, or there would be an up or right arc closer to the origin than e'. Thus there is a directed cycle along the positive x-axis, starting at $(1, 0)$, through e', then back along $y = 1$ to the downward arc from $(1, 1)$ to $(1, 0)$. By first defending $(0, 1)$, we force the fire into this directed cycle, and therefore can contain the fire.

For the third subcase, when all arcs on the positive x-axis face right, all arcs on the positive y-axis face up, and no arcs in the first quadrant (not including the axes) face up or right, Figure 5 shows all of the arcs that have predetermined directions.

If the arc between $(1, -1)$ and $(0, -1)$ points left, it completes a directed cycle including these vertices and $(0, 0)$ and $(1, 0)$. In this case, we could contain the fire with one firefighter per time step. If this arc points right, then it forces all of the arcs on $y = -1$ to the right of this arc to point right as well. It also forces the arc between $(0, -2)$ and $(0, -1)$ to point up, while all other vertical arcs directly to the right of this arc point down. This is shown in Figure 6.
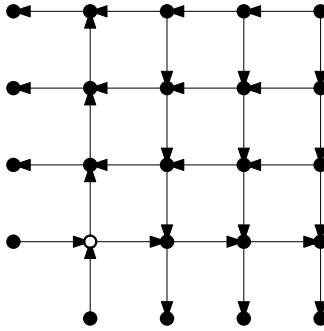
Figure 5. The directions of the arcs in the third subcase.
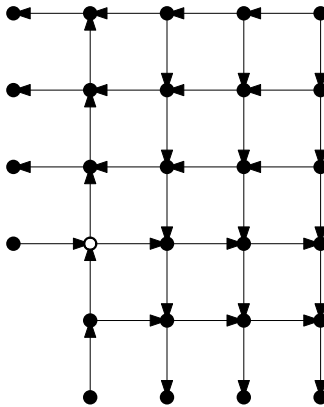
Figure 6. Another level of arcs in the third subcase.

We can continue this process, inductively going level by level down in the fourth quadrant. We will always be able to contain the fire, unless all vertical arcs on the negative y-axis point up, all vertical arcs to the right of the negative y-axis point down and all horizontal arcs below the positive x-axis point right.

By a similar argument to that of the fourth quadrant, in the second quadrant we can always contain the fire unless all horizontal arcs on the negative x-axis point right, all horizontal arcs above the negative x-axis point left, and all vertical arcs in the second quadrant point up.

Notice that, at this point, the arcs in the first, second, and fourth quadrants are the same as in the exception in Figure 1. Since every vertex has in-degree two and out-degree two, the arcs in the third quadrant are forced to be the same as the arcs in the third quadrant of the exception. We can see this by arguing inductively out from the second and fourth quadrants.

Finally, we prove that this exception is in category B. Assume we have one defender per time step. No matter where we put the first defender, the fire will
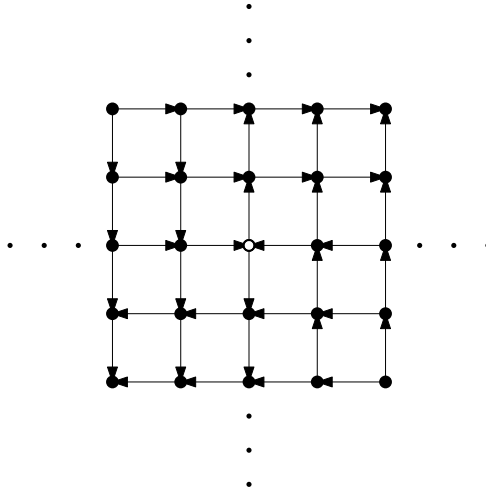
Figure 7. The exception to Theorem 3.2.

spread to at least one of $(1, 0)$ and $(0, 1)$. Without loss of generality, assume the fire moves to $(1, 0)$. By Lemma 2.2, considering $(1, 0)$ to be the origin of an infinite directed quarter-plane contained in the fourth quadrant, one firefighter per time step cannot contain the fire, but a second firefighter at some time step will allow us to contain the fire. If we get the second firefighter at time step $t = 1$, then we can immediately contain the fire.   □

The second case is when the two arcs facing out point in opposite directions. Without loss of generality, assume they point along the positive and negative y-axis. The following theorem classifies which grids are in category A and which grids are in category B in this case.

Theorem 3.2. *Let G be a regular infinite directed grid where the vertical arc directly above the origin faces up and the vertical arc directly below the origin faces down. Then G is a category A grid unless the grid is the exception shown in Figure 7 or a reflection of this figure across the y-axis. These exceptions are in category B.*

Proof. Two cases must be considered. The first case is when both of the horizontal arcs incident on at least one of $(0, 1)$ and $(0, -1)$ point away from the vertex. The second case is when both $(0, 1)$ and $(0, -1)$ have one of their horizontal arcs facing them and one pointing away. These two cases are shown in Figure 8.

In the first case, assume without loss of generality that both horizontal arcs at $(0, 1)$ point away from $(0, 1)$. At least one of the horizontal arcs at $(0, -1)$ points away from $(0, -1)$, and we can assume, without loss of generality, that this arc is the arc directly to its right. If the arc between $(1, 0)$ and $(1, 1)$ faces down, there is

Case 1   Case 2
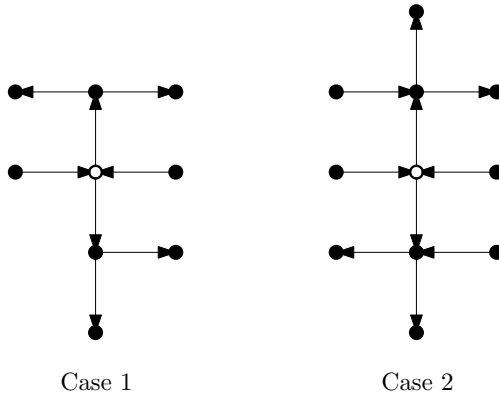
Figure 8. Two cases for Theorem 3.2.

a directed cycle from $(0, 0)$ to $(0, 1)$, to $(1, 1)$, to $(1, 0)$, and back to $(0, 0)$. If this arc faces up, though, it forces the arc between $(1, 0)$ and $(1, -1)$ to face up as well. We then have a directed cycle from $(0, 0)$ to $(0, -1)$, to $(1, -1)$, to $(1, 0)$, and back to $(0, 0)$. Therefore, for any arrangement of the remaining arcs in this first case, the grid can be defended by one firefighter per time step.

For the second case, if the horizontal arcs that point away from $(0, 1)$ and $(0, -1)$ point in the same direction, then by the same argument as that of the first case, the grid can be defended by one firefighter per time step. If the horizontal arcs pointing away from $(0, 1)$ and $(0, -1)$ point in opposite directions, without loss of generality, we can assume the arc between $(0, 1)$ and $(1, 1)$ points right and the arc between $(0, -1)$ and $(-1, -1)$ points left.

Suppose one or more arcs lying in the quarter-plane determined by $x \geq 0$ and $y \geq 1$ points left or down, or one or more arcs lying in the quarter-plane determined by $x \leq 0$ and $y \leq -1$ points right or up. Without loss of generality assume one or more arcs lying in the quarter-plane determined by $x \geq 0$ and $y \geq 1$ points left or down. We place our first defender at $(0, -1)$, forcing the fire to spread to $(0, 1)$. Unless all arcs lying in the quarter-plane determined by $x \geq 0$ and $y \geq 1$ look like the first quadrant in the exception of Theorem 3.1 (i.e., all arcs lying in the quarter-plane determined by $x > 0$ and $y > 1$ point left or down, all arcs directly to the right of $(0, 1)$ point right, and all arcs directly above $(0, 1)$ point up), we now know by Theorem 3.1 that we can contain the fire, treating the arcs lying in the quarter-plane determined by $x \geq 0$ and $y \geq 1$ as the first quadrant in Theorem 3.1.

If all arcs lying in the quarter-plane determined by $x > 0$ and $y > 1$ point left or down, all arcs directly to the right of $(0, 1)$ point right, and all arcs directly above $(0, 1)$ point up, then the arc between $(1, 0)$ and $(1, 1)$ must point down, completing a directed cycle from $(0, 0)$ to $(0, 1)$, to $(1, 1)$, to $(1, 0)$, and back to $(0, 0)$. This case can therefore be defended by one firefighter per time step.

Finally, if all arcs lying in the quarter-plane determined by $x \geq 0$ and $y \geq 1$ point up or right, and all arcs lying in the quarter-plane determined by $x \leq 0$ and $y \leq -1$ point down or left, then all of the directions of these arcs are the same as the directions of these arcs in the exception in Figure 7. Since every vertex has in-degree two and out-degree two and we know the direction of the arcs at the origin, we can see that the remaining vertical arcs in the first quadrant must point up and the remaining vertical arcs in the third quadrant must point down. Then, level by level, the remaining arcs in the second and fourth quadrants (including the axes) are forced to match the directions of the arcs in the exception.

We now prove that this exception is a category B grid. Assume we have one firefighter per time step. No matter where we put the first defender, the fire will spread to at least one of $(0, 1)$ and $(0, -1)$. Without loss of generality, assume the fire moves to $(0, 1)$. If we treat $(0, 1)$ as the origin, then by Lemma 2.2, one firefighter per time step is not enough to contain the fire. By this same lemma, a second firefighter at any time step allows us to contain the fire. If we get the second firefighter at time step $t = 1$, we can contain the fire immediately. Thus this grid is a category B grid. Notice that if we had assumed the arc between $(0, 1)$ and $(1, 1)$ points left and the arc between $(0, -1)$ and $(-1, -1)$ points right, then we would have the reflection of this exception over the y-axis.   □

4. Other infinite directed grids

As a variation of the work done in the previous section, we will now consider an infinite directed grid where all vertices have in-degree two and out-degree two except for a single vertex. We will only investigate the cases when this vertex has in-degree three and out-degree one or in-degree four and out-degree zero. We will think of the construction of one of these grids as a process, starting with a grid where each vertex has in-degree two and out-degree two. We will then change the directions of one or more arcs at a single vertex so that it has the desired degrees and then change arcs at other vertices in such a way that all other vertices still have in-degree two and out-degree two. Note that we may not always make a minimum number of changes in order for this to be the case.

Any time a single arc between u and v is changed in a grid, if all vertices except v are required to maintain their original in-degree and out-degree, then a trail of vertices from v must be changed. If the arc had been facing from v to u, then, when it is changed to point toward v, one of the arcs that had previously been facing away from u must be changed to point toward it. Call the vertex that this arc had previously been facing w. Now, in order for w to continue to have the same in-degree and out-degree, another arc that had previously been facing away from w must now face towards it. This continues, forming a trail of changed

arcs. Moreover, this trail is directed in such a way that, from any point on the trail, we could follow the arcs on the trail back to v. In the other case, when the arc between u and v was facing toward v, then the trail would face the other way, and following the arcs would take us away from v.

Let us now consider the case where v changes to have in-degree three and out-degree one. We will consider this case for most of the rest of this section. Since all vertices except v still have in-degree two and out-degree two, the grid is very similar to the type of grids investigated in Section 3. For this reason, we will refer often to the defense strategies provided for those grids.

Even though it is possible to form more than one trail of changed arcs as we change the degrees of v, we will now suppose that our grid where v has in-degree three and out-degree one contains only one trail of changed arcs. The situation where more than one trail is formed is considered later in this section — in particular, in Figures 10–13. We will show that the single changed trail can only either help move a grid from category B to category A or keep a grid in its original category. It can never bring a grid from category A to category B. We will first prove that a grid cannot go from category A to category B, which implies that all of the grids in category A must remain in category A. We will then determine which category B grids move to category A, and which category B grids stay in category B.

Theorem 4.1. *Suppose we have a category A infinite directed grid where each vertex has in-degree two and out-degree two. If one vertex, v, changes to have in-degree three and out-degree one in such a way that it creates only one trail of changed arcs, then this grid must remain in category A.*

Proof. As discussed above, if there is only one trail of changed arcs, then it must be an infinite directed trail where the arcs point toward v. Call this trail T.

We need to make an observation about how we defend the fire in category A grids where every vertex has in-degree two and out-degree two. In our proofs of Theorems 3.1 and 3.2, when we are able to contain the fire with one firefighter per time step, at each time step the fire could possibly move from a burning vertex to two neighbors since that vertex has out-degree two. We always place our firefighter at one of these two neighbors, and the fire moves to the other neighbor unless that other neighbor has already been burned or defended, in which case we finish containing the fire. Since this is true at every time step, there is at most one new burning vertex at each time step. Thus the burned vertices in all of our containment strategies follow a single directed path from the origin, which we will call P.

If T and P have no vertices in common, then we can use the same defense strategy as would have been used in Theorems 3.1 or 3.2, following P until it has been contained. Otherwise, consider the first vertex on P that is also on T. This situation is shown in Figure 9. This vertex could be v itself, or any other vertex
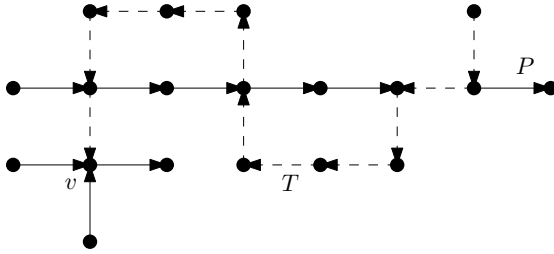
Figure 9. T intersects P (the dashed arcs are T).

on T. We begin by defending as we would have before, forcing the fire along P, until the fire reaches this first shared vertex. At this point we will force the fire to follow T until it reaches v. This is possible because every vertex of T other than v has in-degree two and out-degree two. Once the fire has spread to v, there is at most one vertex to which it can spread since v has out-degree one. We then defend this vertex, if necessary, and therefore contain the fire. In either case, we are still able to contain the fire with one firefighter per time step. The changed grid is therefore still in category A.   □

We will now determine which category B grids are able to become category A grids after the change to v results in one changed trail, T. From Theorem 3.1, one type of category B grid is the grid where all arcs in a single quadrant (including its axes) face away from the origin (without loss of generality, assume this is the first quadrant); the directions of the arcs in the remaining quadrants are irrelevant. If v is in the first quadrant (including its axes), then the fire can be forced to v, at which point we are able to contain the fire. If v is not in the first quadrant, but T contains any arcs that are in the first quadrant, then the vertex w of T that is both in the first quadrant and is closest to v on T must be on an axis. We can force the fire along this axis until it reaches w and then force the fire to follow T to v, where we can contain the fire. If, however, v is not in the first quadrant and T does not affect any arcs in the first quadrant (as an example, consider when v is any vertex in the third quadrant and T consists of precisely the edges to the left of v), then one firefighter per time step will still not be enough to contain the fire. This is the only situation where a category B grid of this type remains in category B.

The other category B grid from Theorem 3.1 is the exception in that theorem (see Figure 1). We will show that, wherever v is on the grid, it will become a category A grid. If v is in the second or fourth quadrants (not including their axes), then we are able to force the fire to v, at which point we are able to contain the fire. If v is in the first quadrant (including the axes), then the construction will create a trail, T, of changed arcs that must intersect an axis at some vertex. We are able to force the fire along that axis to the first vertex on the axis that is also on T. Now we force

the fire along T until we reach v, where the fire can be contained. If v is in the third quadrant (including the axes), since T is changing arcs in such a way that the trail points toward v, it will at some point either enter the second or fourth quadrant (not including their axes) or it will pass through the origin. If T reaches the second or fourth quadrant, we can force the fire to T and then follow T to v, where the fire is contained. If T passes through the origin, then from the very first time step we should force the fire to follow T until it reaches v.

The exception in Figure 7 from Theorem 3.2 is the only category B grid in that theorem (along with its reflection across the y-axis). This grid also becomes a category A grid, regardless of the position of v. For clarity in this proof, we will identify four regions in this grid. Region 1 is where $x \geq 0$ and $y \geq 1$; Region 2 is where $x \leq 0$ and $y \geq -1$; Region 3 is where $x \leq 0$ and $y \leq -1$; Region 4 is where $x \geq 0$ and $y \leq 1$. If v is in Regions 1 or 3 (including the boundaries), then we can force the fire to v and it can be contained. If v is in Regions 2 or 4 (not including the boundaries), then T must either reach Region 1 or 3 or it must pass through the origin. If T reaches Region 1 or 3, then it must reach a boundary in that region; we can force the fire along that boundary to T, and then force the fire to follow T to v. If T passes through the origin, then from the first time step we can force the fire to follow T to v. The only remaining case is when v is at the origin, in which case we are able to contain the fire on the first time step with one firefighter.

We can now see that the only type of category B grid that stays in category B is the grid where all arcs in a quadrant face away from the origin and v does not lie in that quadrant nor does T affect any arcs in that quadrant.

When changing v so that it has in-degree three, out-degree one, and only one trail, T, of changed arcs, we have seen that all category A grids remain in category A, some category B grids remain in category B, and some category B grids become category A grids. It might appear that changing v so that it has in-degree three and out-degree one could only help us contain the fire with one firefighter per time step since it has out-degree one, never permitting category A grids to become category B. However, if v creates more than one trail of changed arcs, it is possible for grids in either category to stay in that category or to switch to the other category. We now provide examples of each situation below. In each example, the white vertex is the origin, and the dashed arcs are the arcs that changed directions.

If we change vertex v so that it has in-degree four and out-degree zero, it creates an even number of two or more trails of changed arcs throughout the grid. If only two trails are created, then they both face toward v, so, similar to Theorem 4.1, they can never change a category A grid to category B. If, however, there are four or more trails created by changing v, it is possible for grids in either category to stay in that category or switch to the other category. Examples of this are similar to those in Figures 10–13.
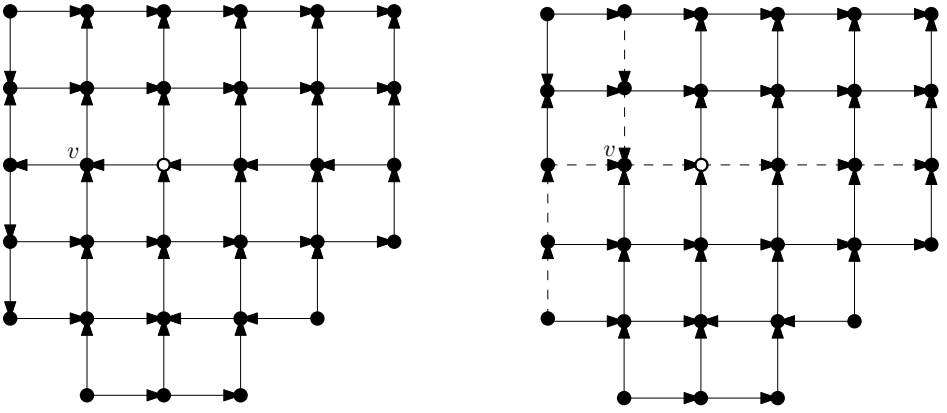
Figure 10. A category A grid that becomes a category B grid.
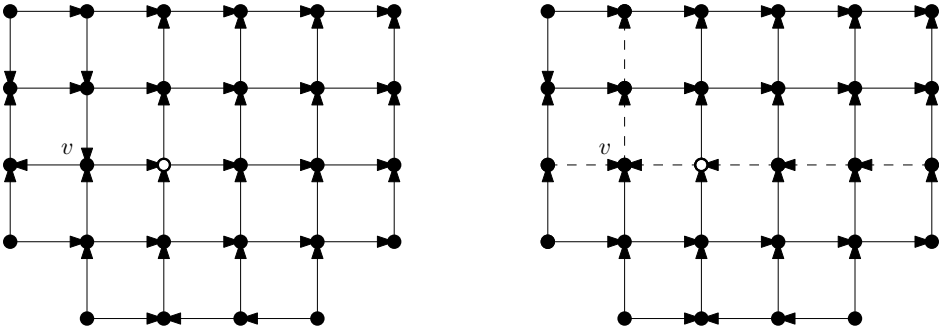
Figure 11. A category B grid that becomes a category A grid.
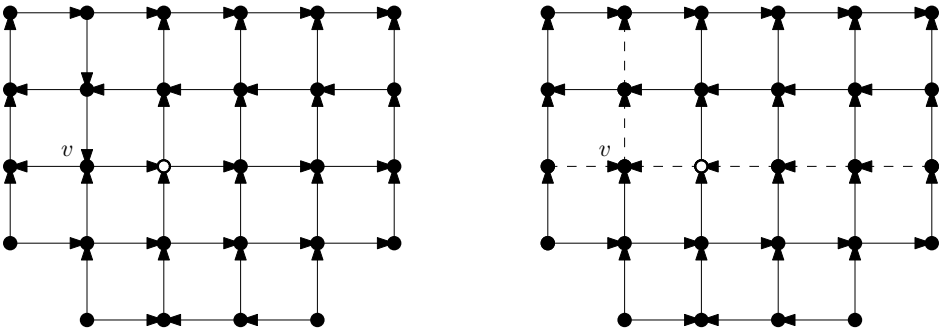
Figure 12. A category A grid that stays category A.

We close with a conjecture for general infinite directed grids. As mentioned in the introduction, we know that two firefighters per time step is sufficient to contain the fire if it begins at the origin. If the grid is regular, by Theorem 1.1, we know that either one firefighter per time step or one firefighter per time step with an additional
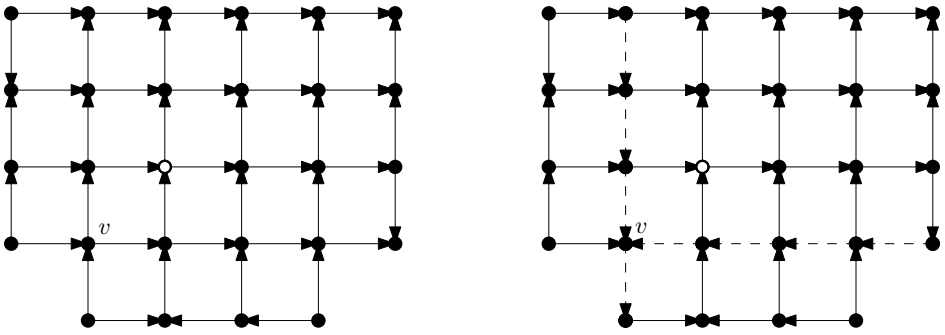
Figure 13. A category B grid that stays category B.

firefighter at some time step is sufficient to contain the fire. In general, we believe the following conjecture holds.

Conjecture 4.2. Let G be an infinite directed grid, and assume that the fire begins at the origin. If we are given one firefighter per time step and an occasional second firefighter is given on some finite number of time steps (the number may depend on the grid), the fire can be contained.

In some related work, Messinger [2008] and Ng and Raff [2008] provide containment strategies for undirected grids utilizing one firefighter on some time steps and two firefighters on other time steps. Their strategies, however, make assumptions as to which time steps a second firefighter will be available. Thus their strategies can be used for some instances of our conjecture, but they do not settle the general case.

The worst possible scenario for an infinite directed grid appears to be the grid where all horizontal arcs in the half-plane $x > 0$ point right, all horizontal arcs in the half-plane $x < 0$ point left, all vertical arcs in the half-plane $y > 0$ point up, and all vertical arcs in the half-plane $y < 0$ point down, seemingly allowing the fire to spread as much as possible. All four of the origin's incident arcs are directed away from the origin, and all other vertices on the x- and y-axes are in-degree one and out-degree three. The remaining vertices have in-degree two and out-degree two. Here is a defense strategy for this grid. In the first time step, we place a firefighter directly to the left of the origin, and we continue to place firefighters vertically above this vertex until a second firefighter is available, which allows us to push the fire to the right instead of simply maintaining it with this continuing vertical line of firefighters. Single firefighters are then again used to maintain the fire in a horizontal fashion until extra firefighters allow us to begin to push the line of defense downwards. In this general pattern, we can corral the fire quadrant by quadrant in a clockwise direction, maintaining the direction of the fire when given only one firefighter, and steering it in a clockwise direction when given an extra firefighter. Using this strategy, we will contain the fire after finitely many time steps.

Acknowledgment

References

[Finbow and MacGillivray 2009] S. Finbow and G. MacGillivray, "The firefighter problem: a survey of results, directions and questions", *Australas. J. Combin.* **43** (2009), 57–77. MR 2010a:05175 Zbl 1179.05112

[Fogarty 2003] P. Fogarty, "Catching the fire on grids", Master's thesis, Department of Mathematics, University of Vermont, 2003, http://www.cems.uvm.edu/~jdinitz/firefighting/fire.pdf.

[Hartke 2004] S. G. Hartke, *Graph-Theoretic Models of Spread and Competition*, Ph.D. thesis, Rutgers, 2004, http://dmac.rutgers.edu/Workshops/WGDataMining/HartkeDissertation.pdf.

[Hartnell 1995] B. Hartnell, "Firefighter! An application of domination", conference paper, 25th Manitoba Conference on Combinatorial Mathematics and Computing, 1995.

[Messinger 2008] M. E. Messinger, "Average firefighting on infinite grids", *Australas. J. Combin.* **41** (2008), 15–28. MR 2009e:05229 Zbl 1178.05068

[Ng and Raff 2008] K. L. Ng and P. Raff, "A generalization of the firefighter problem on $\mathbb{Z} \times \mathbb{Z}$", *Discrete Appl. Math.* **156**:5 (2008), 730–745. MR 2009a:05157 Zbl 1134.05101

[Wang and Moeller 2002] P. Wang and S. A. Moeller, "Fire control on graphs", *J. Combin. Math. Combin. Comput.* **41** (2002), 19–34. MR 2003c:05214 Zbl 1019.05035

biebigha@cord.edu   *Department of Mathematics, Concordia College,*
                           901 8th Street South, Moorhead, MN 56562, United States

leholte@ncsu.edu   *Department of Mathematics, North Carolina State University,*
                           2200 Hillsborough Street, Raleigh, NC 27695, United States

rmwagner@cord.edu   *Department of Mathematics, Concordia College,*
                           901 8th Street South, Moorhead, MN 56562, United States

msp

Induced trees, minimum semidefinite rank, and zero forcing

Rachel Cranfill, Lon H. Mitchell, Sivaram K. Narayan and Taiji Tsutsui

(Communicated by Chi-Kwong Li)

We prove that the ordered subgraph number of a connected graph that has no duplicate vertices is at most three if and only if the complement does not contain a cycle on four vertices. The duality between zero forcing and ordered subgraphs then provides a complementary characterization for positive semidefinite zero forcing. We also provide some necessary conditions for when the minimum semidefinite rank can be computed using tree size.

1. Introduction

Graph theory provides a natural way to describe patterns in the entries of matrices and a large body of research and terminology to help study those patterns. Conversely, matrices that are associated to graphs can provide structural information about the graph. For example, the second-smallest eigenvalue of the Laplacian matrix of a graph is nonzero if and only if the graph is connected [Merris 1995].

The research described in this paper was inspired by the question of finding the smallest possible rank among matrices with a given zero/nonzero (off-diagonal) entry pattern. Depending on the type of matrices one allows (for example, real or complex, symmetric or not), different answers for the same pattern are possible [Berman et al. 2008; IMA-ISU 2010; Barioli et al. 2009], and a complete solution to this problem for any large class of matrices seems difficult. On the other hand, for certain types of patterns (graphs), there are very satisfying complete answers. For example, for trees and positive semidefinite (psd) real symmetric or complex Hermitian matrices, the minimum rank is equal to one less than the number of vertices [van der Holst 2003; Johnson and Duarte 2006]; for trees and symmetric matrices over any field, the minimum rank plus the zero forcing number gives the number of vertices [Chenette et al. 2007; Johnson and Duarte 1999].

One part of our work, described in Section 4, seeks to use the detailed knowledge we have for trees in general graphs. In particular, if a graph contains a tree as an induced subgraph, under what conditions will matrices associated to the larger graph behave like those for the tree with respect to minimum rank?

Rather than looking for trees, participants in the 2004 Research Experience for Undergraduates at Central Michigan University sought to find an alternative that would provide just as much rank information. The result, designed specifically for Hermitian psd matrices, was called *ordered subgraphs* [Hackney et al. 2009]. For some time, it was conjectured that ordered subgraphs would in fact determine minimum rank, but a counterexample on eight vertices was found: the Möbius ladder on eight vertices has psd minimum rank (msr) five and an ordered subgraph (OS) number of four [Mitchell et al. 2010].

Results on ordered subgraphs are of additional interest thanks to their connection to "zero forcing." Defined by the AIM Minimum Rank-Special Graphs Work Group [AIM 2008], zero forcing was also the result of looking for approaches to solving a minimum rank problem, but has since been shown to be of interest in quantum physics [Burgarth et al. 2011]. It turns out that the OS number and the positive semidefinite zero forcing number are two sides of the same coin, as for any graph they sum to the number of vertices [Barioli et al. 2010]. Moreover, the complement of an OS set is a zero forcing set and vice versa. This duality means that our OS results have an equivalent formulation in terms of zero forcing.

One of the many open questions concerning ordered subgraphs (and zero forcing) is how large the class of graphs is for which minimum rank and the ordered subgraph number differ. If the msr of a graph is one or two, then so is the OS number. The Möbius ladder example means that msr three is the remaining case[1] in which we might hope that msr and the ordered subgraph number coincide. In Section 3, we study graphs that have msr 3, show that msr 3 implies OS number 3, and give a characterization of those graphs with OS number 3. Whether OS number equal to 3 implies msr 3 remains open, although we are able to use our work on maximum induced trees from Section 4 to present some partial results in Section 5.

2. Preliminaries

A *graph G* is an ordered pair $(V(G), E(G))$, where $V(G)$ is a set of vertices and $E(G)$ is a set of unordered pairs of vertices. In this paper, we assume all graphs are simple (that is, have no multiple edges or loops). Two vertices u and v are said to be *adjacent* if they share an edge. If u and v are adjacent, we write $uv \in E(G)$.

---

[1]For small rank, that is — some results are known for small nullity as well; see for example [van der Holst 2003].

For any $n \times n$ Hermitian matrix $A = [a_{ij}]$, we associate a simple graph $G(A)$ with vertex set $V(G) = \{v_1, \ldots, v_n\}$ and $v_i v_j \in E(G)$ if and only if $a_{ij} \neq 0$ in A. Note that $G(A)$ is independent of the diagonal elements of A. For a given graph G, we define $\mathscr{P}(G)$ to be the set of all positive semidefinite matrices with graph G. The *minimum semidefinite rank* of G is

$$\mathrm{msr}(G) = \min\{\mathrm{rank}\, A : A \in \mathscr{P}(G)\}.$$

If there is a path between two vertices u and v in G, the *distance* from u to v, $d_G(u, v)$, is the length of the shortest path between u and v. If no such path exists, we say $d_G(u, v) = \infty$.

The *tree size* of a graph G, $\mathrm{ts}(G)$, is the maximum size of a subset of $V(G)$ that induces a tree [Erdős et al. 1986]. Since $\mathrm{msr}(G) = |G| - 1$ if and only if G is a tree, this gives a general lower bound of $\mathrm{msr}(G) \geq \mathrm{ts}(G) - 1$ [Booth et al. 2008].

Let the *neighborhood* of a vertex v in G be $N(v) = \{w \in V(G) : vw \in E(G)\}$, and let the *closed neighborhood* of v be $N[v] = N(v) \cup \{v\}$. We say vertices u and w are *duplicate vertices* if $N[u] = N[w]$.

If $S \subseteq V(G)$ such that all of the vertices in S are pairwise nonadjacent, we say S is an *independent set*. The maximum cardinality of all independent sets of a graph G is called the *independence number* of G and is denoted by $\alpha(G)$ [West 1996, p. 113].

The *union* of two graphs G_1 and G_2, denoted by $G_1 \cup G_2$, is the disconnected graph with vertex set $V(G_1) \cup V(G_2)$ and edge set $E(G_1) \cup E(G_2)$. We frequently write the union of k copies of a graph G as kG. The *join* of G_1 and G_2, written $G_1 \vee G_2$, is the graph with vertex set $V(G_1) \cup V(G_2)$ and edge set consisting of all of the edges in $E(G_1)$ and $E(G_2)$ as well as the edges $\{uv : u \in V(G_1), v \in V(G_2)\}$ [West 1996, p. 118].

Suppose $\vec{V} = \{\vec{v}_1, \ldots, \vec{v}_n\}$ is an n-tuple of vectors in $\mathbb{C}^m$ such that, for $i \neq j$, we have $\langle \vec{v}_i, \vec{v}_j \rangle = 0$ if and only if $v_i v_j \notin E(G)$. We call $\vec{V}$ a *vector representation* of G [Parsons and Pisanski 1989]; the rank of $\vec{V}$ is defined as the dimension of the span of the vectors.

Let $\vec{V} = \{\vec{v}_1, \ldots, \vec{v}_n\}$ be a vector representation of G. If $V = [\vec{v}_1 \cdots \vec{v}_n]$, then $V^* V \in \mathscr{P}(G)$. If $A \in \mathscr{P}(G)$, then $A = B^* B$ for some matrix B with the same rank [Horn and Johnson 1990, p. 407]. Thus, for any $A \in \mathscr{P}(G)$, we can find a vector representation of G that produces A. This implies that finding a vector representation for a graph is equivalent to finding a positive semidefinite matrix of the graph.

Let G be a graph on n vertices and let $S = (v_1, \ldots, v_m)$ be an ordered set of vertices of G. Let G_k be the subgraph of G induced by $\{v_1, \ldots, v_k\}$ for $k \leq m$, and let H_k be the connected component of G_k containing v_k. If for each k there exists a vertex w_k of G such that $w_k \notin G_k$, $w_k v_k \in E(G)$, and $w_k v_l \notin E(G)$ for

all $v_l \in V(H_k)$ with $l \neq k$, we say S is a *vertex set of ordered subgraphs* (OS-*set*) of G [Hackney et al. 2009].

For every v_k in an OS-set, we call its corresponding w_k its OS-*neighbor*. The maximum cardinality of all OS-sets of a graph G is called the OS-*number* of G, denoted by $OS(G)$.

Example 2.1. In the cycle C_4, $OS(C_4) = 2$. Here are some examples of OS-sets of C_4:
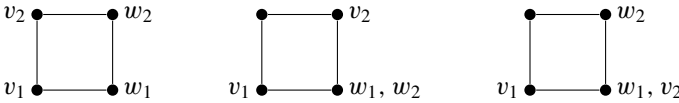
Proposition 2.2 [Hackney et al. 2009]. *If G is a connected graph then* $msr(G) \geq OS(G) \geq ts(G) - 1$. *In particular, if T is a tree, for every $v \in V(T)$, $V(T) \setminus \{v\}$ is an OS-set.*

If H is an induced subgraph of G, then $OS(H) \leq OS(G)$. The OS-number is related to the positive semidefinite zero forcing number, $Z_+(G)$, by $OS(G) + Z_+(G) = |G|$ [Barioli et al. 2010].

3. Graphs with minimum semidefinite rank three

An open question that has been of interest is a complete characterization of all graphs for which $msr(G) = 3$. Some prior results [Booth et al. 2011; AIM 2008] give sufficient conditions, including if $\overline{G} = P_n$ with $n \geq 4$ or $\overline{G} = C_n$ with $n \geq 5$ then $msr(G) = 3$, and a sufficient condition for when $msr(G) \leq 3$:

Proposition 3.1 [Booth et al. 2011]. *If the cycle C_m is not a subgraph of $\overline{G}$ for all $m \geq 4$, then* $msr(G) \leq 3$.

From examples, however, it seems that avoiding C_4 in the complement is enough.

Conjecture 3.2. Let G be a connected graph with no duplicate vertices. Then $msr(G) \leq 3$ if and only if C_4 is not a subgraph of $\overline{G}$.

Remark 3.3. Conjecture 3.2 is not true if the duplicate vertices condition is removed. For example, if G is the graph obtained by identifying an edge of the complete graph on four vertices with an edge of a C_4 (resulting in a graph on six vertices), then a C_4 is a subgraph of $\overline{G}$ but $msr(G) = 3$.

We now prove several results that are related to this conjecture, including that this result holds for the OS-number.

Lemma 3.4. *Let G be a simple connected graph. If $S = (v_1, v_2, v_3, v_4)$ is an OS-set of G, then there is an OS-set S' of G of size four such that $G[S']$ has at least two components and each component has at most two vertices.*

Proof. If $G[S]$ has three or four connected components, the conclusion follows. Otherwise, we consider two cases:

Case 1: $G[S]$ has two connected components, $G[\{v_1, v_2, v_3\}]$ and $G[\{v_4\}]$. Then $w_3 \notin N[v_1] \cup N[v_2]$ and $G[\{v_1, v_2, w_3, v_4\}]$ has at least two components with each component having at most two vertices. Also, $S' = (v_1, v_2, v_4, w_3)$ is an OS-set with OS-neighbors (w_1, w_2, w_4, v_3).

Case 2: Suppose $G[S]$ is connected. Then $w_4 \notin \bigcup_{i=1}^{3} N[v_i]$, and therefore $G[\{v_1, v_2, v_3, w_4\}]$ has at least two components. Furthermore, $S_1 = (v_1, v_2, v_3, w_4)$ is an OS-set with OS-neighbors (w_1, w_2, w_3, v_4), reducing the problem to case 1. $\square$

Remark 3.5. If S_1 and S_2 are OS-sets of G such that there are no edges $vw \in E(G)$ with $v \in S_1$ and $w \in S_2$, then $S_1 \cup S_2$ is an OS-set.

Lemma 3.6. *Let G be a connected graph with no duplicate vertices. If an induced subgraph H of G is isomorphic to $sK_2 \cup tK_1$, then the vertices of H form an OS-set.*

Proof. Clearly, K_1 is an OS-set since G is connected. Let $K_2 = \{v, w\}$. Since G has no duplicate vertices, $N[v] \neq N[w]$. Without loss of generality, we can assume there is a vertex u adjacent to v but not adjacent to w. Then (w, v) is an OS-set with neighbors (v, u). $\square$

Proposition 3.7. *Let G be a connected graph with no duplicate vertices. Then $OS(G) \geq 4$ if and only if $\overline{G}$ contains C_4 as a subgraph.*

Proof. Lemma 3.4 and Lemma 3.6 imply that $OS(G) \geq 4$ if and only if G contains $4K_1$, $2K_1 \cup K_2$, or $2K_2$ as an induced subgraph. However, $\overline{4K_1}$ is K_4, $\overline{2K_1 \cup K_2}$ is K_4 minus an edge, and $\overline{2K_2}$ is C_4, giving the desired result. $\square$

As a consequence of Proposition 3.7, we see the absence of a C_4 subgraph in $\overline{G}$ is necessary for $msr(G) \leq 3$. We believe that this condition is sufficient and can be shown by proving $OS(G) = 3$ if and only if $msr(G) = 3$. We do know, however, that if G is a connected graph without duplicate vertices and $msr(G) \leq 3$, then $msr(G) = ts(G) - 1$ [Booth et al. 2011]. As a result, we have:

Proposition 3.8. *If $msr(G) = 3$, then $OS(G) = 3$ (and $Z_+(G) = |G| - 3$).*

Conjecture 3.9. Suppose G is a connected graph without duplicate vertices. If $OS(G) = 3$, then $msr(G) = 3$.

4. Maximum induced trees

Let T be a maximum induced tree of a graph G. For a vertex w in $V(G)$ such that w is not on T, we define $\mathscr{E}(w)$ to be the edge set of all paths in T between every pair of vertices of T that are adjacent to w.

Prior work on minimum semidefinite rank has yielded a sufficient, but not necessary, condition for when $msr(G) = ts(G) - 1$ [Booth et al. 2008]:

⊛ There exists a maximum induced tree T such that for u and w not on T, $\mathscr{E}(u) \cap \mathscr{E}(w) \neq \varnothing$ if and only if u and w are adjacent in G.

We now present some sufficient conditions for strict inequality.

Proposition 4.1. *Let T be a maximum induced tree of a graph G. If u and w are vertices not on T such that $uw \notin \mathscr{E}(G)$, $|\mathscr{E}(u) \cap \mathscr{E}(w)| = 1$, and u and $w are only adjacent to the longest path P of T that contains $\mathscr{E}(u) \cap \mathscr{E}(w)$, then $\mathrm{msr}(G) > \mathrm{ts}(G) - 1$.*

Proof. The vertices of T not on P belong to an OS-set S. We enlarge S by adding the vertices on P. Let $P = v_1 v_2 \cdots v_i x y v_{i+1} \cdots v_{k-1} v_k$, and without loss of generality assume $xw \in \mathscr{E}(G)$ and $yu \in \mathscr{E}(G)$, where $\{xy\} = \mathscr{E}(u) \cap \mathscr{E}(w)$. We add vertices $v_k, v_{k-1}, \ldots, v_{i+2}, v_{i+1}$ to the set S since we can find OS-neighbors $v_{k-1}, v_{k-2}, \ldots, v_{i+1}, y$, respectively. Then we add w, y, and x in that order to the set followed by $v_i, \ldots, v_2$ since these vertices have OS-neighbors $x, u, v_i, \ldots, v_1$ respectively. The size of this enlarged OS-set is $\mathrm{ts}(G)$. Thus, $\mathrm{msr}(G) \geq \mathrm{OS}(G) > \mathrm{ts}(G) - 1$. □

This leads us to the following result.

Corollary 4.2. *Let T be a maximum induced tree of a graph G. Suppose u and w are vertices not on T such that $uw \notin \mathscr{E}(G)$, $\mathscr{E}(u) \cap \mathscr{E}(w)$ contains only the edge xy where $xw \in \mathscr{E}(G)$, $P = v_1 v_2 \cdots v_i x y v_{i+1} \cdots v_{k-1} v_k$ is the longest path P of T that contains $\mathscr{E}(u) \cap \mathscr{E}(w)$, there exists a path P' on T where $P' = y t_1 t_2 \cdots t_l$ and $t_l u \in \mathscr{E}(G)$, and u and w are adjacent only to vertices of $P \cup P'$. Then $\mathrm{msr}(G) > \mathrm{ts}(G) - 1$.*

Proof. The vertices of T not on P or P' belong to an OS-set S. We enlarge S by adding the vertices of P and P'. We add vertices $v_k, v_{k-1}, \ldots, v_{i+1}$ to the set S since the set of OS-neighbors is $v_{k-1}, v_{k-2}, \ldots, y$, respectively. Then we add $w, y, t_1, \ldots, t_l$ in that order since these vertices have OS-neighbors $x, t_1, t_2, \ldots, t_l, u$, respectively. Also, we add $x, v_i, v_{i-1}, \ldots, v_2$ since the set of OS-neighbors is $v_i, v_{i-1}, \ldots, v_1$, respectively. Thus, by the same argument as in Proposition 4.1, $\mathrm{msr}(G) \geq \mathrm{OS}(G) > \mathrm{ts}(G) - 1$. □

Proposition 4.3. *Let T be a maximum induced tree of a graph G such that T is a star graph. If there exist vertices u and w not on T such that $uw \notin \mathscr{E}(G)$ and $|\mathscr{E}(u) \cap \mathscr{E}(w)| = 1$, then $\mathrm{msr}(G) > \mathrm{ts}(G) - 1$.*

Proof. The vertices of T that are not the center of T and are not adjacent to u or w belong to an OS-set. Let the center vertex of T be x and $\mathscr{E}(u) \cap \mathscr{E}(w) = \{xy\}$. We add vertices of T which are adjacent to u and not on $\mathscr{E}(u) \cap \mathscr{E}(w)$ to the OS-set since all of these vertices have OS-neighbor x. Then we add u and y in that order since they have OS-neighbors y and w. Next, we add vertices that are adjacent

to w and not on $\mathscr{E}(u) \cap \mathscr{E}(w)$ to the OS-set since they also have OS-neighbor x. Thus, the size of OS-set is $\mathrm{ts}(G)$, so $\mathrm{msr}(G) \geq \mathrm{OS}(G) > \mathrm{ts}(G) - 1$. $\qquad\square$

If $\mathscr{E}(u) \cap \mathscr{E}(w) = \varnothing$, we have the following result.

Proposition 4.4. *Let T be a maximum induced tree of a graph G. If there are two vertices $u, w \in V(G)$ such that $u, w \notin V(T)$, $uw \in \mathscr{E}(G)$, and $\mathscr{E}(u) \cap \mathscr{E}(w) = \varnothing$, then $\mathrm{OS}(G) > \mathrm{ts}(G) - 1$. In particular, $\mathrm{msr}(G) > \mathrm{ts}(G) - 1$.*

Proof. Let $G' = G[V(T) \cup \{u, w\}]$. By constructing an OS-set of size $\mathrm{ts}(G)$ in G', we will show that $\mathrm{OS}(G) > \mathrm{ts}(G) - 1$. Let $v_1, \ldots, v_a \in V(T)$ be vertices of degree one in G'. Then $(v_1, \ldots, v_a)$ forms an OS-set of G' with each v_i having corresponding w_i such that w_i is the only vertex adjacent to v_i. Let $F = G[V(G') \backslash \{v_1, \ldots, v_a\}]$. If $v_{a+1}, \ldots, v_l \in V(T)$ such that $\deg_F(v_i) = 1$ for all $i \in \{a+1, \ldots, l\}$, then $(v_1, \ldots, v_a, v_{a+1}, \ldots, v_l)$ forms an OS-set of G' where, for all $i \in \{a+1, \ldots, l\}$, w_i is the unique vertex in F such that $v_i w_i \in \mathscr{E}(F)$. We can repeat this process until all vertices of degree one in $G[V(G') \backslash \{v_1, \ldots, v_l\}]$ have been included in an OS-set of G', say $S = (v_1, \ldots, v_k)$.

Let $\mathscr{V}(u) = \{v \in V(T) : vv' \in \mathscr{E}(u) \text{ for some } v'\}$ and $\mathscr{V}(w) = \{v \in V(T) : vv' \in \mathscr{E}(w) \text{ for some } v'\}$. Without loss of generality, assume that $|\mathscr{V}(u)| \geq |\mathscr{V}(w)|$. Because $|\mathscr{V}(u) \cap \mathscr{V}(w)| \geq 2$ would imply $\mathscr{E}(u) \cap \mathscr{E}(w) \neq \varnothing$, there are two possibilities:

Case 1: $|\mathscr{V}(u) \cap \mathscr{V}(w)| = 1$. Note that if $|\mathscr{V}(u)| = n$ and $|\mathscr{V}(w)| = m$, then $\mathrm{ts}(G) = k + n + m - 1$. Suppose $v \in \mathscr{V}(u) \cap \mathscr{V}(w)$. Since $G[\mathscr{V}(u)]$ is a tree, by Proposition 2.2, $\mathscr{V}(u) \backslash \{v\} = (v_{k+1}, \ldots, v_{k+n-1})$ forms an OS-set. Furthermore, $(v_1, \ldots, v_{k+n-1}, u)$ forms an OS-set since $uw \in \mathscr{E}(G)$ but $v_i w \notin \mathscr{E}(G)$ for all $i \in \{1, \ldots, k+n-1\}$.

Now order vertices $\{x_1, \ldots, x_{m-1}\} = \mathscr{V}(w) \backslash \{v\}$ such that $d_H(x_i, u) \leq d_H(x_{i+1}, u)$ where $H = G[V(T) \cup \{u\}]$. Since for every $i \leq m - 1$ there is a $j > i$ such that $d_H(x_i, u) = d_H(x_j, u) + 1$ and where $x_j x_i \in \mathscr{E}(G)$ but x_j is not adjacent to any other vertex in the connected component of $G[\{x_1, \ldots, x_{j-1}\}]$, we now have an OS-set $(v_1, \ldots, v_{k+n-1}, u, x_1, \ldots, x_{m-1})$ of size $\mathrm{ts}(G)$.

Case 2: $\mathscr{V}(u) \cap \mathscr{V}(w) = \varnothing$. Begin by ordering vertices $u_i \in \mathscr{V}(u)$ by $d_J(u_i, w) \geq d_J(u_{i+1}, w)$ for $i = 1, \ldots, n - 1$ where $J = G[V(T) \cup \{w\}]$.

Let $H = G[V(T) \cup \{u\}]$ and define $\mathscr{V}'(w) = V(T) \backslash (\mathscr{V}(u) \cup S)$. Let v be the unique vertex in $\mathscr{V}'(w)$ such that $d_H(v, u) < d_H(x, u)$ for every $x \in \mathscr{V}'(w)$ where $x \neq v$. If $\mathscr{V}(u) = \{u_1, \ldots, u_n\}$, then, because $\{u_1, \ldots, u_n, v\}$ induces a tree on G, $(u_1, \ldots, u_n)$ forms an OS-set. Moreover, $(v_1, \ldots, v_k, u_1, \ldots, u_n, u)$ forms an OS-set, as $uw \in \mathscr{E}(G)$ but $u_i w \notin \mathscr{E}(G)$ and $v_j w \notin \mathscr{E}(G)$ for any i, j.

Order the vertices in $\mathscr{V}'(w) = \{x_1, \ldots, x_j, v\}$ such that $d_H(x_i, u) \geq d_H(x_{i+1}, u)$ for $i = 1, \ldots, j - 1$. Then $S \cup (u_1, \ldots, u_n, u, x_1, \ldots, x_j)$ is an OS-set that includes u and all vertices on the maximum induced tree except for v. $\qquad\square$

5. OS number three

In this final section, we use our work on maximum induced trees, and, in particular, the condition $\circledast$, to prove that $\mathrm{OS}(G) = 3$ implies $\mathrm{msr}(G) = 3$ for certain graphs.

Proposition 5.1. *Let G be a connected graph without duplicate vertices. If $\overline{G}$ does not contain C_4 as a subgraph then $\mathrm{msr}(G) \leq 3$ or there exists a connected graph G' without duplicate vertices such that*

(1) *G is an induced subgraph of G',*

(2) *$\overline{G'}$ does not contain C_4 as a subgraph,*

(3) *$K_{1,3}$ is an induced subgraph of G', and*

(4) *G' is not $(|G'| - 3)$-connected.*

Proof. For the last claim, if G' is $(|G'| - 3)$-connected then $\mathrm{msr}(G) \leq 3$ [van der Holst 2008; Lovász et al. 1989; 2000].

Case 1: $\alpha(G) = 3$. If necessary, form G' by adding a new vertex adjacent to all vertices of G.

Case 2: $\alpha(G) = 2$. Let $\{u, v\} \subset V(G)$ induce $2K_1$ in G. Form G' by adding a new vertex adjacent to all vertices of G except for u and v. As $\overline{G}$ does not contain K_3 as an induced subgraph, $\overline{G'}$ does not contain C_4 as a subgraph.

Case 3: $\alpha(G) = 1$. Then G is complete and $\mathrm{msr}(G) \leq 1$. $\square$

Suppose that G is a connected graph without duplicate vertices such that $\overline{G}$ does not contain C_4 as a subgraph and $\mathrm{OS}(G) = 3$. From Proposition 5.1, we may assume without loss of generality that $K_{1,3}$ is an induced subgraph of G. Therefore $K_{1,3}$ is a maximum induced tree T of G.

Remark 5.2. Since $\overline{G}$ does not contain C_4 as a subgraph, there are at most three vertices in G not belonging to T that are pairwise disjoint.

Remark 5.3. If u and v are not on T and satisfy $\circledast$, then there exists a vector representation of $G[V(T) \cup \{u, v\}]$ of rank three.

**Proposition 5.4.** *Suppose G is a connected graph without duplicate vertices such that $\overline{G}$ does not contain C_4 as a subgraph and $\mathrm{OS}(G) = 3$. Let $T = K_{1,3}$ be a maximum induced tree of G. If u, v, and w are pairwise nonadjacent vertices not on T such that no two of them satisfy $\circledast$, then $H = G[V(T) \cup \{u, v, w\}]$ has minimum semidefinite rank equal to three.*

Proof. If independent vertices u, v, and w are joined to all vertices of $K_{1,3}$, then $H = K_{1,3} \vee 3K_1$. Thus, its complement consists of $2K_3$. From this observation, since $\overline{G}$ does not contain C_4 as a subgraph, the complement of H has to be one of the following graphs:
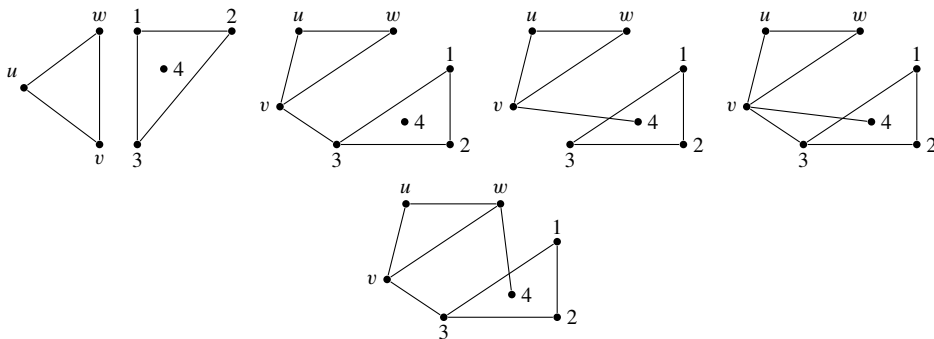
Since all of these graphs are C_m-free for $m \geq 4$, we can use Proposition 3.1 to conclude that $\mathrm{msr}(H) \leq 3$. Since $\mathrm{OS}(H) = 3$, it follows that the $\mathrm{msr}(H) = 3$. $\square$

Acknowledgements

References

[AIM 2008] AIM Minimum Rank – Special Graphs Work Group, "Zero forcing sets and the minimum rank of graphs", *Linear Algebra Appl.* **428**:7 (2008), 1628–1648. MR 2008m:05166 Zbl 1135.05035

[Barioli et al. 2009] F. Barioli, S. M. Fallat, H. T. Hall, D. Hershkowitz, L. Hogben, H. van der Holst, and B. Shader, "On the minimum rank of not necessarily symmetric matrices: a preliminary study", *Electron. J. Linear Algebra* **18** (2009), 126–145. MR 2010e:05176 Zbl 1169.05345

[Barioli et al. 2010] F. Barioli, W. Barrett, S. M. Fallat, H. T. Hall, L. Hogben, B. Shader, P. van den Driessche, and H. van der Holst, "Zero forcing parameters and minimum rank problems", *Linear Algebra Appl.* **433**:2 (2010), 401–411. MR 2011g:15002 Zbl 1209.05139

[Berman et al. 2008] A. Berman, S. Friedland, L. Hogben, U. G. Rothblum, and B. Shader, "Minimum rank of matrices described by a graph or pattern over the rational, real and complex numbers", *Electron. J. Combin.* **15**:1 (2008), Research Paper 25, 19. MR 2008k:05124

[Booth et al. 2008] M. Booth, P. Hackney, B. Harris, C. R. Johnson, M. Lay, L. H. Mitchell, S. K. Narayan, A. Pascoe, K. Steinmetz, B. D. Sutton, and W. Wang, "On the minimum rank among positive semidefinite matrices with a given graph", *SIAM J. Matrix Anal. Appl.* **30**:2 (2008), 731–740. MR 2009g:15003 Zbl 1226.05151

[Booth et al. 2011] M. Booth, P. Hackney, B. Harris, C. R. Johnson, M. Lay, T. D. Lenker, L. H. Mitchell, S. K. Narayan, A. Pascoe, and B. D. Sutton, "On the minimum semidefinite rank of a simple graph", *Linear Multilinear Algebra* **59**:5 (2011), 483–506. MR 2012e:15004 Zbl 1223.05170

[Burgarth et al. 2011] D. Burgarth, D. D'Alessandro, L. Hogben, S. Severini, and M. Young, "Zero forcing, linear and quantum controllability for systems evolving on networks", preprint, 2011. arXiv 1111.1475

[Chenette et al. 2007] N. L. Chenette, S. V. Droms, L. Hogben, R. Mikkelson, and O. Pryporova, "Minimum rank of a tree over an arbitrary field", *Electron. J. Linear Algebra* **16** (2007), 183–186. MR 2008f:05110 Zbl 1142.05335

[Erdős et al. 1986] P. Erdős, M. Saks, and V. T. Sós, "Maximum induced trees in graphs", *J. Combin. Theory Ser. B* **41**:1 (1986), 61–79. MR 87k:05062

[Hackney et al. 2009] P. Hackney, B. Harris, M. Lay, L. H. Mitchell, S. K. Narayan, and A. Pascoe, "Linearly independent vertices and minimum semidefinite rank", *Linear Algebra Appl.* **431**:8 (2009), 1105–1115. MR 2011a:15016 Zbl 1188.05085

[van der Holst 2003] H. van der Holst, "Graphs whose positive semi-definite matrices have nullity at most two", *Linear Algebra Appl.* **375** (2003), 1–11. MR 2004g:05104 Zbl 1029.05099

[van der Holst 2008] H. van der Holst, "Three-connected graphs whose maximum nullity is at most three", *Linear Algebra Appl.* **429**:2-3 (2008), 625–632. MR 2009g:05104 Zbl 1145.05037

[Horn and Johnson 1990] R. A. Horn and C. R. Johnson, *Matrix analysis*, Corrected reprint of the 1985 original ed., Cambridge University Press, 1990. MR 91i:15001 Zbl 0704.15002

[IMA-ISU 2010] IMA-ISU Research Group on Minimum Rank (Institute for Mathematics and its Applications – Iowa State University), "Minimum rank of skew-symmetric matrices described by a graph", *Linear Algebra Appl.* **432**:10 (2010), 2457–2472. MR 2011h:15001

[Johnson and Duarte 1999] C. R. Johnson and A. L. Duarte, "The maximum multiplicity of an eigenvalue in a matrix whose graph is a tree: Invariant factors", *Linear and Multilinear Algebra* **46**:1-2 (1999), 139–144. MR 2000e:05114 Zbl 0929.15005

[Johnson and Duarte 2006] C. R. Johnson and A. L. Duarte, "Converse to the Parter–Wiener theorem: the case of non-trees", *Discrete Math.* **306**:23 (2006), 3125–3129. MR 2007h:05101 Zbl 1114.05061

[Lovász et al. 1989] L. Lovász, M. Saks, and A. Schrijver, "Orthogonal representations and connectivity of graphs", *Linear Algebra Appl.* **114/115** (1989), 439–454. MR 90k:05095 Zbl 0681.05048

[Lovász et al. 2000] L. Lovász, M. Saks, and A. Schrijver, "A correction: "Orthogonal representations and connectivity of graphs" [Linear Algebra Appl. **114/115** (1989), 439–454; MR0986889 (90k:05095)]", *Linear Algebra Appl.* **313**:1-3 (2000), 101–105. MR 2001g:05070

[Merris 1995] R. Merris, "A survey of graph Laplacians", *Linear and Multilinear Algebra* **39**:1-2 (1995), 19–31. MR 97c:05104 Zbl 0832.05081

[Mitchell et al. 2010] L. H. Mitchell, S. K. Narayan, and A. M. Zimmer, "Lower bounds in minimum rank problems", *Linear Algebra Appl.* **432**:1 (2010), 430–440. MR 2010m:15004 Zbl 1220.05077

[Parsons and Pisanski 1989] T. D. Parsons and T. Pisanski, "Vector representations of graphs", *Discrete Math.* **78**:1-2 (1989), 143–154. MR 90k:05104 Zbl 0693.05058

[West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR 96i:05001 Zbl 0845.05001

rachel_cranfill@hmc.edu   *Department of Mathematics, Harvey Mudd College, Claremont, CA 91771, United States*

lmitchell2@vcu.edu   *Department of Mathematics, Virginia Commonwealth University, Richmond, VA 23284-2014, United States*

sivaram.narayan@cmich.edu   *Department of Mathematics, Central Michigan University, Mount Pleasant, MI 48859, United States*

tsutsuit@my.hiram.edu   *Department of Mathematics, Hiram College, Hiram, OH 44234, United States*

A new series for π via polynomial approximations to arctangent

Colleen M. Bouey, Herbert A. Medina and Erika Meza

(Communicated by Kenneth S. Berenhaut)

Using rational functions of the form

$$\left\{ \frac{t^{12m}(t - (2 - \sqrt{3}))^{12m}}{1 + t^2} \right\}_{m \in \mathbb{N}}$$

we produce a family of efficient polynomial approximations to arctangent on the interval $[0, 2 - \sqrt{3}]$, and hence provide approximations to π via the identity $\arctan(2 - \sqrt{3}) = \pi/12$. We turn the approximations of π into a series that gives about 21 more decimal digits of accuracy with each successive term.

1. Introduction

Two of the best-known series for π are

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{k=0}^{\infty} \frac{(4k)!\,(1103 + 26390k)}{(k!)^4\,396^{4k}},$$

devised by Ramanujan about a century ago (see [Baruah et al. 2007; 2009] for history), and

$$\frac{1}{\pi} = \frac{\sqrt{10005}}{4270934400} \sum_{k=0}^{\infty} \frac{(-1)^k (6k)!\,(13591409 + 545140134k)}{(3k)!\,(k!)^3},$$

from the 1980s [Chudnovsky and Chudnovsky 1988]. These series are interesting and important because they converge so rapidly. Indeed, the Ramanujan series gives about 6 more decimal places for π with each successive term and the Chudnovsky series about 13 more decimal places per term [Weisstein n.d.]. The Chudnovsky series was in fact the formula used recently by Yee and Kondo [2011] to compute 10

trillion digits of π, and a modified version of it is used by *Mathematica* to compute a large number of digits of π [Vardi 1991].

Here, in Theorem 2, we present a new series for π that yields about 21 more decimal places per term. The new series is derived from polynomial approximations to the classical arctangent function that come from the integration of rational functions.

2. Polynomial approximations to arctangent

The integration of certain rational functions has proven useful in the approximation of the classical arctangent function, and, because of identities such as $\arctan 1 = \pi/4$, these can produce approximations to π. For example, the family

$$\left\{ \frac{t^{4m}(t-1)^{4m}}{1+t^2} \right\}_{m \in \mathbb{N}}$$

was recently studied in [Medina 2006], where it is shown that it can be used to produce polynomial approximations to $\arctan x$ on the interval [0, 1] whose error is governed by the size of the rational functions on that interval. In this section, we use these methods to produce polynomial approximations to $\arctan x$ on a smaller interval where the size of the integrand is much smaller, and hence the approximations converge much faster.

Consider the sequence of rational functions

$$\frac{t^{a_n}(t-(2-\sqrt{3}))^{b_n}}{1+t^2},$$

where a_n and b_n are integers chosen so that the polynomial division yields a constant remainder, and hence after integration, the arctangent function. We use $2 - \sqrt{3}$ because $\arctan(2-\sqrt{3}) = \pi/12$; thus, if we can approximate arctangent at that value, we can approximate π.

Through trial and error, one finds that 12 is the smallest integer value of the b_n above that yields a constant remainder when the polynomial division is performed.[1] The smallest value for a_n is 2, but in what follows we choose 12 for the sake of symmetry. As Lemma 2 will show, the same is true for multiples of 12; thus, we explore the family of functions

$$\left\{ \frac{t^{12m}(t-\alpha)^{12m}}{1+t^2} \right\}_{m \in \mathbb{N}} \tag{1}$$

where we let $\alpha = 2 - \sqrt{3}$ to facilitate the notation.

---

[1] All computations were done using *Mathematica* 7.0.

The following two lemmas, whose proofs are immediate via initial computations and induction, will facilitate our exploration of the family of rational functions.

Lemma 1. *For any $m \in \mathbb{N}$,*

$$\frac{t^{12m}}{1+t^2} = t^{12m-2} - t^{12m-4} + t^{12m-6} - t^{12m-8} + \cdots - 1 + \frac{1}{1+t^2} = \sum_{n=0}^{6m-1}(-1)^{n+1}t^{2n} + \frac{1}{1+t^2}.$$

Lemma 2. *For any $m \in \mathbb{N}$,*

$$\frac{t^{12m}(t-\alpha)^{12m}}{1+t^2} = q_m(t) + \frac{r_m}{1+t^2}, \tag{2}$$

where $r_m = (-1)^m(4\alpha)^{6m} = (-1)^m(5533696 - 3194880\sqrt{3})^m$, and the q_m are polynomials given recursively by

$$q_m(t) = t^{12}(t-\alpha)^{12}q_{m-1}(t) + r_{m-1}q_1(t),$$

with the initial quotient

$$\begin{aligned}
q_1(t) = {}&-(4\alpha)^6 + (4\alpha)^6 t^2 - (4\alpha)^6 t^4 + (4\alpha)^6 t^6 - (4\alpha)^6 t^8 + (4\alpha)^6 t^{10} \\
&+ (9184097 - 5302440\sqrt{3})t^{12} + 12(564719\sqrt{3} - 978122)t^{13} \\
&+ (8113645 - 4684416\sqrt{3})t^{14} + 8(267909\sqrt{3} - 464032)t^{15} \\
&+ (1200770 - 693264\sqrt{3})t^{16} + 208(780\sqrt{3} - 1351)t^{17} \\
&+ (47554 - 27456\sqrt{3})t^{18} + 8(411\sqrt{3} - 712)t^{19} + (461 - 264\sqrt{3})t^{20} \\
&+ 12(\sqrt{3} - 2)t^{21} + t^{22}.
\end{aligned}$$

The following proposition provides a closed-form formula for the quotients.

Proposition 1. *For each $m \in \mathbb{N}$, define the polynomial quotient $q_m(t) = \sum_{n=0}^{24m-2} a_n t^n$ and the polynomial remainder $r_m \in \mathbb{R}$ via (2). Then*

(i) $a_{2n} = (-1)^{m+1+n}(4\alpha)^{6m}$ *and $a_{2n+1} = 0$ for $0 \le n \le 6m - 1$;*

(ii) $a_{24m-2} = 1$ *and $a_{24m-3} = -\binom{12m}{1}\alpha$ (these being the coefficients of the two highest powers of t in the quotient);*

(iii) $a_{24m-3-2n} = -a_{24m-3-2(n-1)} - \binom{12m}{2n+1}\alpha^{2n+1}$ *for $1 \le n \le 6m - 1$; and*

(iv) $a_{24m-2-2n} = -a_{24m-2-2(n-1)} + \binom{12m}{2n}\alpha^{2n}$ *for $1 \le n \le 6m - 1$.*

Proof. (i) We can rewrite and simplify the function to get

$$\frac{t^{12m}(t-\alpha)^{12m}}{1+t^2} = t^{12m}\left(\frac{(t-\alpha)^{12m}}{1+t^2}\right) = t^{12m}\left(p_m(t) + \frac{(-1)^m(4\alpha)^{6m}}{1+t^2}\right),$$

where $p_m(t)$ is some other quotient polynomial; we also note that Lemmas 1 and 2 together imply that the remainder $(-1)^m (4\alpha)^{6m}$ is indeed correct. Using Lemma 1, we make another substitution and obtain

$$t^{12m} p_m(t) + (-1)^m (4\alpha)^{6m} \left(t^{12m-2} - t^{12m-4} + t^{12m-6} - t^{12m-8} + \cdots - 1 + \frac{1}{1+t^2} \right),$$

which is the result of (i).

(ii) We write $\dfrac{t^{12m}(t-\alpha)^{12m}}{1+t^2} = \dfrac{t^{12m}}{1+t^2}(t-\alpha)^{12m}$. Use Lemma 1 to obtain

$$\left(t^{12m-2} - t^{12m-4} + - \cdots - 1 + \frac{1}{1+t^2} \right)(t-\alpha)^{12m},$$

and the binomial theorem to arrive at

$$\left(t^{12m-2} - t^{12m-4} + - \cdots - 1 + \frac{1}{1+t^2} \right) \sum_{k=0}^{12m} \binom{12m}{k} t^k \alpha^{12m-k} (-1)^k. \quad (3)$$

The coefficients of the two highest powers of t will come from multiplying the two highest powers of t in $(t-\alpha)^{12m}$ with t^{12m-2} in the first factor above.

(iii) To find each new odd coefficient we take the coefficient of the previous highest-order odd term and pair it with one lower power of t on the left of (3); since the signs of t alternate, we negate this. Each new coefficient will have a new lower-order term from the right paired with the highest power on the left. Adding these two, we get the coefficients of the new odd power of t.

(iv) The same argument as in (iii) gives the coefficients of the even powers. □

Since the functions (1) are small in the interval $[0, \alpha]$, integration of (2), after division by r_m, will yield approximations to arctangent on $[0, \alpha]$. That is,

$$\frac{1}{r_m} \int_0^x \frac{t^{12m}(t-\alpha)^{12m}}{1+t^2} \, dt = \frac{1}{r_m} \int_0^x q_m(t) \, dt + \arctan x, \quad (4)$$

and hence

$$P_m(x) = \frac{-1}{r_m} \int_0^x q_m(t) \, dt$$

will approximate arctangent on $[0, \alpha]$ with the error of the approximation given by the integral on the left side of (4), the maximum error occurring when $x = \alpha$. Proposition 1 provides a way to directly compute (after integration) these approximating polynomials; we will provide examples after we analyze their accuracy.

Substituting the largest and smallest values of t into the denominator of the left side of (4), we arrive at the inequality

$$\frac{1}{r_m} \int_0^\alpha \frac{t^{12m}(t-\alpha)^{12m}}{1+\alpha^2} \, dt < \frac{1}{r_m} \int_0^\alpha \frac{t^{12m}(t-\alpha)^{12m}}{1+t^2} \, dt < \frac{1}{r_m} \int_0^\alpha t^{12m}(t-\alpha)^{12m} \, dt. \quad (5)$$

It is now evident that, to further analyze the approximation, we need to compute

$$I_m := \int_0^\alpha t^{12m}(t-\alpha)^{12m}\,dt.$$

This is done via repeated integration by parts:

$$I_m = \int_0^\alpha t^{12m}(t-\alpha)^{12m}\,dt = \frac{((12m)!)^2}{(24m+1)!}\,\alpha^{24m+1}. \tag{6}$$

Since, as already noted, the left side of (4) is the error when $P_m(x)$ approximates $\arctan x$ on $[0,\alpha]$, we will use

$$e_m = \frac{1}{r_m}\int_0^\alpha \frac{t^{12m}(t-\alpha)^{12m}}{1+t^2}\,dt;$$

that is, e_m denotes the error when $P_m(\alpha)$ is used to approximate $\arctan\alpha = \pi/12$. Using this notation, we use (5) with m and $m+1$ to get

$$\frac{1}{(1+\alpha^2)r_m}I_m < e_m < \frac{1}{r_m}I_m \quad \text{and} \quad \frac{1}{(1+\alpha^2)r_{m+1}}I_{m+1} < e_{m+1} < \frac{1}{r_{m+1}}I_{m+1}. \tag{7}$$

Combining these two inequalities we arrive at

$$\frac{e_{m+1}}{e_m} < \frac{(1+\alpha^2)\,r_m\,I_{m+1}}{r_{m+1}\,I_m}, \tag{8}$$

which provides the estimate on how much better the next iterate is compared to the previous one.

Theorem 1. *Define* $e_m = \left|\pi/12 - P_m(\alpha)\right|$, *the error produced in approximating* $\pi/12$ *by the m-th iterate of the new sequence of approximating polynomials. Then, as* $m \to \infty$,

$$\frac{e_{m+1}}{e_m} < \frac{\alpha^{19}}{2^{34}} \approx 7.9063628967 \times 10^{-22} = 0.000000000000000000000079063628967.$$

That is, each iterate gives about 21 more decimal places of accuracy in approximating $\pi/12$.

Proof. Use $|r_m| = (4\alpha)^6$, $1+\alpha^2 = 4\alpha$, (6) and (8) to get

$$\begin{aligned}
\frac{e_{m+1}}{e_m} &< \frac{((12(m+1))!)^2\alpha^{24(m+1)+1}}{(4\alpha)^{6(m+1)}(24(m+1)+1)!} \cdot \frac{(4\alpha)^{6m+1}(24m+1)!}{((12m)!)^2\,\alpha^{24m+1}} \\
&= \frac{((12m+12)(12m+11)\cdots(12m+1))^2\alpha^{24}}{(4\alpha)^5\,(24m+25)(24m+24)\cdots(24m+2)}.
\end{aligned}$$

As $m \to \infty$, this becomes

$$\frac{(12^{12}\,m^{12})^2\alpha^{24}}{4^5\,\alpha^5\,24^{24}\,m^{24}} = \frac{\alpha^{19}}{4^5\,2^{24}} = \frac{\alpha^{19}}{2^{34}}. \qquad \square$$

Example 1. We use the coefficient formulas of Proposition 1 to find approximating polynomials. With $m = 1$,

$$P_1(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11} + \frac{(419-60\sqrt{3})x^{13}}{4096} - \frac{3(362-209\sqrt{3})x^{14}}{14336}$$
$$- \frac{(2916\sqrt{3}-955)x^{15}}{61440} - \frac{(172-99\sqrt{3})x^{16}}{8192} + \frac{(1255+468\sqrt{3})x^{17}}{34816} - \frac{13x^{18}}{4608}$$
$$- \frac{13(61+36\sqrt{3})x^{19}}{38912} - \frac{(172+99\sqrt{3})x^{20}}{10240} + \frac{(5051+2916\sqrt{3})x^{21}}{86016}$$
$$- \frac{3x^{22}}{22528(2-\sqrt{3})^5} + \frac{x^{23}}{94208(2-\sqrt{3})^6}.$$

Then

$$P_1(2-\sqrt{3}) = \frac{57423810140 - 22529108583\sqrt{3}}{70291415040},$$

and numerically we verify that $\left| P_1(2-\sqrt{3}) - \pi/12 \right| < 4.81587 \times 10^{-23}$, or, after multiplication by 12,

$$\left| \frac{57423810140 - 22529108583\sqrt{3}}{5857617920} - \pi \right| < 5.779054023 \times 10^{-22}.$$

Example 2. With $m = 2$,

$$P_2(\alpha) = \frac{30139322553723151897709935 - 115536316730168692893216 6\sqrt{3}}{38685520120050598123929 60},$$

and $\left| P_2(\alpha) - \pi/12 \right| \approx 2.55 \times 10^{-44}$.

3. Converting the iteration into a series

Theorem 1 requires the computation of a new set of polynomial coefficients when we want to obtain an approximation to π with more accuracy. For example, if we have a polynomial that gives n digits of accuracy for π when evaluated at α, then we need to compute a whole new polynomial, and hence a new set of coefficients, in order to obtain $(n + 21)$ more digits of accuracy. Following a technique first developed in [Dalzell 1944] and used recently in [Lucas 2009] to produce a rational series that gives 3–4 more decimal places of accuracy for π with each successive term, we now focus on developing a series that provides the same number of digits (i.e., about 21) per term in computing π as each iteration of the polynomial sequence.

We know that

$$\frac{t^{12}(t-\alpha)^{12}}{1+t^2} = q_1(t) - \frac{(4\alpha)^6}{1+t^2}, \tag{9}$$

which can be rewritten as

$$\frac{1}{1+t^2} = \frac{q_1(t)}{t^{12}(t-\alpha)^{12} + (4\alpha)^6}.$$

Next we factor out $(4\alpha)^6$ on the denominator to get

$$\frac{1}{1+t^2} = \frac{q_1(t)}{(4\alpha)^6} \cdot \frac{1}{1 + \left(\frac{t(t-\alpha)}{2\sqrt{\alpha}}\right)^{12}}.$$

Expanding the right side in a geometric series gives

$$\frac{1}{1+t^2} = \left(\frac{q_1(t)}{(4\alpha)^6}\right) \sum_{n=0}^{\infty} (-1)^n \left(\frac{t(t-\alpha)}{2\sqrt{\alpha}}\right)^{12n}. \tag{10}$$

We integrate both sides on $[0, \alpha]$ and bring the integral inside the sum to get

$$\arctan \alpha = \frac{1}{(4\alpha)^6} \sum_{n=0}^{\infty} \frac{(-1)^n}{(4\alpha)^{6n}} \int_0^{\alpha} q_1(t)\, t^{12n} (t-\alpha)^{12n}\, dt. \tag{11}$$

The polynomial $q_1(t)$ is of degree 22 so we need to compute integrals of the form

$$\int_0^{\alpha} t^{12n+k} (t-\alpha)^{12n}\, dt$$

for $k = 0, \ldots, 22$. This is done using repeated integration by parts; we get

$$\int_0^{\alpha} t^{12n+k} (t-\alpha)^{12n}\, dt = \frac{(12n+k)!\,(12n)!\,\alpha^{24n+k+1}}{(24n+k+1)!}. \tag{12}$$

If we write $q_1(t) = \sum_{k=0}^{22} a_k t^k$, then

$$\frac{\pi}{12} = \frac{1}{(4\alpha)^6} \sum_{n=0}^{\infty} \frac{(-1)^n\, \alpha^{18n+1}\,(12n)!}{4^{6n}} \sum_{k=0}^{22} a_k \frac{(12n+k)!\alpha^k}{(24n+k+1)!}. \tag{13}$$

Simplification of the inside sum leads to the following theorem.

Theorem 2. *We have*

$$\pi = \sum_{n=0}^{\infty} \frac{(-1)^n\,(2-\sqrt{3})^{18n+1}\,((12n)!)^2 (p_1(n) + p_2(n)\sqrt{3})}{2^{12(n+1)-1}(24n+1)!\,q(n)}, \tag{14}$$

where

$$p_1(n) = 293063424013062144n^{11} + 1743144635880815616n^{10}$$
$$+ 4603477509110094336n^9 + 7113505268868220800n^8$$
$$+ 7133195052290432592n^7 + 4863768060244254588n^6$$
$$+ 2295600628029058188n^5 + 747948981593488485n^4$$
$$+ 164336063152773014n^3 + 23098444048852896n^2$$
$$+ 1859706966144526n + 64510302034815,$$

$$p_2(n) = 92656102528843776n^{11} + 553643573938200576n^{10}$$
$$+ 1466739601852815360n^9 + 2269385610499169280n^8$$
$$+ 2272991576208150528n^7 + 1542973536047871648n^6$$
$$+ 721853379546109560n^5 + 231741816550236960n^4$$
$$+ 49765271182018546n^3 + 6762629909208426n^2$$
$$+ 519049199193830n + 16879034409510, \quad and$$

$$q(n) = 18786186952704n^{11} + 111934363926528n^{10} + 295980289228800n^9$$
$$+ 457648310845440n^8 + 458818030927872n^7 + 312432825729024n^6$$
$$+ 147050553999360n^5 + 47683923189760n^4 + 10399859469824n^3$$
$$+ 1446143661248n^2 + 114720643240n + 3904125225.$$

Moreover, if we define the error between the m-th partial sum of the series and π by $e_m = |\pi - S_m|$, then, as $m \to \infty$,

$$\frac{e_{m+1}}{e_m} < \frac{(2-\sqrt{3})^{19}}{2^{34}} \approx 7.9063628967 \times 10^{-22}.$$

Proof. Because of Theorem 1, it suffices to show that

$$\left| \frac{1}{(4\alpha)^6} \sum_{n=m}^{\infty} \frac{(-1)^n}{(4\alpha)^{6n}} \int_0^{\alpha} q_1(t)\, t^{12n}(t-\alpha)^{12n}\, dt \right| = \left| \frac{1}{r_m} \int_0^{\alpha} \frac{t^{12m}(t-\alpha)^{12m}}{1+t^2}\, dt \right|. \quad (15)$$

Using (9) to substitute for $q_1(t)$ and interchanging integration and summation in (15), we obtain

$$\frac{1}{(4\alpha)^6} \int_0^{\alpha} \sum_{n=m}^{\infty} \frac{(-1)^n}{(4\alpha)^{6n}} \left(\frac{t^{12n}(t-\alpha)^{12n}}{1+t^2} \right) (t^{12}(t-\alpha)^{12} + (4\alpha)^6)\, dt,$$

which we can simplify to

$$\frac{1}{(4\alpha)^6} \int_0^{\alpha} \left(\frac{t^{12}(t-\alpha)^{12} + (4\alpha)^6}{1+t^2} \right) \sum_{n=m}^{\infty} \left(\frac{(-1)t^{12}(t-\alpha)^{12}}{(4\alpha)^6} \right)^n\, dt.$$

The sum is a geometric series; after simplification, we get (15), as desired.   □

The new series (14) gives about 21 more decimal places of accuracy with each successive term, though the terms are significantly more complicated and hence more "computationally expensive" than those in either the Ramanujan and Chudnovsky series. We note that all three series require the computation of a single square root, but the powers of $2 - \sqrt{3}$ in the new series do slow down numerical computations. Thus, at this stage, it is fair to say that the Chudnovsky series still provides the fastest numerical tool for computing large numbers of digits of π. Nevertheless, it should be noted that the series (14) is very easy to program (in any language) and provides a viable method for computing digits of π; in fact, we have used it to compute a million digits on a desktop computer.

4. Further remarks

A similar process can be used with the rational functions

$$\left\{ \frac{t^{4m}(t - 1/\sqrt{3})^{6m}}{1 + t^2} \right\}_{m \in \mathbb{N}}$$

to produce polynomial approximations to arctangent on the interval $[0, 1/\sqrt{3}]$, and hence approximations to π, because $\arctan(1/\sqrt{3}) = \pi/6$. These approximations yield 5–6 more decimal places of accuracy with each iteration, and the computations are significantly "less expensive" than those of the sequence herein. (Our research in fact began with the exploration of this other family.)

It is our opinion that the series (14) should be seen as a byproduct of the approximating polynomials P_m which provide good approximations to arctangent on the entire interval $[0, 2 - \sqrt{3}]$. It is possible that the P_m could prove useful for approximating π when used in conjunction with multiple-angle identities such as $\pi/4 = 5 \arctan \frac{1}{7} + 2 \arctan \frac{3}{79}$ [Calcut 2009].

Acknowledgments

References

[Baruah et al. 2007] N. D. Baruah, B. C. Berndt, and H. H. Chan, "Ramanujan's series for $1/\pi$: a survey", *Math. Student* Special Centenary Volume (2007), 1–24. MR 2010d:11151

[Baruah et al. 2009] N. D. Baruah, B. C. Berndt, and H. H. Chan, "Ramanujan's series for $1/\pi$: a survey", *Amer. Math. Monthly* **116**:7 (2009), 567–587. MR 2549375 Zbl 1229.11162

[Calcut 2009] J. S. Calcut, "Gaussian integers and arctangent identities for π", *Amer. Math. Monthly* **116**:6 (2009), 515–530. MR 2010f:11182 Zbl 1229.11164

[Chudnovsky and Chudnovsky 1988] D. V. Chudnovsky and G. V. Chudnovsky, "Approximations and complex multiplication according to Ramanujan", pp. 375–472 in *Ramanujan revisited* (Urbana-Champaign, 1987), Academic Press, Boston, 1988. MR 89f:11099

[Dalzell 1944] D. P. Dalzell, "On 22/7", *J. London Math. Soc.* **19** (1944), 133–134. MR 7,152b Zbl 0060.15306

[Lucas 2009] S. K. Lucas, "Approximations to π derived from integrals with nonnegative integrands", *Amer. Math. Monthly* **116**:2 (2009), 166–172. MR 2478060 Zbl 1179.26008

[Medina 2006] H. A. Medina, "A sequence of polynomials for approximating arctangent", *Amer. Math. Monthly* **113**:2 (2006), 156–161. MR 2006m:41004 Zbl 1132.41304

[Vardi 1991] I. Vardi, *Computational recreations in Mathematica*, Addison-Wesley, Redwood City, CA, 1991. MR 93e:00002 Zbl 0786.11002

[Weisstein n.d.] E. W. Weisstein, "Pi formulas", web page, Wolfram Research MathWorld, Available at http://mathworld.wolfram.com/PiFormulas.html.

[Yee and Kondo 2011] A. J. Yee and S. Kondo, "Round 2... 10 trillion digits of pi", Report, 2011, Available at http://www.numberworld.org/misc_runs/pi-10t/details.html.

cbouey@lion.lmu.edu   *Mathematics Department, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, United States*

hmedina@lmu.edu   *Mathematics Department, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, United States*

emeza2@lion.lmu.edu   *Mathematics Department, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, United States*

msp

A mathematical model of biocontrol of invasive aquatic weeds

John Alford, Curtis Balusek, Kristen M. Bowers and Casey Hartnett

(Communicated by Suzanne Lenhart)

In this paper we modify the classical Lotka–Volterra differential equations to analyze competition between two aquatic plant species, a submersed plant and a free-floating plant. We formulate and analyze a system of three differential equations that control the dynamics of the free-floating plant biomass and both aboveground and belowground biomass for the submersed plant. We investigate our model to understand how plant competition is affected by grass carp herbivory on the submersed plant's aboveground biomass. We analyze both a reduced model, for which the submersed plant is assumed to have constant belowground biomass, and the full model. In each case, we compute stability of equilibria and derive a minimal grass carp stocking rate such that the free-floating plant may dominate the submersed plant. For the reduced model we show that the rate at which grass carp are stocked may exhibit a hysteresis effect.

1. Introduction

Hydrilla verticillata, commonly known as hydrilla, is one of the most invasive aquatic plants in the United States. Hydrilla has a rapid growth rate (as much as 1 inch per day), is typically found in depths up of 15–20 feet, and can grow to be 25 feet long in springs, lakes, marshes, ditches, rivers and tidal zones [Gettys et al. 2009]. Hydrilla is easily spread to a new body of water by just one leaf fragment attached to a boat. Millions of dollars a year are spent on efforts to control and eliminate hydrilla, including herbivory by grass carp and insects (e.g., leaf-mining flies), mechanical harvesters, herbicides, and competition with native aquatic plants [Gettys et al. 2009; Hanlon et al. 2000]. Thus understanding the biology and control of hydrilla is a problem of great significance.

Hydrilla is a submersed plant which is attached to the ground with an extensive root system, but may grow large enough so that its branches form dense mats of plant matter on the surface of the water [Gettys et al. 2009]. A free-floating

plant floats on the surface of the water and has roots that collect nutrients from the water and hang unanchored to the ground. An example of a free-floating plant is *Eichhornia crassipes*, commonly known as water hyacinth. Although water hyacinth is a nonnative, invasive species that must be carefully controlled, it has some desirable qualities. For example, it can be used to purify wastewater [Wolverton and McDonald 1979] and is often used as an ornamental plant for ponds and aquariums [Kay and Hoyle 2001].

When submersed plants and floating plants such as hydrilla and water hyacinth coexist they compete for light, space, and nutrients. The classic mathematical model of two species that compete for a common resource is the Lotka–Volterra differential equations [Edelstein-Keshet 2005; Zeeman 1995; Wangersky 1978]. In this paper we use the Lotka–Volterra competition model to formulate and analyze competition between a submersed plant and a free-floating plant.

Grass carp (or white amur) are fish that are native to rivers in Eastern Asia and may live up to 25 years and grow as much as 10 pounds per year [Gettys et al. 2009]. Large grass carp consume up to 30% of their body weight each day. One of the main biocontrol agents of hydrilla is the sterilized, triploid grass carp. In fact, the triploid grass carp will eat many types of aquatic weeds, but prefer submersed plants such as hydrilla when available [Cuda et al. 2008]. One study [Pine and Anderson 1991] found that given a choice of 12 different types of plants, the water hyacinth was the triploid grass carp's least preferred plant while the top three preferred plants were American pondweed, hydrilla, and elodea, each of which is a submersed plant species.

The rate at which grass carp should be stocked is an active area of research in aquatic plant management [Hanlon et al. 2000]. This rate depends on the feeding rate of the fish and the growth rate and quality of the plants, both of which are influenced by many factors [Cuda et al. 2008; Sutton et al. 2012]. Too few grass carp may be ineffective, whereas too many may completely eliminate all submersed aquatic plants. One study found that 25 to 30 grass carp per hectare of vegetation was necessary to control the undesirable vegetation while maintaining some amount of desirable vegetation [Hanlon et al. 2000]. The stocking rate of grass carp is often recommended based on the percentage of area that has been infested with the submersed plant [Hanlon et al. 2000; Sutton et al. 2012]. In our model we account for herbivory of the submersed plant by grass carp using a single parameter to control the stocking rate of grass carp. We use our model to determine the minimal stocking rate that may result in significant reduction or elimination of submersed plant biomass. The minimal stocking rate is expressed in terms of the relevant parameters that describe the ecosystem.

It is known that plant competition is influenced by herbivory [Van et al. 1998; Center et al. 2005; Tipping et al. 2009]. Our model shows that herbivory of

submersed plant aboveground biomass by grass carp may allow a free-floating plant to out-compete a submersed plant and proliferate. This is an example of the *principle of competitive exclusion* [Zeeman 1995; Wangersky 1978]. We show that, at a critical grass carp stocking rate, a stable ecosystem with large amounts of submersed plant biomass and no free-floating plant biomass may shift to a stable ecosystem with large amounts of free-floating plant biomass and small or no submersed plant biomass. This sudden shift in the stability of an ecosystem has been observed in lakes, coral reefs, woodlands, deserts, and oceans [Scheffer et al. 2001].

Mathematical models of competing aquatic plants and herbivore-plant ecosystems can be found throughout the literature. A model of free-floating and submersed plant dynamics is presented in [Scheffer et al. 2003], but aboveground and belowground biomass for the submersed plant is not distinguished. Competing aquatic plants are modeled in [Shukla 1998] when an undesirable plant is subjected to removal in order to promote the growth of the desirable plant. Experimental data is used in both of these papers to support the models, but neither use Lotka–Volterra dynamics and neither consider herbivory as a plant management strategy. Mathematical models of herbivore-plant dynamics are presented elsewhere, though. For example, in [Wilson et al. 2001] a model for the biocontrol of water hyacinth by insect (weevil) herbivory is considered. In [Gurney and Nisbet 1998], a two-variable Lotka–Volterra predator-prey food chain model is considered for which the herbivore is a predator and the plant is prey. Neither of these two publications model plant competition.

In this paper, we use existing models to formulate differential equations that control the dynamics of aboveground and belowground submersed plant biomass and free-floating plant biomass. We include Lotka–Volterra type competition between the free-floating plant and the aboveground submersed plant and a parameter that controls the mortality of the submersed plant aboveground biomass due to grass carp herbivory. Our paper is outlined as follows. In Section 2 we present the model and nondimensionalize the equations. In Section 2.1 we assume the submersed plant has a constant belowground biomass and analyze a reduced (two-equation) model. In Section 2.2, we consider the full model that incorporates the dynamics for both belowground and aboveground biomass of the submersed plant. In each section we present theoretical results that show how the equilibria and stability of equilibria depend on grass carp stocking rate. In the conclusion, the results are summarized and weaknesses of the model are discussed.

2. The model equations

The model equations are

$$\frac{dB}{dt} = sA - cB\left(1 - \frac{A}{m_A}\right) - d_B B, \qquad (2\text{-}1)$$

$$\frac{dA}{dt} = (cB + r_A A)\left(1 - \frac{A}{m_A}\right) - \alpha_1 AL - d_A A, \tag{2-2}$$

$$\frac{dL}{dt} = r_L L\left(1 - \frac{L}{m_L}\right) - \alpha_2 AL. \tag{2-3}$$

All of the parameters s, c, d_B, r_A, m_A, α_1, d_A, r_L, m_L, α_2 are nonnegative. Here A and B are (respectively) the aboveground and belowground biomass of the submersed plant species and L is the free-floating species biomass. In order to ensure biologically feasible solutions, initial data must be nonnegative. The growth dynamics of the submersed plant in the absence of L are given by the coupled equations (2-1) and (2-2), and for $d_A = 0$, the model is the same as the one in [Turchin 2003; Turchin and Batzli 2001]. The aboveground biomass growth equation (2-2) incorporates logistic growth in the absence of B and exponential growth (regrowth) from energy supplied by the belowground biomass in the absence of A. The parameter d_A in (2-2) controls the mortality of aboveground biomass of the submersed plant. The growth dynamics of the floating plant, given by (2-3), are logistic in the absence of A. Logistic growth has been experimentally verified as a good growth model for water hyacinth [Wilson et al. 2001; 2005]. Competition is modeled as the standard Lotka–Volterra type described in [Edelstein-Keshet 2005] with interaction terms proportional to AL. The competition coefficients α_1 and α_2 control the ability of each plant species to compete with the other and measure how efficient one species is compared to the other at capturing the shared resources.

The parameter d_A has dimensions $(\text{time})^{-1}$ and represents the number of grass carp that are stocked per unit time. As discussed in the introduction, grass carp prefer submersed plants when available and triploid grass carp are sterilized before stocking [Hanlon et al. 2000; Cuda et al. 2008; Pine and Anderson 1991]. Fish-eating predators such as otters and other fish may reduce the number of grass carp, but large grass carp are not affected by predation [Gettys et al. 2009] and grass carp may live 20 or more years [Cuda et al. 2008]. Thus our model assumes that grass carp do not feed on the free-floating plant, there is a limited timespan for biocontrol with large grass carp, and the natality and mortality of grass carp may be ignored.

In order to reduce the number of parameters and understand the important relationships between parameters, we nondimensionalize the model equations by introducing the *dimensionless* variables and parameters

$$x_1 = d_B B (s m_A)^{-1}, \quad y_1 = A m_A^{-1}, \quad x_2 = L m_L^{-1}, \quad \tau = r_L t, \tag{2-4}$$

$$\rho = cs(r_L d_B)^{-1}, \quad \delta_2 = d_B r_L^{-1}, \quad \phi = c r_L^{-1}, \quad \psi = r_A r_L^{-1}, \quad \delta_1 = d_A r_L^{-1}, \tag{2-5}$$

$$\theta_1 = \alpha_1 m_L r_L^{-1}, \quad \theta_2 = \alpha_2 m_A r_L^{-1}. \tag{2-6}$$

After substituting (2-4)–(2-6) into (2-1)–(2-3) we get the system

$$dx_1/d\tau = \delta_2(y_1 - x_1) - \phi x_1(1 - y_1), \tag{2-7}$$

$$dy_1/d\tau = (\rho x_1 + \psi y_1)(1 - y_1) - \theta_1 y_1 x_2 - \delta_1 y_1, \tag{2-8}$$

$$dx_2/d\tau = x_2(1 - x_2) - \theta_2 y_1 x_2. \tag{2-9}$$

Here the variable x_1 controls the (nondimensionalized) submerged plant below-ground biomass dynamics, y_1 controls the (nondimensionalized) submerged plant aboveground biomass dynamics, and x_2 controls the (nondimensionalized) floating plant biomass dynamics.

2.1. *Constant belowground biomass.*

In this section we assume that B is constant and analyze the regrowth model for the submersed plant in the absence of logistic growth as in [Gurney and Nisbet 1998]. Here we replace ρx_1 with a constant β to get

$$dy_1/d\tau = \beta(1 + \psi\beta^{-1}y_1)(1 - y_1) - \theta_1 y_1 x_2 - \delta_1 y_1,$$

for (2-8). We will make the additional assumption that there is a significant amount of belowground biomass and $\psi \ll \beta$. Then these simplifications with (2-8), (2-9) give the system

$$y_1' = \beta(1 - y_1) - \theta_1 y_1 x_2 - \delta y_1, \quad x_2' = x_2(1 - x_2) - \theta_2 y_1 x_2, \tag{2-10}$$

where we have replaced δ_1 with δ, and the prime denotes differentiation with respect to the dimensionless time variable τ. The equilibria are constant solutions and are found by solving the algebraic system that results by setting the right sides of each equation in (2-10) to zero. The long-term behavior of a dynamical system may be determined by equilibria and initial conditions. In general, initial conditions that are close enough to a stable equilibrium will yield solutions that evolve in time to these equilibria. In the remainder of this paper, we perform standard equilibrium and local stability analysis of nonlinear differential equations [Edelstein-Keshet 2005; Strogatz 2001].

For the equilibrium computations, it will be convenient to define the quantities

$$\gamma = 1 + \delta\beta^{-1}, \quad \alpha = \theta_1\beta^{-1}. \tag{2-11}$$

We first consider a graphical analysis of the equilibria in the y_1-x_2 phase plane. The nullclines are curves along which either $y_1' = 0$ or $x_2' = 0$. These curves are

$$x_2 = (1 - \gamma y_1)/(\alpha y_1), \quad x_2 = 0, \quad x_2 = 1 - \theta_2 y_1, \tag{2-12}$$

where the first equation is the y_1-nullcline (when $y_1' = 0$) and the second two equations are the x_2-nullclines (when $x_2' = 0$). When the y_1-nullcline intersects either of the x_2-nullclines for $y_1 \geq 0$ and $x_2 \geq 0$, the point of intersection is an equilibrium. Substituting nonnegative values of y_1 and x_2 into the right side of (2-10) results in a vector field that describes the flow of (2-10) in the phase plane

(that is, the direction of increase or decrease of either y_1 or x_2). The flow along the y_1-nullcline is vertical and the flow along the x_2-nullcline is horizontal.

Figure 1 depicts example phase-plane plots. Each phase plane depends on parameter values. As can be seen from these plots, either one, two, or three equilibria exist. The free-floating plant extinction equilibrium along the $x_2 = 0$-nullcline when $y_1 = \gamma^{-1}$ exists for all parameter values. There may also be one or two equilibria where $x_2 > 0$ and $y_1 > 0$. These are the coexistence equilibria. Note that there are no submersed plant extinction equilibria when $y_1 = 0$. This is clear as we assumed that the belowground biomass is constant and positive.

Motivated by the phase-plane plots we will analyze the equilibria algebraically. We denote the equilibria as $(\hat{y}_1, \hat{x}_2)$. Substituting $\hat{x}_2 = 0$ from (2-12) into the first equation from (2-12) yields the free-floating plant extinction equilibrium

$$(\hat{y}_1, 0) = (\gamma^{-1}, 0). \tag{2-13}$$

Substituting $\hat{x}_2 = 1 - \theta_2 \hat{y}_1$ from (2-12) into the first equation from (2-12) gives a quadratic equation in $\hat{x}_2$ that yields

$$\hat{x}_2^\pm = (2\theta_1)^{-1}\big(\hat{\delta} - \delta \pm \sqrt{(\hat{\delta} - \delta)^2 + 4\theta_1(\delta - \delta_0)}\big), \quad \hat{y}_1^\pm = \theta_2^{-1}(1 - \hat{x}_2), \tag{2-14}$$

where

$$\hat{\delta} = \theta_1 - \beta \quad \text{and} \quad \delta_0 = \beta(\theta_2 - 1). \tag{2-15}$$

After substituting (2-15) into the radicand in (2-14), simple algebra yields

$$(\hat{\delta} - \delta)^2 + 4\theta_1(\delta - \delta_0) = (\delta + \theta_1 + \beta)^2 - 4\theta_1\theta_2\beta,$$

which is zero for two values of δ, one of which is negative as θ_1, θ_2, and β are positive. The radicand in (2-14) may have a positive zero for $\delta = \delta_c$, in which case we get that $\hat{x}_2^c = \hat{x}_2^+ = \hat{x}_2^-$, where

$$\delta_c = 2\sqrt{\theta_1\theta_2\beta} - \theta_1 - \beta, \quad \hat{x}_2^c = (2\theta_1)^{-1}(\hat{\delta} - \delta_c). \tag{2-16}$$

The constants $\hat{\delta}$, δ_0, and δ_c will be used to characterize the stability and existence of equilibria for (2-10). We consider all parameters except δ fixed and positive and $\delta \geq 0$. First, the floating plant equilibria $\hat{x}_2^\pm$ may be nonnegative and real-valued if and only if $\delta \geq \delta_c$ and $\hat{x}_2^+ = \hat{x}_2^-$ when $\delta = \delta_c$ and the radicand is zero. If $\delta > \delta_c$, $\hat{x}_2^+$ increases with δ while $\hat{x}_2^-$ decreases with δ. It is easy to show that $\delta_c \leq \delta_0$. If $\delta = \delta_0$, then either $\hat{x}_2^+$ or $\hat{x}_2^-$ equals zero depending on the sign of $\hat{\delta} - \delta_c$.

The dependence of $\hat{x}_2^\pm$ on δ may be plotted in the δ-$\hat{x}_2$ plane with all other parameters fixed. The resulting curve has the general shape of a parabola which opens to the right. Figure 2 depicts such curves for $\delta_0 > 0$ and two cases where $\delta_c < 0$, $\hat{x}_2^c < 0$ and $\delta_c > 0$, $\hat{x}_2^c > 0$.
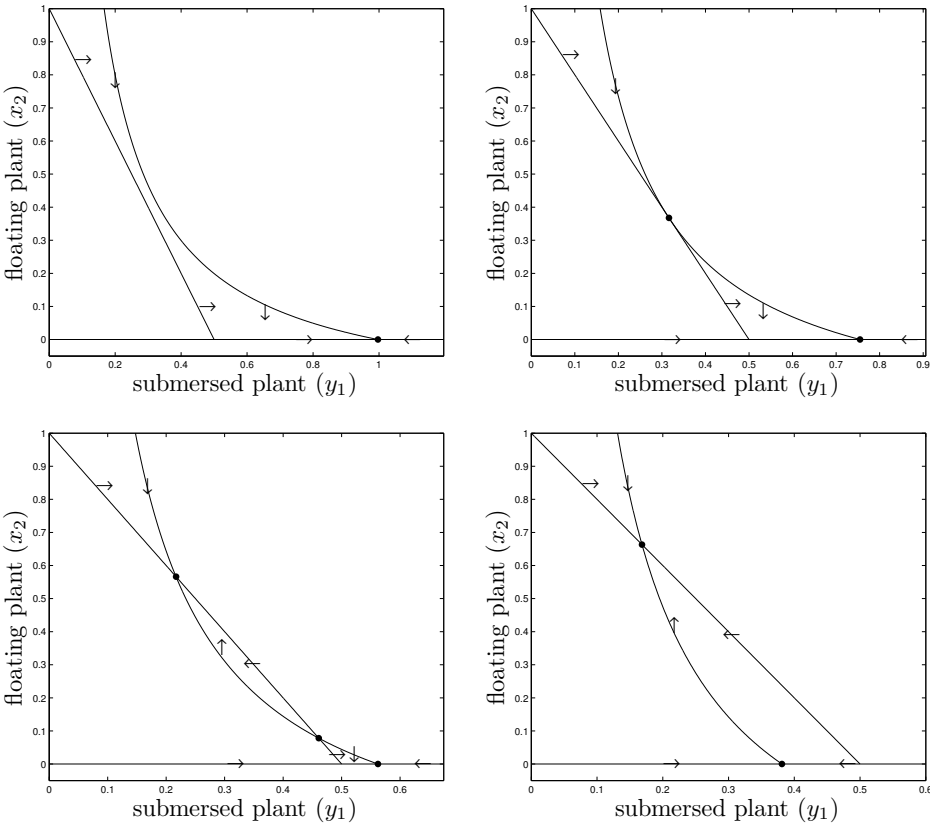
Figure 1. Plots of the y_1-x_2 phase plane for (2-10). The y_1-nullcline (curve) and x_2-nullclines (lines) are from (2-12). The arrows indicate the direction of flow of (2-10) along each nullcline. Equilibria are depicted at the dots where the y_1-nullcline intersects either of the x_2-nullclines. Each phase plane shows the free-floating plant extinction equilibrium at $(\gamma^{-1}, 0)$. There are no other equilibria in the top-left. The phase plane in the top-right shows a coexistence equilibrium for which the nonzero x_2-nullcline is tangential to the y_1-nullcline. The phase planes in the bottom show two (left) and one (right) coexistence equilibria where the x_2-nullcline intersects the y_1-nullcline.

The phase planes plotted in Figure 1 can be explained (qualitatively) by observing the equilibrium curve depicted in the right panel of Figure 2. First, recall that $\gamma = 1 + \beta^{-1}\delta$ defines the free-floating plant extinction equilibrium. Define the functions $f_1(y_1) = (1 - \gamma y_1)/(\alpha y_1)$ and $f_2(y_1) = 1 - \theta_2 y_1$ so that the y_1-nullcline is $x_2 = f_1(y_1)$ and the (nonzero) x_2-nullcline is $x_2 = f_2(y_1)$ from (2-12). If
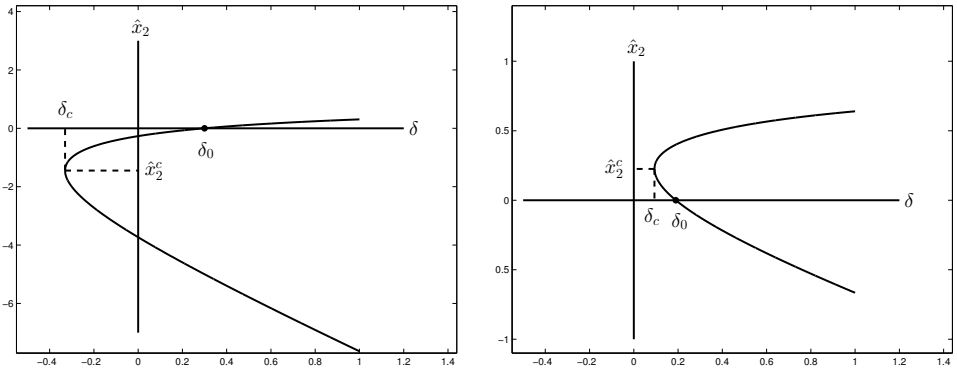
Figure 2. Plots of $\hat{x}_2^{\pm}$ as a function of δ from (2-14). The knee of the curve is $(\delta_c, \hat{x}_2^c)$ from (2-16). For the curve on the left, $\hat{\delta} < \delta_c < 0$, and for the curve on the right, $0 < \delta_c < \hat{\delta}$. The top half of each curve ($\hat{x}_2 > \hat{x}_2^c$) is $\hat{x}_2 = \hat{x}_2^+$ while the bottom half of each curve ($\hat{x}_2 < \hat{x}_2^c$) is $\hat{x}_2 = \hat{x}_2^-$.

$0 \le \delta < \delta_c < \hat{\delta}$, then $f_1(y_1)$ does not intersect $f_2(y_1)$ and the free-floating plant extinction equilibrium is unique. In this case, $0 < \delta < \delta_c$ so that δ is below the knee of the curve in the right panel of Figure 2.

If δ is then increased until $\delta = \delta_c$, then $f_2(y_1)$ is tangent to $f_1(y_1)$ and $f_1(y_1) = f_2(y_1)$ for exactly one value of y_1. This is displayed in the phase plane in the top-right in Figure 1 and corresponds to the knee of the curve in the right panel of Figure 2 where $\delta = \delta_c$ and $\hat{x}_2^- = \hat{x}_2^+ = \hat{x}_2^c$. As δ is increased further, both $\hat{x}_2^+$ and $\hat{x}_2^-$ are real and positive with $\hat{x}_2^- < \hat{x}_2^+$. This corresponds to the phase plane in the bottom-left in Figure 1 and the interval $\delta_c < \delta < \delta_0$ in the right panel of Figure 2. As δ continues to increase until $\delta > \delta_0$ and $\hat{x}_2^- < 0$, there is a single feasible positive equilibrium given by $\hat{x}_2^+$. This corresponds to the phase plane in the bottom-right in Figure 1 and the interval $\delta > \delta_0$ in the right panel of Figure 2.

In order to analyze local stability of the equilibria we compute the linearized stability (Jacobian) matrix for (2-10) which is given by

$$J(\hat{y}_1, \hat{x}_2) = \begin{pmatrix} -\beta - \theta_1 \hat{x}_2 - \delta & -\theta_1 \hat{y}_1 \\ -\theta_2 \hat{x}_2 & 1 - \theta_2 \hat{y}_1 - 2\hat{x}_2 \end{pmatrix}. \tag{2-17}$$

The eigenvalues λ of this matrix satisfy the characteristic equation

$$\lambda^2 - \mathrm{tr}(J(\hat{y}_1, \hat{x}_2))\lambda + \det(J(\hat{y}_1, \hat{x}_2)) = 0.$$

Standard theory [Edelstein-Keshet 2005; Strogatz 2001] is that a necessary and sufficient condition for stability of $(\hat{y}_1, \hat{x}_2)$ is that the eigenvalues of the Jacobian

have negative real parts or

$$\text{tr}(J(\hat{y}_1, \hat{x}_2)) < 0 \quad \text{and} \quad \det(J(\hat{y}_1, \hat{x}_2)) > 0. \tag{2-18}$$

Substituting the free-floating plant extinction equilibrium $\hat{y}_1 = \gamma^{-1}$ and $\hat{x}_2 = 0$ into (2-17) gives

$$\text{tr}(J(\gamma^{-1}, 0)) = 1 - \beta - \delta - \theta_2\gamma^{-1}, \tag{2-19}$$

$$\det(J(\gamma^{-1}, 0)) = -(1 - \theta_2\gamma^{-1})(\beta + \delta). \tag{2-20}$$

Comparing (2-18) and (2-19), (2-20) shows that $(\gamma^{-1}, 0)$ is stable if and only if $\theta_2 > \gamma$ which is equivalent to $\delta < \delta_0$ from (2-15).

We next consider stability of the equilibria $(\hat{y}_1^+, \hat{x}_2^+)$ and $(\hat{y}_1^-, \hat{x}_2^-)$ where we assume $\hat{x}_2^- > 0$. Substitute $\hat{x}_2 = 1 - \theta_2\hat{y}_1$ and (2-17) reduces to

$$J(\hat{y}_1, \hat{x}_2) = \begin{pmatrix} -\beta - \theta_1\hat{x}_2 - \delta & -\theta_1\hat{y}_1 \\ -\theta_2\hat{x}_2 & -\hat{x}_2 \end{pmatrix}, \tag{2-21}$$

so that

$$\text{tr}(J(\hat{y}_1, \hat{x}_2)) = -\beta - \delta - \hat{x}_2(1 + \theta_1), \tag{2-22}$$

$$\det(J(\hat{y}_1, \hat{x}_2)) = \hat{x}_2[\beta + \delta - \theta_1\theta_2\hat{y}_1 + \theta_1\hat{x}_2]. \tag{2-23}$$

It is clear in this case that $\text{tr}(J(\hat{y}_1, \hat{x}_2)) < 0$ as $\hat{x}_2$, δ, β, and θ_1 are all positive. Substitute $\theta_2\hat{y}_1 = 1 - \hat{x}_2$ and, after some algebra, we get that a necessary and sufficient condition for $\hat{x}_2 > 0$ and $\det(J(\hat{y}_1, \hat{x}_2)) > 0$ is $\hat{x}_2 > (1 - \gamma\beta\theta_1^{-1})/2 = (2\theta_1)^{-1}(\hat{\delta} - \delta)$. Thus, if $\delta > \delta_c$ from (2-16), then $\hat{x}_2^+$ is stable and $\hat{x}_2^-$ is unstable.

Table 1 summarizes the conditions on $\delta > 0$ for the existence of equilibria for (2-10) and their (linearized) stability properties. The pair $(\delta_c, \hat{x}_2^c)$ describes the point in the δ-$\hat{x}_2$ plane at the knee of the equilibrium curve when $\hat{x}_2^\pm$ is plotted as a function of δ, as in Figure 2. The first three rows correspond to $\delta_c > \hat{\delta}$ so that the knee of the equilibrium curve is below the δ-axis in the δ-$\hat{x}_2$ plane as depicted in the left panel in Figure 2. The middle three rows correspond to $0 < \delta_c < \hat{\delta}$ and the knee of the equilibrium curve is in the top-right quadrant of the δ-$\hat{x}_2$ plane as in the right panel in Figure 2. For the last three rows $\delta_c < \hat{\delta}$ and $\delta_c < 0$ so that the knee of the equilibrium curve is in the top-left quadrant of the δ-$\hat{x}_2$ plane.

Inspection of the middle three rows of Table 1 shows that when δ_c and $\hat{x}_2^c$ are both positive, as in Figure 2, right, equilibria $(\hat{x}_2^\pm, \hat{y}_2^\pm)$ are created as δ increases through δ_c. This indicates a saddle-node bifurcation [Strogatz 2001] at $\delta = \delta_c$. In this case, there is a simple zero eigenvalue for the Jacobian matrix (2-17) for which $\text{tr}(J(\hat{y}_1, \hat{x}_2)) < 0$ and $\det(J(\hat{y}_1, \hat{x}_2)) = 0$. The bifurcation diagram, plotted in Figure 3, shows $\hat{y}_1$ vs. δ and $\hat{x}_2$ vs. δ and the stability properties of these equilibria.

$(\delta_c, \hat{x}_2^c)$	δ_0	δ	$(\gamma^{-1}, 0)$	$(\hat{y}_1^+, \hat{x}_2^+)$	$(\hat{y}_1^-, \hat{x}_2^-)$
$(-,-)$	$\delta_0 < 0$	$\delta > 0$	unstable	stable	not feasible
$(-,-)$ or $(+,-)$	$\delta_0 > 0$	$0 < \delta < \delta_0$	stable	not feasible	not feasible
$(-,-)$ or $(+,-)$	$\delta_0 > 0$	$\delta > \delta_0$	unstable	stable	not feasible
$(+,+)$	$\delta_0 > 0$	$0 < \delta < \delta_c$	stable	does not exist	does not exist
$(+,+)$	$\delta_0 > 0$	$\delta_c < \delta < \delta_0$	stable	stable	unstable
$(+,+)$	$\delta_0 > 0$	$\delta > \delta_0$	unstable	stable	not feasible
$(-,+)$	$\delta_0 < 0$	$\delta > 0$	unstable	stable	not feasible
$(-,+)$	$\delta_0 > 0$	$0 < \delta < \delta_0$	stable	stable	unstable
$(-,+)$	$\delta_0 > 0$	$\delta > \delta_0$	unstable	stable	not feasible

Table 1. A summary of existence and stability properties of the equilibria from (2-13) and (2-14) as they depend on $\delta > 0$. *Stable* and *unstable* indicate existence of a positive equilibrium whereas *not feasible* indicates the equilibrium exists, but is negative. The constants δ_c, $\hat{x}_2^c$, and δ_0 are given by (2-15) and (2-16).

Figure 3 displays a hysteresis effect. If the free-floating plant is extinct so that $(\hat{y}_1, \hat{x}_2) = (\gamma^{-1}, 0)$ and δ is increased through $\delta = \delta_0$, the free-floating plant extinction equilibrium loses stability. Any small perturbation from the extinction equilibrium (for example, a small remnant of free-floating plant attached to a boat is introduced into the lake) will cause a jump in the ecosystem to the stable coexistence equilibrium $(\hat{y}_1^+, \hat{x}_2^+)$. If $(\hat{y}_1, \hat{x}_2) = (\hat{y}_1^+, \hat{x}_2^+)$ and δ is then decreased, the system does not restabilize to the free-floating plant extinction equilibrium until $\delta = \delta_c$ at the saddle-node bifurcation.

Figure 4 shows simulations of the system (2-10). The parameters obey the middle three rows of Table 1 corresponding to the bifurcation diagram that is plotted in Figure 3. In this case, solutions for $\delta < \delta_c$ quickly (approximately 30 time units) achieve equilibrium at $(\gamma^{-1}, 0)$, while solutions for $\delta > \delta_0$ achieve equilibrium at to $(\hat{y}_1^+, \hat{x}_2^+)$ after approximately 100 time units.

In order to draw meaningful biological conclusions from the analysis, the dimensional forms of the equations and parameters must be considered. The nondimensionalizations are specified in (2-4), (2-5), and (2-6). Table 1 shows that for $\delta > \delta_0$ the free-floating plant extinction equilibrium is unstable. Using (2-5), (2-6), and (2-15), this inequality becomes

$$d_A > cBr_L^{-1}(\alpha_2 m_A r_L^{-1} - 1), \qquad (2\text{-}24)$$

where δ replaced δ_1 in (2-5). That is, the mortality of the aboveground biomass (d_A) should be larger than the production of belowground biomass (cB) scaled by a factor which increases with the competition efficiency of the submersed plant (α_2)
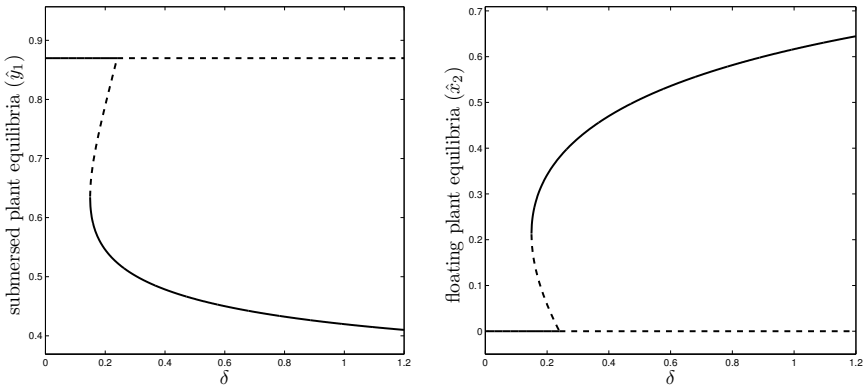
Figure 3. Bifurcation curves in the δ-$\hat{y}_1$ plane (left) and δ-$\hat{x}_2$ plane (right) for $\delta \geq 0$, $\hat{y}_1 \geq 0$, and $\hat{x}_2 \geq 0$ where $\beta = 1$, $\theta_1 = 2$, and $\theta_2 = 1.24$. Stable equilibria are plotted solid whereas unstable equilibria are plotted dashed. The submersed plant carrying capacity equilibria $\hat{y}_1 = (1 + \beta^{-1}\delta)^{-1}$ is the top curve in the left panel and the free-floating plant extinction equilibria $\hat{x}_2 = 0$ is the horizontal line in the right panel. The coexistence equilibria $\hat{y}_1 = \hat{y}_1^{\pm}$ make up the bottom curve (solid $\hat{y}_1^+$ and dashed $\hat{y}_1^-$) in the left panel and $\hat{x}_2 = \hat{x}_2^{\pm}$ make up the top curve (solid $\hat{x}_2^+$ and dashed $\hat{x}_2^-$) in the right panel. The coexistence equilibria coalesce when $\hat{y}_1^+ = \hat{y}_1^-$ and $\hat{x}_2^+ = \hat{x}_2^-$ at a saddle-node bifurcation for $\delta = \delta_c = 0.15$ from (2-16). Here $\delta_0 = 0.24$ and there is a region of bistability for $\delta_c < \delta < \delta_0$.

and the carrying capacity of aboveground biomass (m_A) and decreases with the growth rate of the free-floating plant (r_L). The minimal stocking rate is quantified by (2-24). Any plant management strategy that can reduce the right side of (2-24) results in a smaller number of grass carp necessary to destabilize the ecosystem towards free-floating plant dominance. If the quantity in parentheses can be made negative, for example by increasing the growth rate of r_L, grass carp will not be needed at all as the free-floating plant extinction equilibrium is stable for $\delta_A = 0$ (corresponding to row 1 and row 7 in Table 1 where $\delta_0 < 0$).

2.2. *Nonconstant belowground biomass.* In the previous section, the belowground biomass was assumed positive. This precludes the existence of a submersed plant extinction equilibrium. In this section we investigate the full model (2-7), (2-8), (2-9) and show that there is a stable submersed plant extinction equilibrium. As in the case for constant belowground biomass, there are multiple equilibria which will be denoted by $(\hat{x}_1, \hat{y}_1, \hat{x}_2)$ and which depend on the various parameters. Setting the
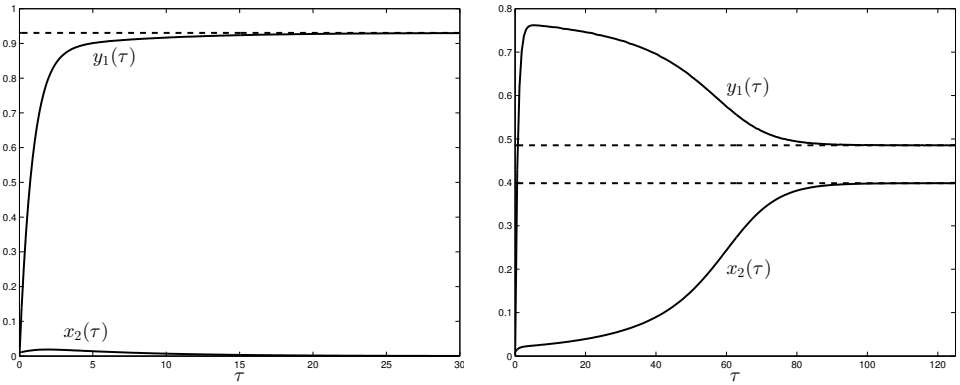
Figure 4. Simulations of the system (2-10) where the parameters are as in Figure 3 with $\beta = 1$, $\theta_1 = 2$, and $\theta_2 = 1.24$. For both plots the initial conditions are $(y_1(0), x_2(0)) = (0, 0.01)$. In the left panel, $\delta = 0.0748 < \delta_c$, and in the right panel, $\delta = 0.264 > \delta_0$. The dashed horizontal lines are the stable equilibria at $y_1 = (1+\beta^{-1}\delta)^{-1}$ and $x_2 = 0$ in the left panel and $y_1 = \hat{y}_1^+$ and $x_2 = \hat{x}_2^+$ in the right panel.

right side of (2-7) to zero yields

$$\hat{x}_1 = \hat{\delta}_2 \hat{y}_1 (1 + \hat{\delta}_2 - \hat{y}_1)^{-1}, \quad \hat{\delta}_2 = \phi^{-1}\delta_2. \tag{2-25}$$

If we next substitute (2-25) into the right side of (2-8) and use (2-9), then we get that the equilibria $\hat{y}_1$ and $\hat{x}_2$ obey

$$\hat{y}_1 \big([\psi(1 - \hat{y}_1) - \delta_1 - \theta_1 \hat{x}_2](\phi + \delta_2 - \phi\hat{y}_1) + \rho\delta_2(1 - \hat{y}_1)\big) = 0, \tag{2-26}$$

$$\hat{x}_2(1 - \hat{x}_2 - \theta_2 \hat{y}_1) = 0. \tag{2-27}$$

We first consider the case $\hat{y}_1 = 0$ and the submersed plant is extinct. This yields two possibilities. The case $(0, 0, 0)$ is extinction of both species and the case $(0, 0, 1)$ is extinction of the submersed plant with the free-floating plant at carrying capacity.

We now consider the equilibria such that $\hat{x}_1 > 0$, $\hat{y}_1 > 0$ and the submersed plant is not extinct. First, note that (2-25) implies that $\phi + \delta_2 - \phi\hat{y}_1 > 0$ and from (2-26) we see that the feasible equilibria must obey $0 < \hat{y}_1 < 1$ as all of the parameters are nonnegative. For the coexistence equilibria $\hat{x}_1 > 0$, $\hat{y}_1 > 0$, $\hat{x}_2 > 0$ and neither the submersed plant nor the free-floating plant is extinct. In this case, (2-27) gives that $\hat{x}_2 = 1 - \theta_2 \hat{y}_1$ and substituting this into (2-27) yields the equation

$$\nu\hat{y}_1^2 + (\xi - 1 - \nu(1 + \hat{\delta}_2))\hat{y}_1 + 1 + \hat{\delta}_2 - \xi\kappa = 0, \tag{2-28}$$

where

$$\nu = 1 - \theta_1\theta_2\psi^{-1}, \quad \xi = \psi^{-1}(\delta_1 + \theta_1 - \rho\hat{\delta}_2), \quad \kappa = 1 + \frac{(\delta_1 + \theta_1)\hat{\delta}_2}{\delta_1 + \theta_1 - \rho\hat{\delta}_2}. \quad (2\text{-}29)$$

We will use (2-28) in Theorem 1 to examine coexistence equilibria under a constrained parameter set.

In order to analyze stability of equilibria, we consider the Jacobian matrix $J(\hat{x}_1, \hat{y}_1, \hat{x}_2)$ which is given by

$$\begin{pmatrix} -\delta_2 - \phi(1 - \hat{y}_1) & \delta_2 + \phi\hat{x}_1 & 0 \\ \rho(1 - \hat{y}_1) & -\rho\hat{x}_1 - \theta_1\hat{x}_2 - \delta_1 + \psi - 2\psi\hat{y}_1 & -\theta_1\hat{y}_1 \\ 0 & -\theta_2\hat{x}_2 & 1 - 2\hat{x}_2 - \theta_2\hat{y}_1 \end{pmatrix}. \quad (2\text{-}30)$$

We will use (2-30) and the results of the equilibria computations to show the following theorem.

Theorem 1. *If $\delta_1 > \psi + \rho\hat{\delta}_2$ and $\theta_2 < \min\{1, \theta_1^{-1}\psi\}$, then $(0, 0, 1)$ is the only feasible stable equilibrium of* (2-7), (2-8), (2-9).

Proof. First consider the free-floating plant extinction equilibrium $(\hat{x}_1, \hat{y}_1, 0)$ where $\hat{x}_1 \geq 0$ and $\hat{y}_1 \geq 0$. The Jacobian from (2-30) is $J(\hat{x}_1, \hat{y}_1, 0)$ whose last row is the vector $(0, 0, 1 - \theta_2\hat{y}_1)$. Thus $J(\hat{x}_1, \hat{y}_1, 0)$ has one eigenvalue equal to $1 - \theta_2\hat{y}_1$. In this case, inspection of (2-25) and (2-26) yields that $0 \leq \hat{y}_1 < 1$ as all parameters are positive and all equilibria must be nonnegative. The assumption $\theta_2 < 1$ shows that $1 - \theta_2\hat{y}_1 > 0$ so that $(\hat{x}_1, \hat{y}_1, 0)$ is unstable.

We next consider $(0, 0, 1)$, the submersed plant extinction equilibrium when the free-floating plant is at carrying capacity. Substituting this into the Jacobian (2-30) results in the matrix

$$J(0, 0, 1) = \begin{pmatrix} -\delta_2 - \phi & \delta_2 & 0 \\ \rho & \psi - \theta_1 - \delta_1 & 0 \\ 0 & -\theta_2 & -1 \end{pmatrix}, \quad (2\text{-}31)$$

and the eigenvalues obey

$$(1 + \lambda)\left(\lambda^2 + (\theta_1 + \delta_1 - \psi + \delta_2 + \phi)\lambda + (\delta_2 + \phi)(\theta_1 + \delta_1 - \psi) - \rho\delta_0\right) = 0. \quad (2\text{-}32)$$

Thus $\lambda = -1$ or

$$\lambda = \left(-\gamma \pm \sqrt{\gamma^2 - 4[(\delta_2 + \phi)(\theta_1 + \delta_1 - \psi) - \rho\delta_0]}\right)/2, \quad (2\text{-}33)$$

where $\gamma = \theta_1 + \delta_1 - \psi + \delta_2 + \phi$ which is positive as it was assumed that $\delta_1 > \psi$. Therefore, nonreal eigenvalues have negative real parts. If the eigenvalues are real, they will both be negative if $(\delta_2 + \phi)(\theta_1 + \delta_1 - \psi) - \rho\delta_0 > 0$ which is equivalent to $\delta_1 > \rho\hat{\delta}_2(1 + \hat{\delta}_2)^{-1} + \psi - \theta_1$ where $\hat{\delta}_2 = \delta_2\phi^{-1}$. The assumption $\delta_1 > \psi + \rho\hat{\delta}_2$ shows that both eigenvalues are negative in this case and $(0, 0, 1)$ is stable.

Coexistence equilibria obey $\hat{x}_1 > 0$, $\hat{y}_1 > 0$, $\hat{x}_2 > 0$ and are found by solving (2-28) for $\hat{y}_1$. The solutions of (2-28) are

$$\hat{y}_\pm = \frac{v(1+\hat{\delta}_2)+1-\xi \pm \sqrt{(v(1+\hat{\delta}_2)+1-\xi)^2 - 4v(1+\hat{\delta}_2-\xi\kappa)}}{2v}. \quad (2\text{-}34)$$

The parameters v, ξ, and κ are defined in (2-29). The assumption $\theta_2 < \theta_1^{-1}\psi$ implies that v is positive. The assumption $\delta_1 > \psi + \rho\hat{\delta}_2$ implies that $\xi > 1$ and $\kappa > 1 + \hat{\delta}_2$. Therefore the radicand in (2-34) is positive, $\hat{y}_\pm$ are real, $y_- < 0$, and $y_+ > 0$. Thus y_- is not feasible. Expanding the expression in the radicand of (2-34) yields that

$$(\xi - 1)^2 + v^2(1+\hat{\delta}_2)^2 - 2v(1+\hat{\delta}_2) - 2v\xi(1+\hat{\delta}_2) + 4v\xi\kappa,$$

which is larger than $(\xi - 1 + v(1+\hat{\delta}_2))^2$ using the fact that $4v\xi\kappa > 4v\xi(1+\hat{\delta}_2)$. It follows that $y_+ > 1 + \hat{\delta}_2$ so y_+ is not feasible since $0 < \hat{y}_1 < 1$ for coexistence. $\square$

Figure 5 shows the time courses for simulations of (2-7), (2-8), (2-9) when the hypotheses of Theorem 1 are obeyed. Substituting (2-4), (2-5), and (2-6) into the assumptions in Theorem 1 yields that

$$d_A > r_A + s, \quad r_L > \max\left\{\alpha_2 m_A, \frac{\alpha_1 \alpha_2 m_A m_L}{r_A}\right\}. \quad (2\text{-}35)$$
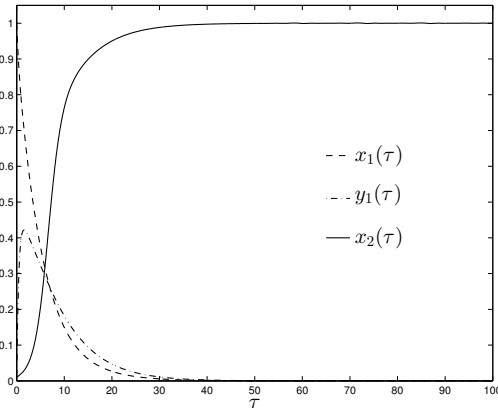
Figure 5. Simulation of the system (2-7), (2-8), (2-9) where the parameters are $\theta_1 = 0.1$, $\psi = 1$, $\rho = 1$, $\delta_2 = 0.1$, $\phi = 0.25$, $\delta_1 = 1.1 \cdot (\psi + \rho\hat{\delta}_2)$, and $\theta_2 = 0.9 \cdot \min\{1, \theta_1^{-1}\psi\}$. The initial conditions are $(x_1(0), y_1(0), x_2(0)) = (1, 0, 0.01)$. For these values of δ_1 and θ_2, the hypotheses of Theorem 1 are obeyed and $(0, 0, 1)$ is the only feasible stable equilibrium of (2-7)–(2-9).

If the growth rate of the free-floating plant r_L may be enhanced by nutrient loading as described in [Scheffer et al. 2003], it may be possible that the second inequality in (2-35) is satisfied.

3. Conclusion

We have presented a modified Lotka–Volterra competition model (2-1)–(2-3) for two competing aquatic plants where one species is a submersed plant while the other is a free-floating plant. We investigated how herbivory by grass carp affects the competitive abilities of the submersed and free-floating plants. In Section 2.1 we analyzed a reduced model (2-10) by phase-plane methods and computed equilibria and stability of these equilibria. We derived conditions in (2-35) on the grass carp stocking rate d_A so that the free-floating plant extinction equilibrium is unstable and free-floating plants may dominate the ecosystem. In addition, we showed that grass carp stocking may exhibit a hysteresis effect whereby grass carp may be decreased below the critical level at which the free-floating plant extinction equilibrium loses stability and suppression of the submersed plant biomass may still be achieved. This is depicted in the bifurcation diagram in Figure 3. In Section 2.2 we included the belowground biomass dynamics of the submersed plant. We proved Theorem 1 which provides sufficient conditions (2-35) on the grass carp stocking rate d_A and free-floating plant growth rate r_L that guarantee the free-floating plant carrying capacity equilibrium is the only feasible equilibrium and is locally stable.

Although the model (2-1)–(2-3) is qualitative and not intended to give a detailed quantitative description of the biology, it may be analyzed without extensive numerical computations and the results are amenable to biological interpretation and experimentation. For example, (2-35) shows that the minimal stocking rate is the sum of the growth rate of the aboveground biomass for the submersed plant (r_A) and the rate at which the aboveground biomass supplies energy for growth of the belowground biomass (s). Both of these quantities depend on the particular species of submersed and floating plant being considered, but they may be measured experimentally and an experimentally determined stocking rate may then be compared with the minimal stocking rate predicted here. Similarly, the predicted hysteresis effect may be experimentally verified just as in [Scheffer et al. 2003].

Finally, we discuss some model weaknesses and future work. Grass carp were assumed to graze on aboveground biomass at a rate proportional to the amount of aboveground biomass, with d_A the proportionality constant, resulting in the term $d_A A$ in (2-2). This is a linear functional response [Turchin 2003] in grass carp herbivory. The hyperbolic or Holling's type II functional response [Turchin 2003] is $k N A (D + A)^{-1}$. Here A is the aboveground biomass of the submersed plant, k is the

maximum killing rate, N is the number of grass carp, and D is the prey (submersed plant) density at which the killing rate is half of the maximum. This functional response models a saturation of the grass carp feeding rate so that grass carp have a maximum rate of consumption (kN) of submersed plant biomass. Future work will include analysis of a model with hyperbolic functional response for the grass carp. We have also assumed spatial heterogeneity in our formulation of the model using ordinary differential equations. Future investigations will be to include modeling spatial heterogeneities in the ecosystem with partial differential equations.

References

[Center et al. 2005] T. D. Center, T. K. Van, F. A. Dray, Jr., S. J. Franks, M. T. Rebelo, P. D. Pratt, and M. B. Rayamajhi, "Herbivory alters competitive interactions between two invasive aquatic plants", *Biol. Control* **33** (2005), 173–185.

[Cuda et al. 2008] J. P. Cuda, R. Charudattan, M. J. Grodowitz, R. M. Newman, J. F. Shearer, M. L. Tamayo, and B. Villegas, "Recent advances in biological control of submersed aquatic weeds", *J. Aquat. Plant Manage.* **46** (2008), 15–32.

[Edelstein-Keshet 2005] L. Edelstein-Keshet, *Mathematical models in biology*, Classics in Applied Mathematics **46**, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005. Reprint of the 1988 original. MR 2131632 Zbl 1100.92001

[Gettys et al. 2009] L. A. Gettys, W. T. Haller, and M. Bellaud (editors), *Biology and control of aquatic plants: A best management practices handbook*, 2nd ed., Aquatic Ecosystem Restoration Foundation, Marietta, GA, 2009.

[Gurney and Nisbet 1998] W. S. C. Gurney and R. M. Nisbet, *Ecological Dynamics*, Oxford University Press, 1998.

[Hanlon et al. 2000] S. G. Hanlon, M. V. Hoyer, C. E. Cichra, and D. E. Canfield, Jr., "Evaluation of macrophyte control in 38 Florida lakes using triploid grass carp", *J. Aquat. Plant Manage* **38** (2000), 48–54.

[Kay and Hoyle 2001] S. H. Kay and S. T. Hoyle, "Mail order, the internet, and invasive aquatic weeds", *J. Aquat. Plant Manage.* **39** (2001), 88–91.

[Pine and Anderson 1991] R. T. Pine and L. W. J. Anderson, "Plant preferences of the triploid grass carp", *J. Aquat. Plant Manage.* **29** (1991), 80–82.

[Scheffer et al. 2001] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, and B. Walker, "Catastrophic shifts in ecosystems", *Nature* **413** (2001), 591–596.

[Scheffer et al. 2003] M. Scheffer, S. Szabó, A. Gragnani, E. H. van Nes, S. Rinaldi, N. Kautsky, J. Norberg, R. M. M. Roijackers, and R. J. M. Franken, "Floating plant dominance as a stable state", *Proc. Natl. Acad. Sci.* **100** (2003), 4040–4045.

[Shukla 1998] V. P. Shukla, "Modelling the dynamics of wetland macrophytes: Keoladeo National Park wetland, India", *Ecol. Model* **109** (1998), 99–114.

[Strogatz 2001] S. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*, Westview Press, Boulder, CO, 2001.

[Sutton et al. 2012] D. L. Sutton, V. V. Vandiver, Jr., and J. E. Hill, "Grass carp: A fish for biological management of hydrilla and other aquatic weeds in Florida", Bulletin 867, Department of Fisheries and Aquacultural Sciences, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida, 2012, http://edis.ifas.ufl.edu/pdffiles/FA/FA04300.pdf.

[Tipping et al. 2009] P. W. Tipping, L. Bauer, M. R. Martin, and T. D. Center, "Competition between *Salvinia minima* and *Spirodela polyrhiza* mediated by nutrient levels and herbivory", *Aquat. Bot.* **90** (2009), 231–234.

[Turchin 2003] P. Turchin, *Complex population dynamics: A theoretical/empirical synthesis*, Monographs in Population Biology **35**, Princeton University Press, 2003. MR 2005f:92024 Zbl 1062.92077

[Turchin and Batzli 2001] P. Turchin and G. O. Batzli, "Availability of food and the population dynamics of arvicoline rodents", *Ecology* **82** (2001), 1521–1534.

[Van et al. 1998] T. K. Van, G. S. Wheeler, and T. D. Center, "Competitive interactions between Hydrilla (*Hydrilla verticillata*) and Vallisneria (*Vallisneria americana*) as influenced by insect herbivory", *Biological Control* **11** (1998), 185–192.

[Wangersky 1978] P. J. Wangersky, "Lotka–Volterra Population Models", *Annu. Rev. Ecol. Syst.* **9** (1978), 189–218.

[Wilson et al. 2001] J. R. Wilson, M. Rees, N. Holst, M. B. Thomas, and G. Hill, "Water hyacinth population dynamics", pp. 96–104 in *Biological and integrated control of water hyacinth,* Eichhornia crassipes*: Proceedings of the Second Meeting of the Global Working Group for the Biological and Integrated Control of Water Hyacinth* (Beijing, China, 9–12 October 2000), edited by M. H. Julien et al., 2001.

[Wilson et al. 2005] J. R. Wilson, N. Holst, and M. Rees, "Determinants and patterns of population growth in water hyacinth", *Aquat. Bot.* **81** (2005), 51–67.

[Wolverton and McDonald 1979] B. C. Wolverton and R. C. McDonald, "The water hyacinth: From prolific pest to potential provider", *Ambio* **8** (1979), 2–9.

[Zeeman 1995] M. L. Zeeman, "Extinction in competitive Lotka–Volterra systems", *Proc. Amer. Math. Soc.* **123**:1 (1995), 87–96. MR 95c:92019 Zbl 0815.34039

jalford@shsu.edu   *Mathematics and Statistics, Sam Houston State University, P.O. Box 2206, Hunstville, TX 77341-2206, United States*

cab035@shsu.edu   *Mathematics and Statistics, Sam Houston State University, P. O. Box 2206, Huntsville, TX 77341-2206, United States*

kxp004@shsu.edu   *Department of Chemistry, Sam Houston State University, P. O. Box 2117, Huntsville, TX 77341-2117, United States*

cxh016@shsu.edu   *Mathematics and Statistics, Sam Houston State University, P.O. Box 2206, Huntsville, TX 77341-2206, United States*

Irreducible divisor graphs for numerical monoids

Dale Bachman, Nicholas Baeth and Craig Edwards

(Communicated by Scott Chapman)

The factorization of an element x from a numerical monoid can be represented visually as an irreducible divisor graph $G(x)$. The vertices of $G(x)$ are the monoid generators that appear in some representation of x, with two vertices adjacent if they both appear in the same representation. In this paper, we determine precisely when irreducible divisor graphs of elements in monoids of the form $N = \langle n, n + 1, \ldots, n + t \rangle$ where $0 \le t < n$ are complete, connected, or have a maximum number of vertices. Finally, we give examples of irreducible divisor graphs that are isomorphic to each of the 31 mutually nonisomorphic connected graphs on at most five vertices.

1. Introduction and preliminaries

Irreducible divisor graphs related to commutative rings were introduced and studied in [Coykendall and Maney 2007] and later studied in [Maney 2008; Axtell and Stickles 2008; Axtell et al. 2011]. In these papers, the authors represent elements of commutative rings using graphs which provide information about factorization properties of these elements. The general goal is to use graph-theoretic information to study factorization properties in the ring. As a notable example, it was shown in [Coykendall and Maney 2007; Axtell et al. 2011] that an atomic domain is a unique factorization domain precisely when every irreducible divisor graph over that ring is complete (equivalently, connected). We note that graphical representations of numerical semigroups have also been useful in computing a minimal set of relations, as in [Rosales 1996].

In this paper, we study irreducible divisor graphs of elements in numerical monoids — additive submonoids of the nonnegative integers. Our results indirectly apply to irreducible divisor graphs of elements of the form x^n in a polynomial ring of the form $\mathbb{F}[x^{n_1}, x^{n_2}, \ldots, x^{n_t}]$ where $\mathbb{F}$ is a field, x is an indeterminate and $n_1 < n_2 < \cdots < n_t$ are positive integers. By considering a specific family of monoids (and hence commutative rings) we are able to provide more precise information

about which graphs can be realized as irreducible divisor graphs of elements in various monoids and hence rings.

In this section we formally introduce irreducible divisor graphs of elements in numerical monoids and give some preliminary results that both motivate and provide useful tools for later sections. In Section 2 we consider numerical monoids generated by intervals of positive integers. Using the results of [García-Sánchez and Rosales 1999], where numerical monoids generated by intervals were thoroughly studied, we are able to classify exactly when the irreducible divisor graph of an element is complete and/or connected. We conclude Section 2 by presenting a method that can be used to determine whether or not a connected graph can be realized as the irreducible divisor graph of an element in a numerical monoid generated by a given interval. In Section 3 we show, by way of examples, that every connected graph with between one and five vertices can be realized as the irreducible divisor graph of an element in some numerical monoid. This leads us to ask the following question:

Question 1.1. Can every connected graph be realized as the irreducible divisor graph of an element in some numerical monoid?

Throughout, $\mathbb{N}$ will denote the set of all positive integers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Recall that a numerical monoid is an additive submonoid of $\mathbb{N}_0$. More precisely, if $0 < n_1 < n_2 < \cdots < n_t$ are t positive integers such that for all $i \in \{2, \ldots, t\}$, $n_i = a_1 n_1 + \cdots + a_{i-1} n_{i-1}$ has no nonnegative integer solutions $\{a_1, a_2, \ldots, a_{i-1}\}$, then

$$N = \langle n_1, n_2, \ldots, n_t \rangle = \{a_1 n_1 + \cdots + a_t n_t : a_i \in \mathbb{N}_0\} \subseteq \mathbb{N}_0$$

is the *numerical monoid* minimally generated by the set $\{n_1, n_2, \ldots, n_t\}$. We now give a formal definition of the irreducible divisor graph of an element in a numerical monoid, mimicking the definition of the irreducible divisor graph of a nonzero nonunit of an atomic domain.

Definition 1.2. Let $N = \langle n_1, n_2, \ldots, n_t \rangle$ be a minimally generated numerical monoid. If $x \in N$, the *irreducible divisor graph of x*, denoted by $G_N(x)$, is defined as follows:

(1) The vertex set $V[G_N(x)]$ of $G_N(x)$ consists of the n_i for all i such that there exist $a_1, a_2, \ldots, a_t \in \mathbb{N}_0$ with $x = \sum_{j=1}^{t} a_j n_j$ and $a_i \neq 0$.

(2) The *edge set* $E[G_N(x)]$ of $G_N(x)$ has an edge from n_i to n_j for all pairs (i, j) for which there exist $a_1, a_2, \ldots, a_t \in \mathbb{N}_0$ with $x = \sum_{k=1}^{t} a_k n_k$, and $a_i, a_j \neq 0$.

(3) We put $A_i - 1 \geq 0$ loops on vertex n_i, where $A_i = \max\{a_i : x = \sum_{k=1}^{t} a_k n_k$ for some $a_1, \ldots, a_t \in \mathbb{N}_0\}$.

Thus, if $x \notin N$, the graph $G_N(x)$ is empty (has no vertices or edges). We write $G(x)$ in place of $G_N(x)$ if N is clear from context. Although we represent an edge as (n_i, n_j), this is not to be considered as an ordered pair and (n_j, n_i) represents the same edge.

This definition is consistent with the definition from [Coykendall and Maney 2007], in that if R is the semigroup ring $R = \mathbb{F}[y^{n_1}, y^{n_2}, \ldots, y^{n_t}]$ for some field $\mathbb{F}$ and some variable y, the graphs $G_N(x)$ and $G_R(y^x)$ are isomorphic.

Example 1.3. Let N have minimal generating set $\{5, 11, 12, 13, 14\}$ and let $x = 30$. In N we can express x only as $x = 5 + 11 + 14$, $x = 5 + 12 + 13$ and $x = 6 \cdot 5$. Thus $G(30)$ contains 5, 11, 12, 13 and 14 as vertices, with edges connecting vertices 5 and 11, 5 and 12, 5 and 13, 5 and 14, 11 and 14, and 12 and 13. Moreover, there are 5 loops on vertex 5, since $30 = 6 \cdot 5$. Thus, the irreducible divisor graph of $x = 30$ in $N = \langle 5, 11, 12, 13, 14 \rangle$ is as follows:

$$
\begin{array}{c}
11 \\
\diagup \;\; | \;\; \overset{\frown}{\underset{\smile}{5}} \\
14 - 5 \; 12 \\
| \diagup \\
13
\end{array}
$$

The following equivalent definition of an irreducible divisor graph will be useful when determining which vertices and edges occur in an irreducible divisor graph $G(x)$ and will be used extensively in the following sections.

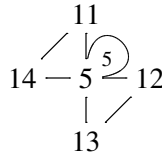**Definition 1.4.** Let $N = \langle n_1, n_2, \ldots, n_t \rangle$ be a numerical monoid. If $x \in N$, the *irreducible divisor graph of x*, denoted by $G_N(x)$, is defined as follows:

(1) $n_i \in V[G(x)]$ if and only if $x - n_i \in N$.

(2) $(n_i, n_j) \in E[G(x)]$ if and only if $x - (n_i + n_j) \in N$.

Remark 1.5. Let $x \in N$, where $N = \langle n_1, n_2, \ldots, n_t \rangle$ and $\{n_1, n_2, \ldots, n_t\}$ is a minimal generating set for N, and let $M = \langle rn_1, rn_2, \ldots, rn_t \rangle$.

(1) Clearly $rx \in M$, and $\{rn_1, rn_2, \ldots, rn_t\}$ is a minimal generating set for M.

(2) For any i,
$$n_i \in V[G_N(x)] \iff rn_i \in V[G_M(rx)].$$

(3) For any distinct i and j,
$$(n_i, n_j) \in E[G_N(x)] \iff (rn_i, rn_j) \in E[G_M(rx)].$$

Thus it is sensible, when studying irreducible divisor graphs of numerical monoids, to study only *primitive* numerical monoids — those for which the generating set is relatively prime. For the balance of this article (except for some examples in Section 3) we consider numerical monoids of the form $\langle n, n+1, \ldots, n+t \rangle$. These

are primitive, and the relationship described above allows results to be applied to associated nonprimitive numerical monoids as well.

For a primitive numerical monoid N, the *Frobenius number*, $\mathcal{F}(N)$, of N is the largest natural number not in N. The following easy proposition, whose proof we leave to the reader, gives extreme conditions for when an irreducible divisor graph is either complete (all pairs of vertices are adjacent) or is completely devoid of edges. This result tells us is that the problem of describing $G_N(x)$ for a given numerical monoid N is finite — once x is large enough, it is obvious that $G_N(x)$ contains all possible vertices and edges. We will improve this result for certain classes of numerical monoids in Section 2.

Proposition 1.6. *Let $N = \langle n_1, n_2, \ldots, n_t \rangle$ be a primitive minimally generated numerical monoid with $n_1 < n_2 < \cdots < n_t$.*

(1) *If $x > \mathcal{F}(N) + n_{t-1} + n_t$, then $G(x)$ is complete.*

(2) *If $x < 2n_1$, then $G(x)$ has no edges.*

An example shows that the converses of (1) and (2) in Proposition 1.6 are false. Let $N = \langle 12, 13, 14 \rangle$. Then $G(65)$ is complete because $65 = (12) + 3(13) + (14)$. However, $\mathcal{F}(N) = 71$ and $65 < \mathcal{F}(N) + 13 + 14$. Moreover, $G(29)$ is an empty graph since $29 = 12a + 13b + 14c$ has no nonnegative integer solutions (a, b, c). However, $29 \geq 2 \cdot 12$.

2. Numerical monoids generated by intervals

In this section we study numerical monoids generated by intervals; that is, minimally generated by the set $\{n, n+1, \ldots, n+t\}$, where $n \geq 1$ and $0 \leq t \leq n-1$. For the balance of this paper we will use the notation $[a, b]$ (where $a \leq b$) to represent the interval of natural numbers $\{a, a+1, \ldots, b\}$. We start with two results that we will apply often.

Proposition 2.1 [García-Sánchez and Rosales 1999, Lemma 1 and Corollary 5]. *Let $n, t \in \mathbb{N}$ and let $N = \langle n, \ldots, n+t \rangle$.*

(1) *$x \in N$ if and only if $x \in [pn, p(n+t)]$ for some $p \in \mathbb{N}$.*

(2) *$\mathcal{F}(N) = \left\lceil \frac{n-1}{t} \right\rceil n - 1$.*

For ease of discussion, we name the intervals (as subsets of the natural numbers) contained in the monoid. We let $N_p = [pn, p(n+t)]$, where $p \in \mathbb{N}_0$, and note that $|N_p| = pt + 1$. We define a *gap* in N to be a maximal (with respect to set containment) nonempty interval of natural numbers that is not contained in N. In order to help with visualization, Figure 1 shows the intervals and gaps for the monoid $N = \langle n, \ldots, n+t \rangle$. Notice in particular that the first gap (between N_0 and N_1) has size $n-1$, and that subsequent gaps decrease in size by t.
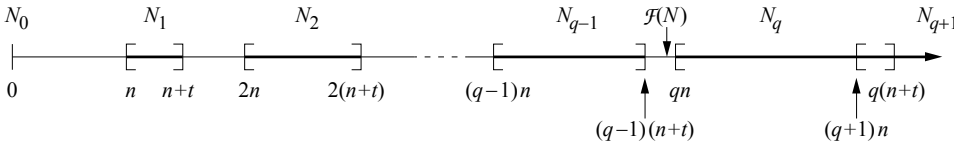
Figure 1. Intervals contained in the monoid $N = \langle n, \ldots, n+t \rangle$.

Graph-theoretic properties of $G(x)$. The next result shows that the only irreducible divisor graphs of elements in a numerical monoid generated by an interval containing no loops are disjoint unions of components each isomorphic to K_1 or K_2, the complete graphs on one and two vertices. Since loops almost always occur, we omit consideration of loops in the sequel.

Proposition 2.2. *Let $n \in \mathbb{N}$ and $N = \langle n, n+1, n+2, \ldots, n+t \rangle$ where $0 \leq t \leq n-1$. If $x \in N$, then $G(x)$ has no loops if and only if $G(x)$ is isomorphic to a disjoint union of components each isomorphic to K_1 or K_2.*

Proof. If $x \in N$, then by Proposition 2.1 $x \in [pn, p(n+t)]$ for some positive integer p. Thus $x = pn+k$ where $0 \leq k \leq pt$. First assume $p \geq 3$ and write $x = pn+ps+r$ where either $0 \leq s < t$ and $0 \leq r < p$ or else $s = t$ and $r = 0$. If $0 \leq s < t$ and $0 \leq r < p$, then $x = r(n+s+1)+(p-r)(n+s)$. Since $p \geq 3$, either $r \geq 2$ or $p-r \geq 2$. Thus there is at least one loop on either the vertex $n+s$ or the vertex $n+s+1$. If $x = p(n+t)$ then there are $p-1 \geq 2$ loops on vertex $n+t$. Therefore, if $p \geq 3$, $G(x)$ contains at least one loop.

If $p = 1$, then $x = n+i$ where $i \in [0, t]$ and $G(x)$ is isomorphic to K_1. If $p = 2$, then $x = 2n+j$ where $1 \leq j \leq 2t-1$. If j is even, then $x = 2n+j = 2(n+j/2)$, resulting in a loop on the vertex $n+j/2$. If j is odd, then note that, for any $n+i \in V[G(x)]$, $x-(n+i) = 2n+j-(n+i) = n+j-i$ and hence $0 \leq j-i \leq t$. Thus $x-[(n+i)+(n+j-i)] = 0$ and so $n+i$ is adjacent only to $n+j-i$. As this holds for all i with $n+i \in V[G(x)]$, $G(x)$ consists of multiple components isomorphic to K_2, which by definition has no loops. $\square$

The next set of theorems — our main results — give complete classifications of when $G(x)$ has $t+1$ vertices, is connected with $t+1$ vertices, or is complete with $t+1$ vertices whenever $x \in \langle n, n+1, \ldots, n+t \rangle$.

Proposition 2.3. *Let $N = \langle n, \ldots, n+t \rangle$, where $n > 1$ and $0 < t < n$. Then $G(x)$ has $t+1$ vertices if and only if $x \in [(p+1)n+t, (p+1)n+pt]$ with $p > 0$. Moreover, if $x > \mathcal{F}(N)+n+t$ then $G(x)$ has $t+1$ vertices.*

Proof. By Definition 1.4, vertex $n+i$ is in the graph if and only if $x-n-i \in N$. Thus the $t+1$ vertices $\{n, \ldots, n+t\}$ are in the graph if and only if

$$S := [x-n-t, x-n] \subset N.$$

Since $N = \bigcup_{p \geq 0} N_p$ (by Proposition 2.1) and since $|N_p| \geq t+1$ for $p > 0$, we have $S \subset N_p$ for some $p > 0$ when $pn \leq x - n - t$ and $x - n \leq p(n+t)$, i.e., $x \in [(p+1)n + t, (p+1)n + pt]$.

The last condition expresses the case when x is sufficiently large that the integers in S are all larger than $\mathcal{F}(N)$; since there are no gaps above this point, $S \subseteq N$. This is also the point above which

$$[(p+1)n + t, (p+1)n + pt] \cap [(p+2)n + t, (p+2)n + (p+1)t] \neq \varnothing. \qquad \square$$

Proposition 2.4. *Let $N = \langle n, \ldots, n+t \rangle$, where $n > 1$ and $0 < t < n$. Then $G(x)$ is complete on $t+1$ vertices if and only if $x \in [(p+2)n + 2t - 1, (p+2)n + pt + 1]$ for $p \geq 0$ (if $t = 1$), $p > 0$ (if $t = 2$) and $p > 1$ otherwise. Moreover, if $x > \mathcal{F}(N) + 2n + 2t + 1$ then $G(x)$ is complete on $t+1$ vertices.*

Proof. By Definition 1.4 the graph is complete if and only if $x - (n+i) - (n+j) \in N$ for each pair of distinct i and j in $[0, t]$, that is, when $S = [x - (n+t) - (n+t-1), x - n - (n+1)] \subset N$. Note that $|S| = 2t - 1$ and $|N_p| = pt + 1 \geq 2t - 1$ when $p \geq (2t-2)/t$, which produces the bounds on p. When N_p is large enough to contain S, it is also required that $pn \leq x - (n+t) - (n+t-1)$ and $x - n - (n+1) \leq p(n+t)$ which implies $x \in [(p+2)n + 2t - 1, (p+2)n + pt + 1]$.

As in Proposition 2.3, the second condition occurs when all elements of S are larger than $\mathcal{F}(N)$, that is, $S \subset N$ whenever $x - 2n - 2t + 1 > \mathcal{F}(N)$. $\qquad \square$

The goal now is to give a result analogous to Propositions 2.3 and 2.4 for connected graphs with $t+1$ vertices. First we require two technical lemmas which relate the vertex degrees of $G(x)$ to the set $S = [x - 2n - 2t + 1, x - 2n - 1]$. We then use this set to characterize when $G(x)$ is connected on $t+1$ vertices.

Lemma 2.5. *Let $N = \langle n, \ldots, n+t \rangle$, where $n > 1$ and $0 < t < n$, and let $S = [x - 2n - 2t + 1, x - 2n - 1]$. Then*

(1) *If S contains an interval of length $t+1$ that is contained in N then $G(x)$ has a vertex of degree t.*

(2) *If S contains an interval of length $t+1$ that is disjoint from N then $G(x)$ has a vertex of degree 0.*

Proof. Let $S_k = [x - 2n - k - t, x - 2n - k]$ be an interval of length $t+1$ in S.

For the first statement, we can find k so that $S_k \subset N$. The edge $(n+k, n+j)$ is in $E[G(x)]$ if and only if $x - 2n - k - j \in N$. Since $S_k \subset N$, $x - 2n - k - j \in N$ for $0 \leq j \leq t$. Thus (ignoring loops on $n+k$) the vertex $n+k$ has degree t.

For the second statement, we can find k so that S_k is disjoint from N. As above, we see that vertex $n+k$ is not adjacent to any other vertex. $\qquad \square$

If not for the vertices n and $n+t$, the preceding statements could each be made into equivalences. In fact, these vertices will require examination during the course

of the next proof; we did not complicate the statement of Lemma 2.5 because these special cases each occur only once.

Lemma 2.6. *Let $N = \langle n, \ldots, n+t \rangle$, where $n > 1$ and $0 < t < n$, and let $S = [x - 2n - 2t + 1, x - 2n - 1]$. Then $G(x)$ is connected on $t + 1$ vertices if and only if $|S \cap N| \geq t$.*

Proof. We note that an edge $(n+i, n+j)$ is in $E(G(x))$ when $x - (n+i) - (n+j) \in N$. Since $x - 2n - 2t + 1 \leq x - 2n - i - j \leq x - 2n - 1$, $E(G(x))$ is characterized by the intersection of S and N.

Furthermore, either $S \cap N \subset N_p$ or $S \cap N \subset N_p \cup N_{p+1}$ for some p. To see this, we assume that $S \cap N_p \neq \varnothing$. Then $x - 2n - 2t + 1 \leq p(n+t)$, and hence $x - 2n - 1 \leq (p+1)(n+t) - n + t - 2 < (p+1)(n+t)$. Thus the largest element of S is smaller than the largest element of N_{p+1}. We may therefore consider two cases:

Case 1: $S \cap N = S \cap N_p$ for some p. We divide this case into three subcases: $|S \cap N| > t$, $|S \cap N| = t$ or $|S \cap N| < t$.

In the first subcase, we notice that there is an interval of length at least $t + 1$ in $S \cap N$ (in fact, $S \cap N$ is a single interval), so by Lemma 2.5 there is a vertex of degree t and hence $G(x)$ is connected on $t + 1$ vertices.

For the second subcase we assume $|S \cap N| = t$. Since $|N_p| = pt + 1$, we certainly have $|N_p| \neq t$ unless $p = 0$ and $t = 1$, in which case $G(x) = K_2$, which is connected on two vertices. Otherwise, $S \cap N \subset N_p$, so $|N_p| > t$. Since both N_p and S are intervals, $S \cap N_p$ comprises precisely either the first t elements of N_p or the last. If $S \cap N_p = [x - 2n - 2t + 1, x - 2n - t]$ then $\deg(n+t) = t$, while if $S \cap N_p = [x - 2n - t, x - 2n - 1]$ then $\deg(n) = t$.

In the last subcase, we note that if $S \cap N = \varnothing$ then there is an interval of length at least $t + 1$ (namely, all of S) that is not contained in N, so by Lemma 2.5 $G(x)$ is not connected. We assume for the balance of this case that $S \cap N$ is nonempty.

If $|N_p| > t$, that is, $p > 0$, then since $|S \cap N_p| < t$, S cannot extend the interval N_p in two directions, hence the intersection of S with the complement of N is a single interval. Thus there is an interval of length at least $t + 1$ that is not in N, so by Lemma 2.5 there is a vertex of degree 0 and $G(x)$ is not connected.

If $p = 0$, then $S \cap N = \{0\}$, and the degree of each vertex is at most 1. If $t > 1$, this shows that $G(x)$ is not connected. If $t = 1$, then the hypothesis of the subcase $|S \cap N| < t$ is not satisfied.

Case 2: S intersects the two intervals N_{p-1} and N_p, as shown in Figure 2. We choose k so that $x - 2n - k = pn$, the smallest element of N_p, and we let $S \cap N = [x - 2n - 2t + 1, x - 2n - k - j] \cup [x - 2n - k, x - 2n - 1]$.

We divide this case into the three subcases $|S \cap N| > t - 1$ (i.e., $|S \cap N| \geq t$), $|S \cap N| = t - 1$ and $|S \cap N| < t - 1$.
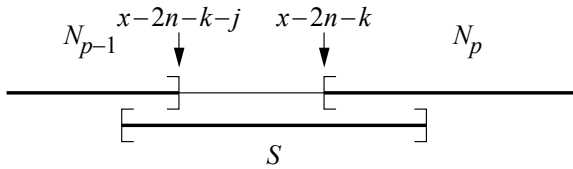
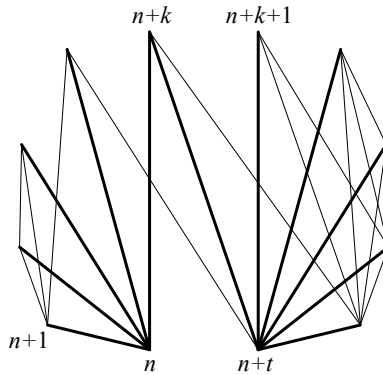**Figure 2.** Case 2: S overlaps two intervals.

Figure 3. $G_N(x)$, in Case 2 when $|S \cap N| \geq t$, with a connected subgraph highlighted.

The graph for the first subcase is shown in Figure 3. In this case $j \leq t$, and the verification that the darkened subgraph exists is straightforward. In particular, the element of S associated with the edge $(n+k, n+t)$ is $x-2n-k-t$, so this element and the ones associated with the other darkened edges involving $n+t$ are contained in the lower portion of $S \cap N$, while the ones associated with the edges involving n are contained in the upper portion.

If $|S \cap N| < t-1$, the gap between N_{p-1} and N_p contains at least $t+1$ consecutive integers, so by Lemma 2.5 there is a vertex of degree 0, and $G(x)$ is not connected.

We are left with the subcase $|S \cap N| = t-1$. The graph for this case is shown in Figure 4, and we verify that the subgraph on vertices $\{n, \ldots, n+k\}$ and that on $\{n+k+1, \ldots, n+t\}$ have no edges between them. Indeed, the missing edges between the subgraphs are associated with the elements $x-(n+k)-(n+t) = x-2n-k-t$ through $x-n-(n+k+1) = x-2n-k-1$, none of which is in N. $\qquad \square$

Proposition 2.7. *Let* $N = \langle n, \ldots, n+t \rangle$, *where* $n > 1$ *and* $0 < t < n$. *Then* $G(x)$ *is connected on* $t+1$ *vertices if and only if at least one of the following conditions holds*:

(1) $x \in [(p+2)n+t, (p+2)n+(p+1)t]$ *for* $p \geq 0$ *(if* $t = 1$*) and* $p > 0$ *otherwise.*

(2) $x > C(N)$, *where* $C(N) = \mathcal{F}(N)+2n+t+1$ *if* t *divides* $n-1$, *and* $C(N) = \mathcal{F}(N)+n+t+1$ *otherwise.*
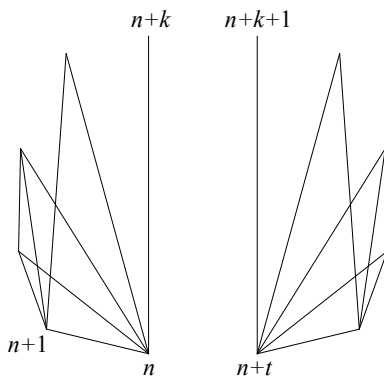
Figure 4. $G_N(x)$, in Case 2 when $|S \cap N| = t - 1$.

Proof. We define $S = [x - 2n - 2t + 1, x - 2n - 1]$ as before and recall that by Lemma 2.6, $G(x)$ is connected on $t + 1$ vertices if and only if $|S \cap N| \geq t$.

If S intersects exactly one interval N_p, then $|S \cap N| \geq t$ when the smallest element of S is close enough to the left end of the interval, that is, $x - 2n - 2t + 1 \geq pn - (t - 1)$, or is not too close to the right end, that is, $x - 2n - 2t + 1 \leq p(n + t) - (t - 1)$. These inequalities give the first condition, and the conditions on p follow from the requirement that $|N_p| \geq t$.

If S spans a gap of size larger than $t - 1$, then $G(x)$ is not connected, while if S spans a gap of size at most $t - 1$ then $G(x)$ is connected. Since consecutive gaps decrease in size by t (refer to Figure 1), the last gap, $\mathcal{G}$, has size at most t. Assume that $S \cap \mathcal{G} \neq \emptyset$. If $|S \cap \mathcal{G}| < t$, then $G(x)$ is connected. If $|S \cap \mathcal{G}| = t$, that is, $\mathcal{F}(N) \in S$, then $G(x)$ is not connected. Moreover, the last gap has size less than t if and only if t does not divide the size of the first gap, namely that between N_0 and N_1, which has size $n - 1$. In this case, $G(x)$ is connected on $t + 1$ vertices for all $x > y$ satisfying $y - 2n - 2t + 1 = np - (t - 1)$ where N_p is the last interval *before* $\mathcal{F}(N)$, that is, if $\mathcal{F}(N) = qn - 1$, then $p = q - 1$. If the last gap is of size t, the relevant p belongs to the interval *after* $\mathcal{F}(N)$, that is, $p = q$. □

Note that Proposition 2.7 is worded differently from Propositions 2.3 and 2.4. In Proposition 2.7, when t does not divide $n - 1$, there are values of x that do not satisfy the first condition, but do produce connected graphs.

The following corollary is a concise restatement of the previous results in the case $t = n - 1$. Though it follows from these results, the direct proof is more straightforward, so it is sketched.

Corollary 2.8. *Let* $n > 2$, $N = \langle n, \ldots, 2n - 1 \rangle$ *and* $x \in N$.

(1) $G(x)$ *has n vertices if and only if* $x \geq 3n - 1$.

(2) *The following statements are equivalent.*

  (a) $G(x)$ *is connected with n vertices.*
  (b) $\deg(n) = n - 1$.
  (c) $x \geq 4n - 1$.

 (3) $G(x)$ *is complete on n vertices if and only if $x \geq 5n - 3$.*

Proof. Notice that $N = \{0\} \cup [n, \infty)$.

  For (1) we require that $[x - (2n - 1), x - n] \subset N$, which is true precisely when $x - (2n - 1) \geq n$.

  For (2) we note that $\deg(n) = n - 1$ (omitting loops, as usual), when $[x - n - (2n - 1), x - n - (n - 1)] \subset N$, which is true precisely when $x - 3n + 1 \geq n$, so conditions (b) and (c) are equivalent. It is clear that in this case $G(x)$ is connected. Conversely, if $G(x)$ is connected then vertex $2n - 1$ is adjacent to at least one other vertex, that is, $x - (2n - 1) - (n + j) \in N$ for some $j \in [0, n]$, so $x - (3n - 1) \geq x - (3n - 1) - j \geq n$, and the inequality (c) is established.

  For (3) we note that vertices $2n - 1$ and $2n - 2$ must be adjacent, so $x - 4n + 3 = x - (2n - 1) - (2n - 2) \geq n$, which produces the inequality. Moreover, if the inequality is satisfied all pairs of vertices are adjacent since $x - (n + i) - (n + j) \geq x - 4n + 3$ if i and j are distinct integers in $[0, n - 1]$. □

Remark 2.9. For $n = 2$, statements (2) and (3) in Corollary 2.8 would not quite be correct, because the set S comprises the single element $x - 5$, and can thus coincide with $N_0 = \{0\}$. Thus, in addition to the ranges listed, $G(5)$ is complete (and therefore connected).

Constructions. The goal of this section is to address the following question: "When N is a numerical monoid generated by an interval, which connected graphs occur as $G(x)$ for some $x \in N$?" Throughout, we assume $N = \langle n, n+1, \ldots, n+t \rangle$ with $0 \leq t \leq n - 1$ and require $G(x)$ to have $t + 1$ vertices. It remains an open question as to what graphs can be realized when not all generators are required to occur as a vertex.

  There are $\binom{t+1}{2}$ ways to choose two distinct values $n + i, n + j \in [n, n+t]$ and yet only $2t - 1$ distinct sums $(n + i) + (n + j)$. By Definition 1.4, vertices $n + i$ and $n + j$ are adjacent in $G(x)$ if $x - [(n + i) + (n + j)] \in N$. Thus, to determine the number of edges that can occur in the irreducible divisor graph $G(x)$ for some $x \in \langle n, n+1, \ldots, n+t \rangle$ we consider the $2t - 1$ possible sums in $[2n + 1, 2n + 2t - 1]$ along with Proposition 2.1.

  We have no general result for what graphs occur when $t > 4$, but the methods of this section may be extended for larger values of t. We now show how to determine which connected 5-vertex graphs with exactly four edges can be realized as $G(x)$ for $x \in N = \langle n, n+1, \ldots, n+4 \rangle$. The results of the remaining cases are outlined in Section 3.

a	Number of edges	Edges
$2n+1$	1	$(n, n+1)$
$2n+2$	1	$(n, n+2)$
$2n+3$	2	$(n, n+3)$, $(n+1, n+2)$
$2n+4$	2	$(n, n+4)$, $(n+1, n+3)$
$2n+5$	2	$(n+1, n+4)$, $(n+2, n+3)$
$2n+6$	1	$(n+2, n+4)$
$2n+7$	1	$(n+3, n+4)$

Table 1. Edges associated with values of $x - a$.

Using Definition 1.4 we can determine which of the $\binom{4+1}{2} = 10$ possible edges occur in $G(x)$ by considering which values $x - \big((n+i)+(n+j)\big)$ are in N as distinct i and j range over the set $\{0, 1, 2, 3, 4\}$. Since $(n+i)+(n+j) \in [2n+1, 2n+7]$, we may summarize the relationships among values $x - a$ and edges in $G(x)$ as in Table 1.

We will use this table as a guide for constructing irreducible divisor graphs $G(x)$ with $x \in \langle n, n+1, n+2, n+3, n+4 \rangle$ such that $G(x)$ has exactly 5 vertices and exactly four edges. By Proposition 2.1, the smallest number of consecutive positive integers in N is 5. Moreover, the number of consecutive integers in N must be $p(n+4) - pn + 1 = 4p + 1$ for some $p \in \mathbb{N}$ and the length of a sequence of consecutive integers not in N must be $(p+1)n - p(n+4) - 1 = n - 4p - 1$ for some integer p with $1 \leq p \leq (n-1)/4$; that is, the gap sizes are congruent to $n-1$ modulo 4.

Referring to Table 1, we see that in order to guarantee exactly 4 edges in $G(x)$, we need to have either 3 or 4 consecutive integers not in N. Indeed, the set $[x - (2n+7), x - (2n+1)]$, which we called S in Lemmas 2.5 and 2.6, must intersect N in at most two intervals; see Figure 2. In the former case we are left with 4 edges exactly when $x - (2n+5)$, $x - (2n+4)$, and $x - (2n+3)$ are not in N. In the latter case we have 4 edges exactly when either $x - (2n+7)$, $x - (2n+6)$, $x - (2n+5)$ and $x - (2n+4)$ are not in N or $x - (2n+4)$, $x - (2n+3)$, $x - (2n+2)$ and $x - (2n+1)$ are not in N.

Suppose first that $x - (2n+5)$, $x - (2n+4)$, and $x - (2n+3)$ are not in N and hence $x - (2n+1)$, $x - (2n+2)$, $x - (2n+6)$, and $x - (2n+7)$ are in N. That is

$$E[G(x)] = \big\{(n, n+1), (n, n+2), (n+2, n+4), (n+3, n+4)\big\}.$$

To guarantee exactly 3 consecutive integers not in N we need, from Proposition 2.1, $n - 4p - 1 = 3$ where $p \geq 1$. In order for the correct three consecutive values to be outside of N, we require $x - (2n+6) = p(n+4)$ since $x - (2n+6)$ is the largest value in N preceding this sequence. Since $n = 4p+4$, $x = \frac{1}{4}n^2 + 2n + 2$ and we

obtain the graph $G\left(\frac{1}{4}n^2+2n+2\right)$ in $N=\langle n,n+1,n+2,n+3,n+4\rangle$ whenever $n=4k$ with $k>1$.

$$n+1 \text{———} n \text{———} n+2 \text{———} n+4 \text{———} n+3$$

Now suppose that either $x-(2n+3)$, $x-(2n+2)$, $x-(2n+1)\in N$ or $x-(2n+7)$, $x-(2n+6)$, $x-(2n+5)\in N$. In the first case,

$$E[G(x)]=\{(n,n+1),(n,n+2),(n,n+3),(n+1,n+2)\}$$

in which case $G(x)$ has only 4 vertices. In the second case,

$$E[G(x)]=\{(n+3,n+4),(n+2,n+4),(n+1,n+4),(n+2,n+3)\}$$

	N	x
1	$\langle n\rangle$, $n>0$	pn, $p>0$
2	$\langle n,n+1\rangle$, $n>1$	$2n+1$
3	$\langle n,n+1,n+2\rangle$, $n=2k$, $k>1$	$\frac{1}{2}n^2+2n$
4	$\langle n,n+1,n+2\rangle$, $n>3$	$x\in[pn+3,p(n+2)-3]$, $p>3$
4	$\langle 3,4,5\rangle$	$x>11$
6	$\langle n,..,n+3\rangle$, $n=3k$, $k>1$	$\frac{1}{3}n^2+2n+2$
7	$\langle n,..,n+3\rangle$, $n=3k$, $k>1$	$\frac{1}{3}n^2+2n+3$
8	$\langle n,..,n+3\rangle$, $n=3k+2$, $k>0$	$\frac{1}{3}n^2+\frac{7}{3}n+2$
9	$\langle n,..,n+3\rangle$, $n=3k+2$, $k>0$	$\frac{1}{3}n^2+\frac{7}{3}n$
10	$\langle n,..,n+3\rangle$, $n>4$	$x\in[pn+5,p(n+3)-5]$, $p>3$
10	$\langle 4,5,6,7\rangle$	$x>16$
13	$\langle n,..,n+4\rangle$, $n=4k$, $k>1$	$\frac{1}{4}n^2+2n+2$
16	$\langle n,..,n+4\rangle$, $n=4k$, $k>1$	$\frac{1}{4}n^2+2n+3$
19	$\langle n,..,n+4\rangle$, $n=4k$, $k>1$	$\frac{1}{4}n^2+2n$
22	$\langle n,..,n+4\rangle$, $n=4k+3$, $k>0$	$\frac{1}{4}n^2+\frac{9}{4}n+2$
26	$\langle n,..,n+4\rangle$, $n=4k+3$, $k>0$	$\frac{1}{4}n^2+\frac{9}{4}n+4$
28	$\langle n,..,n+4\rangle$, $n=4k+3$, $k>0$	$\frac{1}{4}n^2+\frac{9}{4}n+5$
29	$\langle n,..,n+4\rangle$, $n=4k+2$, $k>0$	$\frac{1}{4}n^2+\frac{5}{2}n+4$
30	$\langle n,..,n+4\rangle$, $n=4k+2$, $k>0$	$\frac{1}{4}n^2+\frac{5}{2}n+6$
31	$\langle n,..,n+4\rangle$, $n>5$	$x\in[pn+7,p(n+4)-7]$, $p>3$
31	$\langle 5,6,7,8,9\rangle$	$x>21$

Table 2. Construction families. The first column refers to the numbering in Figure 5. We use the abbreviation $\langle n,..,n+3\rangle$ for $\langle n,n+1,n+2,n+3\rangle$, and likewise for $\langle n,..,n+4\rangle$.

and again we have a graph with only 4 vertices. Therefore, the graph shown above is the only graph with 5 vertices and 4 edges that can be realized as $G(x)$ for some $x \in \langle n, n+1, n+2, n+3, n+4 \rangle$.

Similar arguments can be made to determine which connected graphs on $t+1$ vertices can be realized as $G(x)$ with $x \in \langle n, n+1, \ldots, n+t \rangle$, and these conditions are listed in Table 2 on the previous page.

3. Connected graphs with at most five vertices

In this section we give examples showing that each of the 31 nonisomorphic connected graphs with one to five vertices can be realized as the irreducible divisor graph of an element in a primitive minimally generated numerical monoid. In Figure 5, if the positive integers $n_1, \ldots, n_t$ occur as vertices in the graph $G(x)$,
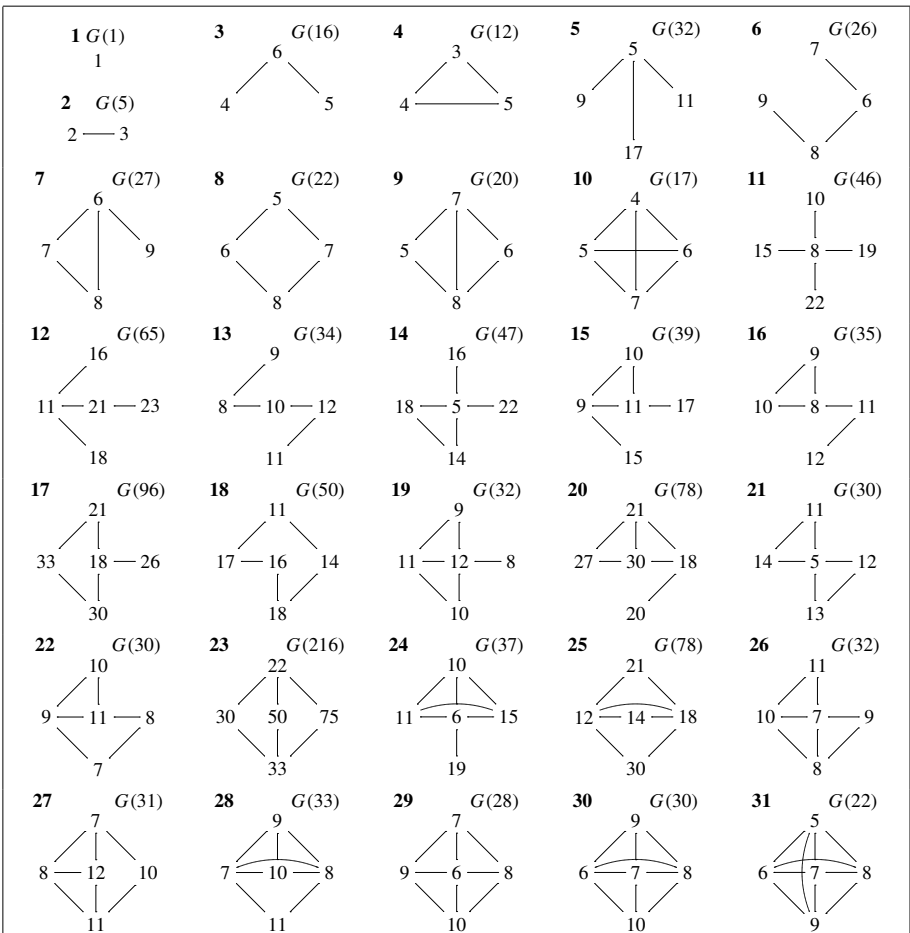

Figure 5. Connected graphs with at most five vertices.

then $x \in N = \langle n_1, \ldots, n_t \rangle$. In Table 2 we give, when possible, a family of examples realizing a given graph using the methods of Section 2. When such a family is not given, it is because that graph cannot be realized as the irreducible divisor graph of an element in a numerical monoid generated by an interval.

Acknowledgement

References

[Axtell and Stickles 2008] M. Axtell and J. Stickles, "Irreducible divisor graphs in commutative rings with zero divisors", *Comm. Algebra* **36**:5 (2008), 1883–1893. MR 2010c:13004 Zbl 1142.13003

[Axtell et al. 2011] M. Axtell, N. R. Baeth, and J. Stickles, "Irreducible divisor graphs and factorization properties of domains", *Comm. Algebra* **39**:11 (2011), 4148–4162. MR 2855118

[Coykendall and Maney 2007] J. Coykendall and J. Maney, "Irreducible divisor graphs", *Comm. Algebra* **35**:3 (2007), 885–895. MR 2008a:13001 Zbl 1114.13001

[García-Sánchez and Rosales 1999] P. A. García-Sánchez and J. C. Rosales, "Numerical semigroups generated by intervals", *Pacific J. Math.* **191**:1 (1999), 75–83. MR 2000i:20095 Zbl 1009.20069

[Maney 2008] J. Maney, "Irreducible divisor graphs, II", *Comm. Algebra* **36**:9 (2008), 3496–3513. MR 2009h:13001 Zbl 1153.13300

[Rosales 1996] J. C. Rosales, "An algorithmic method to compute a minimal relation for any numerical semigroup", *Internat. J. Algebra Comput.* **6**:4 (1996), 441–455. MR 97f:20080 Zbl 0863.20026

dbachman@ucmo.edu   *Department of Mathematics and Computer Science, University of Central Missouri, W. C. Morris 222, Warrensburg, MO 64093, United States*

baeth@ucmo.edu   *Mathematics and Computer Science, University of Central Missouri, W. C. Morris 222, Warrensburg, MO 64093, United States*

cedwards6573@yahoo.com   *Department of Mathematics, University of Oklahoma, Physical Sciences Center 423, Norman, OK 73019, United States*

msp

An application of Google's PageRank
to NFL rankings

Laurie Zack, Ron Lamb and Sarah Ball

(Communicated by Charles R. Johnson)

We explain the PageRank algorithm and its application to the ranking of football teams via the GEM method. We then modify and extend the GEM method with the addition of more football statistics to look at the possibility of using this method to more accurately rank teams. Lastly, we compare both methods by aggregating each statistical ranking using the cross-entropy Monte Carlo algorithm.

1. Introduction

Over the last few decades, abundant research has been done in the mathematics of rankings. There are numerous ranking methods in the field of sports, such as the Massey ratings and Colley matrix, which have been used by the Bowl Championship Series to rank Division I collegiate football teams [BCS 2011]. The search engine Google also uses a mathematical algorithm to compute PageRank, a ranking method used to determine which websites should appear above others in its search results. Google receives 71% of all internet search requests, while the next leading search engine receives only 14% of the requests [SEO 2010], and its PageRank algorithm is one of the main reasons it is the leading search engine on the internet.

There are many factors that determine which websites come up first when you search for something through an internet search engine. On Google, one of those factors is a webpage's PageRank score, and it is this idea of PageRank that set Google apart from other search engines when it was created. The PageRank algorithm assigns a score to each webpage in order to rank the pages according to usefulness. In theory, the most relevant and important pages should come up first in the search results [Wills 2006].

The general concept of the algorithm is to model a random web surfer, starting on one webpage and then clicking on different links to make his or her way through the web. The most "important" webpages are those that have a higher probability of

being seen by the random surfer [Wills 2006]. For a page to have higher probability of being seen, either more webpages have to link to that page, or other highly ranked webpages have to link to it.

2. The mathematics behind PageRank

The exact code and formula for Google's PageRank algorithm are kept secret and it is only known what was first used during the development of Google and PageRank. The algorithm that will be used throughout this paper to show how PageRank is calculated is the one that was originally used by Sergey Brin and Lawrence Page [Brin and Page 1998; Page et al. 1999], the creators of Google, and is most likely not the same one used today. To show how Google calculates PageRank let's consider an internet with only four webpages: A, B, C, and D. The web link diagram below shows how the webpages link to each other, where each arrow represents a link from one page to another. For example, webpage C links to both A and D, but not to B.
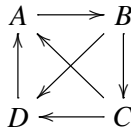
This web link diagram is turned into a web hyperlink matrix H, where

$$H_{ij} = \begin{cases} 1 & \text{if } i \text{ links to } j, \\ 0 & \text{if } i \text{ does not link to } j. \end{cases}$$

Therefore,

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

for this example.

Next, a row stochastic matrix S is formed from H and is then used to model the random web surfer with the equation $G = \alpha S + (1 - \alpha) y v$, where α is defined as the *dampening factor*, y is a column vector of ones, and v is called the *personalization vector*. The vector v is a probability distribution vector, and is currently unknown, but during the development of Google $v = \left(\frac{1}{n} \ \frac{1}{n} \ \cdots \ \frac{1}{n} \right)$ was used [Brin and Page 1998; Page et al. 1999]. The *dampening factor* models the random web surfer's ability to move to a different webpage by means other than following a link, with probability $(1 - \alpha)$. The dampening factor used by Brin and Page during early development was $\alpha = 0.85$. In most research done since 1998, values of α range between 0.85 and 0.99 [Wills 2006]. For this example and throughout the paper,

$\alpha = 0.85$ will be used, and, because there are four webpages in this example, $v = \left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}\right)$. Using the equation $G = 0.85S + 0.15yv$ we obtain the Google matrix G:

$$G = \begin{pmatrix} \frac{3}{80} & \frac{71}{80} & \frac{3}{80} & \frac{3}{80} \\ \frac{3}{80} & \frac{3}{80} & \frac{37}{80} & \frac{37}{80} \\ \frac{37}{80} & \frac{3}{80} & \frac{3}{80} & \frac{37}{80} \\ \frac{71}{80} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \end{pmatrix}.$$

The PageRank vector π is then found by computing the corresponding left eigenvector satisfying $\pi G = \pi$, and, since G is row stochastic, 1 is the dominant eigenvalue, which means π can always be computed [Bryan and Leise 2006]. The i-th entry of π is known as the PageRank score for webpage i. For this particular matrix, the PageRank vector is approximately (0.306 0.297 0.164 0.233). Therefore, the webpage ranking listed from most important to least important is A, B, D, C.

It should be noted that this method is highly inefficient for large matrices, and in 2010 it was estimated that there were approximately a trillion webpages [Kelly 2010]. With such a large and sparse matrix, the power method can be used fairly efficiently to approximate eigenvectors (i.e., to find the PageRank vector) [Bryan and Leise 2006].

3. Dangling node

With the internet constantly growing, many webpages do not link to the majority of the others. In fact, many of them have no out links at all (e.g, postscript files, images). These webpages are known as dangling nodes, and their prevalence leads to a hyperlink matrix which contains mostly zeros. For example, suppose we have the following web link diagram:
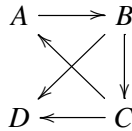
Webpage D would be considered a dangling node, and, in the hyperlink matrix H, row four would be a row of zeros; therefore the matrix would no longer be row stochastic and 1 would no longer be a possible dominant eigenvalue. To fix this, several options exist, one of which is to insert a *personalization vector*, w, into the dangling node rows. It is unknown what Google actually does, but, for this paper, we model the random web surfer's options when on webpage D by assuming he or she has an equal chance to select any other webpage by typing in its URL or to just stay on webpage D, making $w = \left(\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}\right)$. With D being our dangling

NO	48	PHL	22	NO	10	CAR	23	NO	30	CAR	20
NO	35	ATL	27	NO	26	ATL	23	PHL	38	CAR	10
PHL	34	ATL	7	ATL	28	CAR	20	ATL	19	CAR	28

Table 1. Sample 2009 scores.

node we would obtain the following new web hyperlink matrix:

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Calculating as before, the PageRank vector becomes (0.197 0.271 0.219 0.312), producing the ranking D, B, C, A.

4. Using PageRank to rank football teams: GEM 1 method

Applying a similar method to the PageRank algorithm, Govan, Meyer, and Albright [Govan et al. 2008] developed a method called the *GEM method* (which we denoted here by *GEM 1*), using the margin of victory $(v_1 - v_2)$ to weight the "link" between two football teams, where v_1 and v_2 are the teams' scores against each other. As a small sample, the scores in Table 1 were taken from games played in 2009.

By calculating the margin of victory we can create the following link diagram, where each link has a weight equal to the margin of victory:
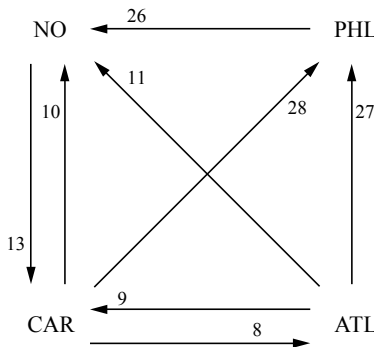
For example, if New Orleans (NO) played Philadelphia (PHL) and the score was NO-48 and PHL-22, a directed arrow would point towards NO with a weight of 26. If a pair of teams played two games and the same team won both times, the weight assigned to the link is the sum of the margins of victory for the two games.

This link diagram then corresponds to the following hyperlink matrix:

$$H = \begin{array}{c} \\ NO \\ PHL \\ ATL \\ CAR \end{array} \overset{\begin{array}{cccc} NO & PHL & ATL & CAR \end{array}}{\left(\begin{array}{cccc} 0 & 0 & 0 & 13 \\ 26 & 0 & 0 & 0 \\ 11 & 27 & 0 & 9 \\ 10 & 28 & 8 & 0 \end{array} \right)}.$$

Next, we continue as before to make it row stochastic and follow the PageRank algorithm to get the final ranking. We obtain (0.330 0.252 0.087 0.332) for the PageRank vector, which produces the ranking 1. CAR, 2. NO, 3. PHL, 4. ATL.

5. Ranking football teams: GEM 2 method

We then modified the GEM method to create what we have termed the *GEM 2 method*. Instead of using the margin of victory to weight one arrow for each game, we used both scores to produce two weighted arrows. Since NO scored 48 points against PHL and PHL scored 22 points against NO, the link diagram will now have one arrow directed from PHL to NO with a weight of 48 and another directed from NO to PHL with a weight of 22. If a pair of teams played two games, we summed each team's scores from the two games. Using the data provided in Table 1, we created a new link diagram as follows:
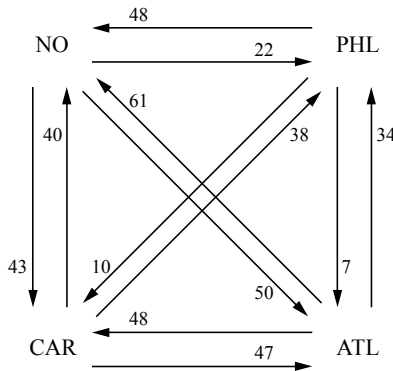
From this diagram the following hyperlink matrix H was then created:

$$H = \begin{array}{c} \\ NO \\ PHL \\ ATL \\ CAR \end{array} \overset{\begin{array}{cccc} NO & PHL & ATL & CAR \end{array}}{\left(\begin{array}{cccc} 0 & 22 & 50 & 43 \\ 48 & 0 & 7 & 10 \\ 61 & 34 & 0 & 48 \\ 40 & 38 & 47 & 0 \end{array} \right)}.$$

The PageRank algorithm gives the PageRank vector (0.317 0.200 0.248 0.335) and the ranking 1. CAR, 2. NO, 3. ATL, 4. PHL.

	Score	Total Yardage	Time of Possession	Turnovers	Actual NFL Ranking
1.	NO	DAL	GB	PHL	NO
2.	NYG	NO	MIN	CAR	MIN
3.	PHL	NYG	DAL	NO	DAL
4.	MIN	MIN	NO	GB	GB
5.	ATL	ATL	NYG	SF	PHL
6.	GB	GB	CAR	TB	ARI
7.	CAR	PHL	ATL	CHI	ATL
8.	DAL	CAR	DET	DET	CAR
9.	CHI	CHI	TB	ATL	SF
10.	ARI	TB	ARI	ARI	NYG
11.	TB	WAS	WAS	DAL	CHI
12.	SF	SEA	CHI	NYG	SEA
13.	WAS	ARI	STL	MIN	WAS
14.	DET	DET	SF	WAS	TB
15.	SEA	STL	SEA	SEA	DET
16.	STL	SF	PHL	STL	STL

Table 2. Final rankings compared to actual rankings using GEM 2.

6. Extended GEM 1 and GEM 2 methods

We collected data on the score, total yardage, turnovers, and time of possession for each regular season game for all 16 teams in the NFL National Football Conference in 2009 [ESPN 2009]. We created four separate H matrices, one for each of the statistics, then proceeded as in Section 5 following the GEM 2 method and the PageRank algorithm using $v = (1/16 \ \ 1/16 \ \cdots \ \ 1/16)$ as our personalization vector. Following the same process as before, we produced a ranking for each statistic collected. However, when calculating turnovers, since it is a negative statistic, we chose to orient the directed arrows in the reverse direction.

Table 2 shows the final rankings for each statistic using the GEM 2 method, and also includes the actual end of the regular season rankings.

In comparison, Table 3 shows the final rankings for each statistic also compared with the actual end of the regular season rankings using the original GEM 1 method.

7. Results

For both GEM 1 and GEM 2, we compared the Kendall rank correlation for each statistic versus the actual rankings which are shown in Table 4. The Kendall rank correlation is defined by $r = (n_c - n_d)/(n(n-1)/2)$, where n_c is the number of

	Score	Total Yardage	Time of Possession	Turnovers	Actual NFL Ranking
1.	DAL	GB	GB	PHL	NO
2.	GB	CHI	DAL	NO	MIN
3.	PHL	MIN	CAR	DAL	DAL
4.	MIN	DAL	MIN	TB	GB
5.	CAR	CAR	SEA	CAR	PHL
6.	NYG	PHL	CHI	GB	ARI
7.	NO	ARZ	NO	NYG	ATL
8.	ARZ	NYG	ARZ	CHI	CAR
9.	TB	NO	ATL	SF	SF
10.	SEA	TB	TB	STL	NYG
11.	ATL	DET	NYG	MIN	CHI
12.	SF	SEA	STL	WAS	SEA
13.	CHI	STL	DET	ARZ	WAS
14.	WAS	SF	WAS	ATL	TB
15.	DET	ATL	PHL	DET	DET
16.	STL	WAS	SF	SEA	STL

Table 3. Final rankings compared to actual rankings using GEM 1.

Statistic	Correlation
SCORE1	0.63
SCORE2	0.60
YARD1	0.38
YARD2	0.67
TIME1	0.32
TIME2	0.35
TURN1	0.32
TURN2	0.25

Table 4. Kendall rank correlations versus actual rankings.

concordant pairs and n_d is the number of discordant pairs in the two rankings. For simplicity, labels of the form STAT1 refer to the GEM 1 method and labels of the form STAT2 refer to the GEM 2 method. Based on the r-values, we can see that each method performed better than the other in different statistics. The ranks were then aggregated using the cross-entropy Monte Carlo algorithm with the distance measure equal to the Kendall tau distance, as this algorithm promotes combining several ordered lists in a proper and efficient manner [Pihur et al. 2009; de Boer et al.

	Aggregate GEM 1	Aggregate GEM 2	Aggregate* GEM 1	Aggregate* GEM 2	Actual NFL Ranking
1.	GB	NO	GB	NO	NO
2.	DAL	GB	DAL	NYG	MIN
3.	CAR	NYG	MIN	MIN	DAL
4.	MIN	MIN	CAR	GB	GB
5.	PHL	DAL	CHI	DAL	PHL
6.	NO	CAR	NO	ATL	ARZ
7.	NYG	ATL	ARZ	CAR	ATL
8.	CHI	CHI	TB	ARZ	CAR
9.	ARZ	ARZ	NYG	CHI	SF
10.	TB	TB	SEA	TB	NYG
11.	SEA	PHL	PHL	PHL	CHI
12.	SF	DET	ATL	WAS	SEA
13.	ATL	SF	DET	DET	WAS
14.	STL	WAS	STL	SF	TB
15.	DET	STL	SF	SEA	DET
16.	WAS	SEA	WAS	STL	STL
r-value	0.53	0.55	0.47	0.60	-

Table 5. Aggregated rankings for both GEM 1 and GEM 2 vs. actual rankings.

2005]. With the aggregate rankings, the GEM 2 method performed only slightly better than the original GEM 1 method, with respective Kendall rank correlations of $r = 0.55$ and $r = 0.53$.

We then decided to take out the least-correlated statistic and aggregate the rankings again. We aggregated twice with the GEM 1 method, once without TIME and once without TURN, since both had equally low r-values, and for the GEM 2 method, we aggregated without TURN. When ignoring the least-correlated statistic, the GEM 2 method performed considerably better, with a Kendall rank correlation of $r = 0.60$, compared to the GEM 1 method, $r = 0.45$ when omitting TURN and $r = 0.47$ when omitting TIME. The aggregated rankings when TIME is omitted from GEM 1 and TURN is omitted from GEM 2 are shown in Table 5 along with the original aggregated rankings and the actual end of season rankings and are denoted by Aggregate*.

There is plenty of other variability in the overall approach to this application of PageRank. We could use more statistics or choose different statistics which are better predictors of overall outcome. We also could use a different dampening factor or modify the personalization vector which could improve the rankings as well.

Nonetheless, it is possible to use this method to produce and compute rankings for any sport or anything else from which a link structure can be created.

References

[BCS 2011] Bowl Championship Series, "Bowl championship series official website", webpage, 2011, http://www.bcsfootball.org.

[de Boer et al. 2005] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method", *Ann. Oper. Res.* **134** (2005), 19–67. MR 2006f:90053 Zbl 1075.90066

[Brin and Page 1998] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", *Computer networks and ISDN systems* **30**:1 (1998), 107–117.

[Bryan and Leise 2006] K. Bryan and T. Leise, "The $25,000,000,000 eigenvector: The linear algebra behind Google", *SIAM Rev.* **48**:3 (2006), 569–581. MR 2008b:15030 Zbl 1115.15007

[ESPN 2009] ESPN NFL, NFL schedule for 2009, 2009, http://espn.go.com/nfl/schedule/_/year/2009.

[Govan et al. 2008] A. Y. Govan, C. D. Meyer, and R. Albright, "Generalizing Google's PageRank to rank National Football League teams", in *Proceedings of the SAS Global Forum*, SAS Global Users Group/SAS Institute, Cary, NC, 2008.

[Kelly 2010] K. Kelly, *What technology wants*, Viking, New York, 2010.

[Page et al. 1999] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web", tech report, Stanford InfoLab, 1999, http://ilpubs.stanford.edu:8090/422/.

[Pihur et al. 2009] V. Pihur, S. Datta, and S. Datta, "RankAggreg, an R package for weighted rank aggregation", *BMC Bioinformatics* **10** (2009), 62.

[SEO 2010] SEO Consultants Directory, "Top search engines for 2010", webpage, 2010, http://www.seoconsultants.com/search-engines/.

[Wills 2006] R. S. Wills, "Google's PageRank: The math behind the search engine", *Math. Intelligencer* **28**:4 (2006), 6–11. MR 2272767

lzack@highpoint.edu   *High Point University, 833 Montlieu Avenue,
                                   High Point, NC 27262, United States*

rlamb@highpoint.edu   *High Point University, 833 Montlieu Avenue,
                                   High Point, NC 27262, United States*

ball.sarah.elizabeth@gmail.com   *High Point University, 833 Montlieu Avenue,
                                   High Point, NC 27262, United States*

msp

Fool's solitaire on graphs

Robert A. Beeler and Tony K. Rodriguez

(Communicated by Joseph Gallian)

In recent work by Beeler and Hoilman, the game of peg solitaire is generalized to arbitrary boards. These boards are treated as graphs in the combinatorial sense. Normally, the goal of peg solitaire is to minimize the number of pegs remaining at the end of the game. In this paper, we consider the open problem of determining the *maximum* number of pegs that can remain at the end of the game, under the restriction that we must jump whenever possible. In this paper, we give bounds for this number. We also determine it exactly for several well-known families of graphs. Several open problems regarding this number are also given.

1. Introduction

Peg solitaire is a table game which traditionally begins with "pegs" in every space except for one which is left empty (i.e., a "hole"). If in some row or column two adjacent pegs are next to a hole (as in Figure 1), then the peg in x can jump over the peg in y into the hole in z. The peg in y is then removed. Usually, the goal is to remove every peg but one. If this is achieved, then the board is considered solved [Beasley 1985; Berlekamp et al. 2003]. However, in this paper we consider the open problem of determining the *maximum* number of pegs that can remain at the end of the game under the caveat that we jump whenever possible. We refer to this variation as the *fool's solitaire problem*.

In [Beeler and Hoilman 2011], the notion of peg solitaire was generalized to graphs. A graph, $G = (V, E)$, is a set of vertices, V, and a set of edges, E. Because of the restrictions of peg solitaire, we will assume that all graphs are finite undirected graphs with no loops or multiple edges. In particular, we will *always* assume that
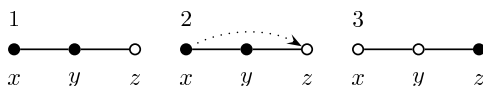

Figure 1. A typical jump in peg solitaire.

graphs are connected. For all undefined graph theory terminology, refer to [West 1996]. In particular, $n(G)$ denotes the *order* of the graph G, that is, the number of vertices in the graph.

If there are pegs in vertices x and y and a hole in z, then we allow x to jump over y into z provided that $xy, yz \in E$. The peg in y is then removed. In general, the game begins with a *starting state* $S \subset V$ which is a set of vertices that are empty. A *terminal state* $T \subset V$ is a set of nonadjacent vertices that have pegs at the end of the game. A terminal state T is *associated* with starting state S if T can be obtained from S by a series of jumps. We will assume that S consists of a single vertex.

The *fool's solitaire number* of a graph G, denoted by $\mathrm{Fs}(G)$, is the cardinality of the largest terminal state T that is associated with a starting state consisting of a single hole. A terminal state T is a *fool's solitaire solution* if the cardinality of T is equal to $\mathrm{Fs}(G)$. The *dual* of a peg configuration T, denoted by T', is the state resulting from reversing the roles of pegs and holes.

The objective of this paper is to gain insight on the fool's solitaire number for graphs. To do this, we will determine bounds of the fool's solitaire number for graphs and find the fool's solitaire number for various classes of graphs. In analyzing the terminal states of a graph, the following theorem is useful.

Theorem 1.1 [Beeler and Hoilman 2011]. *Suppose that S is a starting state of G with associated terminal state T. Let S' and T' be the duals of S and T, respectively. It follows that T' is a starting state of G with associated terminal state S'.*

The following is an immediate corollary that will prove useful.

Corollary 1.2. *On a graph G, there exists some vertex $s \in V(G)$ such that, when $S = \{s\}$, there exists some series of jumps that will yield T as a terminal state if and only if the dual T' of T is solvable to one peg.*

This result provides an alternative method of checking if a suspected terminal state is obtainable. Generally, to determine if a terminal state T of a graph G is obtainable, you simply solve the dual.

2. Upper bounds on Fs(G)

In this section, we present upper bounds for $\mathrm{Fs}(G)$. We begin with a simple, but useful, theorem involving the independence number of a graph. An *independent set* of vertices is a set of mutually nonadjacent vertices. The *independence number* is the maximum size of an independent set in a graph [West 1996].

Theorem 2.1. *For any graph G, $\mathrm{Fs}(G) \leq \alpha(G)$, where $\alpha(G)$ is the independence number of G.*

Proof. By definition, any terminal state is an independent set of vertices. Thus the maximum independent set has at least as many vertices as the largest terminal state. Ergo, $\text{Fs}(G) \leq \alpha(G)$. □

While Theorem 2.1 seems almost trivial, the bound given is sharp for many graphs, as will be discussed in Section 4. Another upper bound involving the domination number follows. In a graph G, a set $S \subseteq V(G)$ is a *dominating set* if every vertex not in S has a neighbor in S. The *domination number* is the minimum size of a dominating set in G.

Theorem 2.2. *For any graph G, $\text{Fs}(G) \leq n(G) - \gamma(G)$, where $\gamma(G)$ is the domination number of G.*

Proof. We begin by showing that the dual of any terminal state is a dominating set. Let T be any terminal state of a graph G. Note that T is an independent set of $V(G)$. Consider T', the dual of T. Since each vertex in a dominating set dominates itself, every vertex not in T is dominated. Also, by definition of an independent set, every vertex in T is adjacent only to vertices in T', so these vertices are dominated as well. Thus T' is a dominating set.

We now show that $\text{Fs}(G) \leq n(G) - \gamma(G)$. Note that $\text{Fs}(G) = |T| = n(G) - |T'|$. Since T' is a dominating set by the argument above, we have that $\gamma(G) \leq |T'|$. Hence $\text{Fs}(G) = n(G) - |T'| \leq n(G) - \gamma(G)$. □

The upper bound given in Theorem 2.1 can be improved for several classes of graphs.

Theorem 2.3. *Let G be a graph. If for every maximum independent set A the dual of A is an independent set with at least two vertices, then $\text{Fs}(G) \leq \alpha(G) - 1$.*

Proof. Suppose to the contrary that $\text{Fs}(G) = \alpha(G)$. This implies that A is a terminal state for some maximum independent set A. Thus, by Corollary 1.2, G would be solvable from starting state A'. Because the dual of A is also an independent set, it follows that no moves are possible from this starting state. Hence either $|A'| = 1$ or $\text{Fs}(G) \leq \alpha(G) - 1$. Since we assume that A' has at least two vertices, $\text{Fs}(G) \leq \alpha(G) - 1$. □

3. Families of graphs

In this section, we present the fool's solitaire number of certain families of graphs. As usual, P_n, C_n, and K_n will denote the path, the cycle, and the complete graph on n vertices, respectively. Let $K_{n,m}$ denote the complete bipartite graph with $V = X \cup Y$, $X = \{x_1, \ldots, x_n\}$, and $Y = \{y_1, \ldots, y_m\}$, where $n \geq m$. In particular, $K_{1,n}$ is called a *star*. The n-dimensional hypercube is denoted by Q_n.

Note that, if $\text{Fs}(G) = \alpha(G)$, it suffices to provide the series of peg solitaire jumps that will yield a solution. If $\text{Fs}(G) = \alpha(G) - 1$, it suffices to demonstrate that

$\mathrm{Fs}(G) \neq \alpha(G)$ and to provide the series of peg solitaire jumps that will yield a terminal state with cardinality $\alpha(G) - 1$.

The following proposition is obvious, but included for the sake of completeness.

Proposition 3.1. *The fool's solitaire number for the complete graph on n vertices is one.*

We now consider complete bipartite graphs.

Proposition 3.2. *For the star $K_{1,n}$, $\mathrm{Fs}(K_{1,n}) = n$.*

Proof. Note that $\alpha(K_{1,n}) = n$. Placing the hole in the center makes it so that no moves are available. Thus $\mathrm{Fs}(G) = n$. □

Theorem 3.3. *For the complete bipartite graph $K_{n,m}$, if $n, m > 1$, then $\mathrm{Fs}(K_{n,m}) = n - 1$.*

Proof. We begin by showing that $\mathrm{Fs}(K_{n,m}) \neq n$. For the complete bipartite graph $K_{n,m}$, note that $\alpha(K_{n,m}) = n$. The only maximum independent set of $K_{n,m}$ is X, which has independent set Y as its dual. Since $|Y| = m > 1$, $\mathrm{Fs}(K_{n,m}) \leq n - 1$ by Theorem 2.3.

We claim that $T = X - \{x_1\}$ is the fool's solitaire solution. Hence we must show that $T' = Y \cup \{x_1\}$ is reducible to a single peg. For $i = 1, \ldots, \lfloor m/2 \rfloor$, we let the $(2i - 1)$-st move be from x_1 over y_{2i-1} into x_2. Similarly, the $2i$-th jump is from x_2 over y_{2i} into x_1. If m is odd, then we make an additional jump from x_1 over y_m into x_2. Since $K_{n,m}$ is solvable from starting state T', it follows that $\mathrm{Fs}(K_{n,m}) = n - 1$ by Corollary 1.2. □

We will now consider the solutions to paths and cycles. When discussing these graphs, we will label the vertices of the graphs with elements of the set $\{0, 1, \ldots, n - 1\}$ in the obvious way. Also note that P_2 and P_3 are isomorphic to $K_{1,1}$ and $K_{1,2}$, respectively. As the fool's solitaire number of these graphs was determined in Proposition 3.2, we do not consider these cases below.

Theorem 3.4. *For the path on n vertices, if $n > 3$, then $\mathrm{Fs}(P_n) = \lfloor n/2 \rfloor$.*

Proof. Note the independence number of a path on n vertices is $\lceil n/2 \rceil$.

We begin by showing that, if n is odd, then $\mathrm{Fs}(P_n) < \lceil n/2 \rceil$. There is only one independent set with cardinality $\lceil n/2 \rceil$, namely $\{0, 2, 4, \ldots, n-3, n-1\}$. Because the dual of this set is an independent set with at least two vertices, $\mathrm{Fs}(P_{2k+1}) \leq \lfloor n/2 \rfloor$ by Theorem 2.3.

To obtain the fool's solitaire solution for P_n (regardless of whether n is even or odd), begin with the hole in 0. The i-th move will be to use the peg in $2i$ to jump over $2i - 1$ into $2i - 2$. This will remove $\lceil n/2 \rceil$ pegs. It follows that $\mathrm{Fs}(P_n) = \lfloor n/2 \rfloor$. □

Theorem 3.5. *For the cycle on n vertices, $\mathrm{Fs}(C_n) = \lfloor \frac{n-1}{2} \rfloor$.*

Proof. Note that $\alpha(C_n) = \lfloor n/2 \rfloor$. We begin by showing that if n is even, then $\mathrm{Fs}(C_n) < n/2$. Let $n = 2k$, where $k \in \mathbb{Z}$. Up to automorphism on the vertices, C_{2k} has one maximum independent set of vertices. Since the dual of this set is an independent set with at least two vertices, it follows that $\mathrm{Fs}(C_{2k}) \leq k - 1$ by Theorem 2.3.

To obtain the fool's solitaire solution of C_n (regardless of whether n is even or odd), begin with the hole in 0. The i-th move will be to use the peg in $2i$ to jump over $2i - 1$ into $2i - 2$. This can be repeated k times, removing $\lceil n/2 \rceil$ pegs. If n is even, we make an additional jump from 0 over $n - 1$ into $n - 2$. In either case, $\mathrm{Fs}(C_n) = \lfloor \frac{n-1}{2} \rfloor$. $\qquad\square$

We will now consider the hypercube on 2^n vertices, Q_n. As usual, each vertex will be labeled with an element from the set $\{0, 1, \ldots, 2^n - 1\}$, with two vertices being adjacent if and only if their binary expansions differ by one bit.

Theorem 3.6. *The fool's solitaire number of the n-dimensional hypercube for $n \geq 2$ is $\mathrm{Fs}(Q_n) = 2^{n-1} - 1$.*

Proof. We first show that $\mathrm{Fs}(Q_n) \neq \alpha(Q_n) = 2^{n-1}$. Up to automorphism on the vertices, there is a unique maximum independent set of vertices, namely the set of all vertices whose binary expansions have an even number of ones. As the dual of this set is an independent set with at least two vertices, $\mathrm{Fs}(Q_n) \leq 2^{n-1} - 1$.

Note that Q_n is Hamiltonian with an even number of vertices [Harary et al. 1988]. Relabel the vertices of Q_n along a Hamiltonian cycle with the numbers $0, 1, \ldots, 2^n - 1$ in the obvious way. Note that the odd-numbered vertices correspond to the vertices with an odd number of ones in their binary expansions. Hence, the odd-numbered vertices form a maximum independent set in Q_n. We claim that $\{1, 3, \ldots, 2^n - 3\}$ is the fool's solitaire solution. Hence we must show that the dual of this set, $\{2^n - 1, 0, 2, 4, \ldots, 2^n - 2\}$, is reducible to a single peg. Begin by jumping from $2^n - 1$ over 0 into 1. For the remaining $2^{n-1} - 1$ moves, the i-th move is from $2i - 1$ over $2i$ into $2i + 1$, where $i = 1, \ldots, 2^{n-1} - 1$. Hence $\mathrm{Fs}(Q_n) = 2^{n-1} - 1$. $\qquad\square$

4. Lower bounds on Fs(G)

In Section 2, we gave several upper bounds on the fool's solitaire number. Unfortunately, lower bounds on the fool's solitaire number are more difficult to prove in general. However, a useful proposition follows.

Proposition 4.1. *Suppose that H is obtained from G by appending a vertex that is not adjacent to any vertex in the fool's solitaire solution of G. It follows that $\mathrm{Fs}(H) \geq \mathrm{Fs}(G) + 1$.*
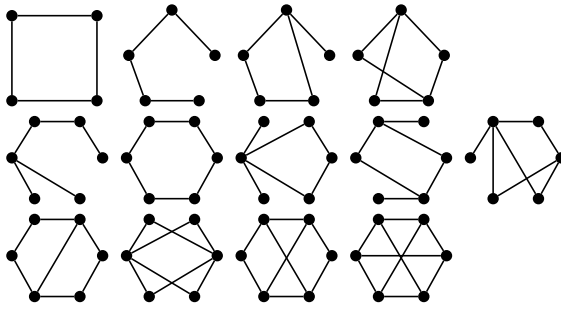
Figure 2. Graphs with $n(G) \leq 6$ such that $\mathrm{Fs}(G) = \alpha(G) - 1$.

Proof. Suppose that H is obtained from G by appending a vertex v' to G such that $vv' \notin E(G)$ for all $v \in T$, where T is the fool's solitaire solution of G. We obtain a terminal state of H with $|T| + 1$ vertices by finding the fool's solitaire solution on the subgraph induced by the vertices of G. Since v' is not adjacent to any vertex in T, it follows that $T \cup \{v'\}$ is a valid terminal state of H. This terminal state has $\mathrm{Fs}(G) + 1$ vertices. Hence, $\mathrm{Fs}(H) \geq \mathrm{Fs}(G) + 1$.   □

To aid in a more general result, an exhaustive computer search of all terminal states associated with a single vertex starting state was performed on all 143 nonisomorphic connected graphs with six vertices or less. The algorithm is implemented on the first author's website [Beeler and Norwood n.d.].

Lists of graphs of small order were obtained from the appendix of [Harary 1969]. The independence numbers of these graphs were verified using the Small Graph Database [Grout n.d.].

Of the 143 connected graphs with six vertices or less, 130 of them satisfy $\mathrm{Fs}(G) = \alpha(G)$. The remaining thirteen graphs satisfy $\mathrm{Fs}(G) = \alpha(G) - 1$. These graphs are given in Figure 2.

Based on this and the results of Section 3, we present the following conjecture.

Conjecture 4.2. For all connected graphs G,

$$\alpha(G) - 1 \leq \mathrm{Fs}(G) \leq \alpha(G).$$

While we were unable to prove this, Proposition 4.1 may prove useful for an inductive proof of this conjecture.

5. Open problems

Let H be a graph obtained from G by deleting an edge of G. We note that $\alpha(H) \geq \alpha(G)$ for all graphs G. Thus, a natural conjecture is that $\mathrm{Fs}(H) \geq \mathrm{Fs}(G)$ for all graphs G. However, this is not the case. Using the aforementioned exhaustive computer search on all graphs with six vertices or less, three were found in which
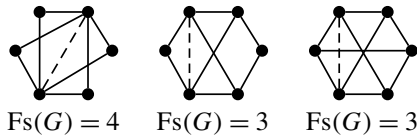
$$\text{Fs}(G) = 4 \qquad \text{Fs}(G) = 3 \qquad \text{Fs}(G) = 3$$

Figure 3. Graphs in which edge deletion lowers $\text{Fs}(G)$.

edge deletion actually *lowers* the fool's solitaire number. These graphs are given in Figure 3. In each of these cases, deleting the dashed edge will lower the fool's solitaire number by one.

Some natural open questions motivated by this observation include:

(i) How much can edge deletion lower the fool's solitaire number?

(ii) Let $\text{ED}(n)$ be the number of nonisomorphic graphs with n vertices such that edge deletion lowers the fool's solitaire number. If n is large enough, does $\text{ED}(n) = 0$? Let $i(n)$ be the number of nonisomorphic graphs with n vertices. What can be said about $\lim_{n\to\infty} \text{ED}(n)/i(n)$?

One of the major results in [Beeler and Hoilman 2011] was to show that the cartesian product of solvable graphs was likewise solvable. What can be said about $\text{Fs}(G \,\square\, H)$ in terms of $\text{Fs}(G)$ and $\text{Fs}(H)$?

Acknowledgments

References

[Beasley 1985] J. D. Beasley, *The ins and outs of peg solitaire*, Recreations in Mathematics **2**, Oxford University Press, Eynsham, 1985. MR 87c:00002

[Beeler and Hoilman 2011] R. A. Beeler and D. P. Hoilman, "Peg solitaire on graphs", *Discrete Math.* **311**:20 (2011), 2198–2202. MR 2012g:05153 Zbl 1230.05211

[Beeler and Norwood n.d.] R. A. Beeler and H. Norwood, "Solitaire solver: peg solitaire on graphs solver applet", Software, East Tennesse State University, Johnson City, TN, http://faculty.etsu.edu/BEELERR/solitaire.

[Berlekamp et al. 2003] E. R. Berlekamp, J. H. Conway, and R. K. Guy, *Winning ways for your mathematical plays*, vol. 2, 2nd ed., A K Peters, Natick, MA, 2003. MR 2004d:91001 Zbl 1011.00009

[Grout n.d.] J. Grout, "Graph database", Drake University, Des Moines, IA, http://artsci.drake.edu/grout/graphs.

[Harary 1969] F. Harary, *Graph theory*, Addison-Wesley, Reading, MA, 1969. MR 41 #1566 Zbl 0182.57702

[Harary et al. 1988] F. Harary, J. P. Hayes, and H.-J. Wu, "A survey of the theory of hypercube graphs", *Comput. Math. Appl.* **15**:4 (1988), 277–289. MR 89i:05230 Zbl 0645.05061

[West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR 96i:05001 Zbl 0845.05001

beelerr@etsu.edu   *Department of Mathematics and Statistics, East Tennessee State University, Johnson City, TN 37614, United States*

ztkr2@goldmail.etsu.edu   *Department of Mathematics and Statistics, East Tennessee State University, Johnson City, TN 37614, United States*

Newly reducible iterates in families of quadratic polynomials

Katharine Chamberlin, Emma Colbert, Sharon Frechette,
Patrick Hefferman, Rafe Jones and Sarah Orchard

(Communicated by Michael Zieve)

We examine the question of when a quadratic polynomial $f(x)$ defined over a number field K can have a newly reducible n-th iterate, that is, $f^n(x)$ irreducible over K but $f^{n+1}(x)$ reducible over K, where f^n denotes the n-th iterate of f. For each choice of critical point γ, we consider the family

$$g_{\gamma,m}(x) = (x - \gamma)^2 + m + \gamma, \quad m \in K.$$

For fixed $n \geq 3$ and nearly all values of γ, we show that there are only finitely many m such that $g_{\gamma,m}$ has a newly reducible n-th iterate. For $n = 2$ we show a similar result for a much more restricted set of γ. These results complement those obtained by Danielson and Fein (*Proc. Amer. Math. Soc.* **130**:6 (2002), 1589–1596) in the higher-degree case. Our method involves translating the problem to one of finding rational points on certain hyperelliptic curves, determining the genus of these curves, and applying Faltings' theorem.

1. Introduction

Let K be a number field and $f(x) \in K[x]$. By the n-th iterate $f^n(x)$ of $f(x)$, we mean the n-fold composition of f with itself. Determining the factorization of $f^n(x)$ into irreducible polynomials has proven to be an important problem. From a dynamical perspective, it is a question about the inverse orbit of zero, namely $O^-(0) := \bigcup_{n \geq 1} f^{-n}(0)$. This set has significance in various ways; for instance, it accumulates at every point of the Julia set of f [Beardon 1991, p. 71]. The field of arithmetic dynamics seeks to understand sets such as $O^-(0)$ from an algebraic perspective, and finding the factorization of $f^n(x)$ fits into this scheme: a nontrivial factorization arises from an "unexpected" algebraic relation among

elements of $O^-(0)$. In addition, understanding the factorization of $f^n(x)$ has proven to be a key obstacle in determining the Galois groups of $f^n(x)$ (see [Hamblen et al. 2013; Jones 2008] or [Jones and Manes 2011] for the case of some rational functions). These Galois groups provide a sort of dynamical analogue to the well-studied ℓ-adic Galois representations [Boston and Jones 2007].

In general, the factorization of the iterates of f can exhibit a wide variety of behaviors. For instance, in [Fein and Schacher 1996, Lemma 1.1] it is shown that for each $n \geq 1$ and $d \geq 2$, there exists a number field K such that, for some $f(x) \in K[x]$ of degree d, $f^{n+1}(x)$ is *newly reducible*; that is, $f^n(x)$ is irreducible over K but $f^{n+1}(x)$ is reducible over K. More specifically, it follows from [Stoll 1992, p. 243] and [Fein and Schacher 1996, Lemma 1.1] that if $f(x) = x^2 + m$ for $m \in \mathbb{Z}_{>0}$, $m \equiv 1, 2 \mod 4$, then for any fixed $n \geq 1$ there exists a number field K such that $f^{n+1}(x)$ is newly reducible over K. But what happens when we fix the number field K to start with, and ask about the factorization of $f^n(x)$ as n grows? Many authors have examined this question, in general with the aim of giving criteria that ensure all iterates are irreducible (see, e.g., [Jones 2012; Odoni 1985, Section 4]). Most usefully for our purposes, Danielson and Fein [2002] consider the case when $f(x) = x^d + m$, for $d \geq 2$. They show, for instance, that if $m \in \mathbb{Z}$ and $f(x)$ is irreducible, then all iterates of f are irreducible. In fact they only assume that K is the quotient field of a unique factorization domain R, and in this case they show that certain strong diophantine conditions must be satisfied when $f^n(x)$ is irreducible and $f^{n+1}(x)$ is reducible. In particular, for $K = \mathbb{Q}$, they take $S(d, n)$ to be the set of $m \in \mathbb{Q}$ such that $f^{n+1}(x)$ is newly reducible. Further, let $S(d) = \bigcup_{n \geq 1} S(d, n)$. In [Danielson and Fein 2002, Theorem 7] it is shown that $S(2, 1)$ (and thus $S(2)$) is infinite, $S(3, n)$ is finite for all $n \geq 1$, and $S(d)$ is finite for d odd, $d \geq 5$. Moreover, the abc conjecture implies that $S(d)$ is finite for d even, $d \geq 4$.

One goal of the present paper is to determine whether $S(2, n)$ is finite for $n \geq 2$. Our main result, however, is significantly more general. Consider the family of polynomials

$$g_{\gamma,m}(x) = (x - \gamma)^2 + m + \gamma, \quad \gamma, m \in K, \tag{1-1}$$

where K is a number field. Denote the ring of integers of K by $\mathbb{O}_K$. Our main result is the following:

Theorem 1. *Let K be a number field, $v_{\mathfrak{p}}$ the valuation attached to a prime $\mathfrak{p}$ of $\mathbb{O}_K$, and $g_{\gamma,m}(x)$ as in (1-1). If one of the following holds, then there are only finitely many m such that $g_{\gamma,m}^n(x)$ is irreducible over K and $g_{\gamma,m}^{n+1}(x)$ is reducible over K:*

(1) *$n \geq 3$ and there exists a prime $\mathfrak{p}$ of $\mathbb{O}_K$ with $v_{\mathfrak{p}}(2) = e \geq 1$ and $v_{\mathfrak{p}}(\gamma) = s$ with $s \neq -e2^i$ for all $i \geq 1$;*

(2) *$n = 2$ and $\gamma = r/4$ for for $r \in \mathbb{Z}$ such that $-200 \leq r \leq 200$.*

In particular, when $K = \mathbb{Q}$, part (1) of Theorem 1 holds when $v_2(\gamma)$ is not of the form -2^j for $j \geq 1$. Hence when $\gamma = 0$, we obtain that $S(2, n)$ is finite for $n \geq 2$ (in the notation of [Danielson and Fein 2002]); in other words, for each $n \geq 2$ there are at most finitely many $m \in \mathbb{Q}$ such that $x^2 + m$ has a newly reducible $(n+1)$-st iterate. In Proposition 10, we show further that $S(2, 3)$ is empty. Note also that part (1) of Theorem 1 applies whenever γ belongs to the ring of integers of K, and in particular for $\gamma \in \mathbb{Z}$. In fact, part (1) holds whenever γ is taken so that

$$g_{\gamma,m}^i(\gamma) \in K[m] \text{ does not have repeated roots for any } i \geq 1. \qquad (1\text{-}2)$$

(See Theorem 6, Proposition 9, and the discussion immediately before Proposition 9.) Condition (1-2) is the same as the condition appearing in [Faber et al. 2009] for the *preimage curve* $Y^{\text{pre}}(i, -\gamma)$, given by the vanishing of the polynomial

$$(g_{0,m}^i(x) + \gamma) \in K[x, m],$$

to be nonsingular for all $i \geq 1$. In Proposition 9, we give a new criterion ensuring that (1-2) holds for given γ, thereby improving [Faber et al. 2009, Proposition 4.8]. The full strength of condition (1-2) is not required to prove part (1) of Theorem 1; see the remark following the proof of Proposition 9.

For given K, denote by $S(2, n, \gamma)$ the set of $m \in K$ such that $g_{\gamma,m}^{n+1}(x)$ is newly reducible. Thus Theorem 1 establishes the finitude of $S(2, n, \gamma)$ for $n \geq 2$ and certain γ. In Theorem 3, we show that for each $\gamma \in K$, the set $S(2, 1, \gamma)$ is infinite, and we explicitly describe its elements. In the case $\gamma = 0$, this result follows from [Danielson and Fein 2002, Proposition 2]. When $n \geq 2$, the sets $S(2, n, \gamma)$ may still be nonempty, even for $K = \mathbb{Q}$. For instance, when $f(x) = x^2 - x - 1$, corresponding to $\gamma = \frac{1}{2}$ and $m = -\frac{7}{4}$, we have that $f(x)$ and $f^2(x)$ are irreducible but

$$f^3(x) = (x^4 - 3x^3 + 4x - 1)(x^4 - x^3 - 3x^2 + x + 1), \qquad (1\text{-}3)$$

and thus $-\frac{7}{4} \in S\left(2, 2, \frac{1}{2}\right)$. For $K = \mathbb{Q}$, the sets $S(2, n, \gamma)$ are likely to be empty for $n \geq 3$, since as we will see they correspond to rational points on high-genus curves. However, without effective algorithms to find such points, a new approach will be required to precisely determine $S(2, n, \gamma)$.

To prove Theorem 1, we first examine the case where $n \geq 3$ and use the fact that comparing constant terms of a hypothetical nontrivial factorization of $g_{\gamma,m}^{n+1}(x)$ gives rise to K-rational points on a hyperelliptic curve (at least for the γ satisfying part (1) of Theorem 1). This allows us to use Faltings' theorem to conclude that $S(2, n, \gamma)$ is finite for these γ and for $n \geq 3$. We then examine the case $n = 2$ using a system of equations generated from a factorization of the third iterate. After defining certain cases for this system, we use Faltings' theorem on a plane curve arising from the Gröbner basis of the system to show that $S(2, 2, \gamma)$ is finite for certain γ.

2. The case $n = 1$

Before we approach the main theorem, let's examine the case where $n = 1$. It is possible for $g_{\gamma,m}^2(x)$ to be reducible and $g_{\gamma,m}(x)$ irreducible:

Example 2. Let $\gamma = 0$, $m = -\frac{4}{3}$, and $K = \mathbb{Q}$. Then

$$g_{0,-\frac{4}{3}}(x) = x^2 - \frac{4}{3}$$

is irreducible over $\mathbb{Q}$ since $\frac{4}{3}$ is not a rational square. However, we have

$$g_{0,-\frac{4}{3}}^2(x) = \left(x^2 - \frac{4}{3}\right)^2 - \frac{4}{3} = \left(x^2 - 2x + \frac{2}{3}\right)\left(x^2 + 2x + \frac{2}{3}\right).$$

Because it has degree 4, $g_{\gamma,m}^2(x)$ could a priori have nontrivial factors of degree 1, 2, or 3. We will show in Corollary 5 that if $g_{\gamma,m}(x)$ is irreducible, then the only nontrivial factorization for $g_{\gamma,m}^2(x)$ is $p_1(x)p_2(x)$, with $\deg p_1(x) = \deg p_2(x) = 2$.

Theorem 3. *We have $g_{\gamma,m}(x)$ irreducible and $g_{\gamma,m}^2(x)$ reducible if and only if either*

(1) $\gamma \neq \frac{1}{4}$ *and* $m = (c_1^4 - 4\gamma)/(4 - 4c_1^2)$, *where* $c_1 \in K \setminus \{-1, 1\}$ *and* $(4\gamma - c_1^2)/(1 - c_1^2)$ *is not a square in K; or*

(2) $\gamma = \frac{1}{4}$ *and* $-4m - 1$ *is not a square in K.*

In particular, for each $\gamma \in K$, the set $S(2, 1, \gamma)$ is infinite.

Remark. It is interesting to note that when $\gamma = \frac{1}{4}$, we have

$$g_{1/4,m}^2(x) = \left(x^2 - \frac{3}{2}x + \left(m + \frac{13}{16}\right)\right)\left(x^2 + \frac{1}{2}x + \left(m + \frac{5}{16}\right)\right), \qquad (2\text{-}1)$$

and so $g_{1/4,m}^2(x)$ is reducible for all $m \in K$. This phenomenon has already been noticed, albeit in somewhat different language, in [Faber et al. 2009, Remark 2.6 and p. 94].

Proof. Suppose that $g_{\gamma,m}(x)$ is irreducible and $g_{\gamma,m}^2(x)$ is reducible, so that $g_{\gamma,m}^2(x) = p_1(x)p_2(x)$. Write $p_1(x) = (x - \gamma)^2 + b_1(x - \gamma) + b_0$ and $p_2(x) = (x - \gamma)^2 + c_1(x - \gamma) + c_0$, where b_i, $c_i \in K$, and note that

$$g_{\gamma,m}^2(x) = (x - \gamma)^4 + 2m(x - \gamma)^2 + m^2 + m + \gamma. \qquad (2\text{-}2)$$

Comparing coefficients in the equality $g_{\gamma,m}^2(x) = p_1(x)p_2(x)$ gives the following system of equations:

(a) $c_1 + b_1 = 0$;

(c) $b_1c_0 + b_0c_1 = 0$;

(b) $c_0 + b_1c_1 + b_0 = 2m$;

(d) $b_0c_0 = m^2 + m + \gamma$.

Clearly $b_1 = -c_1$ from (a), and then from (c) we have $c_1(b_0 - c_0) = 0$. If $c_1 = 0$, then from (b) we obtain $c_0 + b_0 = 2m$. Squaring both sides and subtracting four times

equation (d), one verifies that $-m - \gamma = \frac{1}{4}(c_0 - b_0)^2$. As this is a square, $g_{\gamma,m}(x)$ is reducible (see (1-1) on page 482), and from this contradiction we conclude that $c_1 \neq 0$, and hence $b_0 = c_0$. See (3-1) in the proof of Theorem 6 for a generalization of this statement. From (b) and (d) we now derive the following system of two equations:

(e) $2c_0 - c_1^2 - 2m = 0$;

(f) $c_0^2 - m^2 - m - \gamma = 0$.

Solving (e) for c_0 and substituting the result into (f) gives

$$c_1^4 + 4mc_1^2 - 4m - 4\gamma = 0. \tag{2-3}$$

Note that $c_1 = \pm 1$ if and only if $\gamma = \frac{1}{4}$. Thus in the case where $\gamma \neq \frac{1}{4}$, we may solve (2-3) for m to obtain $m = (c_1^4 - 4\gamma)/(4 - 4c_1^2)$. Because $g_{\gamma,m}(x)$ is assumed to be irreducible, we have that $-m - \gamma$ is not a square in K, and one computes $-m - \gamma = (c_1^2(4\gamma - c_1^2))/(4(1 - c_1^2))$. In the case where $\gamma = \frac{1}{4}$, we may take $c_1 = \pm 1$ and $c_0 = (1 + 2m)/2$ to get a solution to equations (e) and (f) (this is the same as the factorization in (2-1)). Hence $g_{1/4,m}^2(x)$ is reducible for all $m \in K$. Since $g_{1/4,m}(x)$ is assumed to be irreducible, $-m - \gamma = -m - \frac{1}{4}$ cannot be a square in K, which holds if and only if $-4m - 1$ is not a square in K.

Assume now that either of the conditions in the statement of Theorem 3 hold. Then $-m - \gamma$ is not a square in K, so $g_{\gamma,m}(x)$ is irreducible. The other hypotheses ensure that equations (e) and (f) above have solutions in K, and hence $g_{\gamma,m}^2(x)$ is reducible. $\qquad \square$

Note that when $\gamma = 0$, taking $c_1 = 2$ in Theorem 3 yields Example 2. We also remark that in the case of $\gamma = 0$, taking $c_1 = 2z$ in Theorem 3 yields Proposition 2 of [Danielson and Fein 2002], at least in the case where K is a number field. (Note that there the polynomial under consideration is $x^2 - m$, and hence the results differ by a minus sign.)

3. The case $n \geq 3$

Having handled the case $n = 1$, we now address the case where $n \geq 3$. We postpone the case $n = 2$ until Section 4 because the curves we must analyze have genus one, while for $n \geq 3$ the curves that arise have genus at least two, allowing us to apply Faltings' theorem.

Understanding the roots of $g_{\gamma,m}^{n+1}(x)$ is central to our analysis. In general, if β_i is a root of $g_{\gamma,m}^n(x)$, then the two roots of $g_{\gamma,m}(x) - \beta_i$ are roots of $g_{\gamma,m}^{n+1}(x)$. Calling them α_i^+ and α_i^-, we have $\alpha_i^+ = \gamma + \sqrt{\beta_i - m - \gamma}$ and $\alpha_i^- = \gamma - \sqrt{\beta_i - m - \gamma}$. Note that

$$2\gamma - \alpha_i^+ = 2\gamma - \left(\gamma + \sqrt{\beta_i - m - \gamma}\right) = \gamma - \sqrt{\beta_i - m - \gamma} = \alpha_i^-.$$

The following picture summarizes the relation of the roots to one another. Note that they are arranged in a tree.
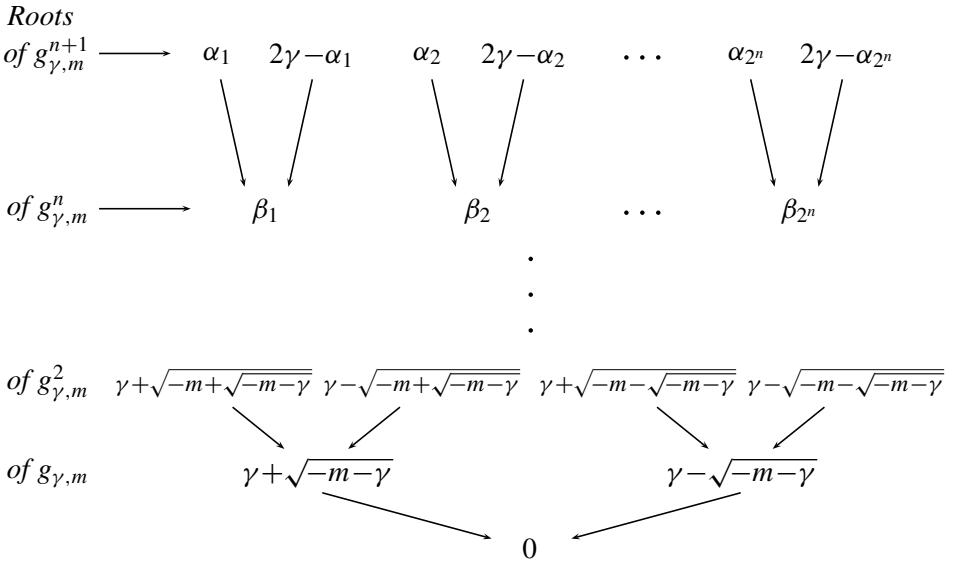
In this section we establish two principal results on the structure of hypothetical factors in the case where $g_{\gamma,m}^{n+1}(x)$ is newly reducible. Our first result is similar to [Jones and Boston 2012, Proposition 2.6].

Theorem 4. *Let $g_{\gamma,m}(x) = (x - \gamma)^2 + m + \gamma$ with $\gamma, m \in K$. Suppose $g_{\gamma,m}^n(x)$ is irreducible, and $g_{\gamma,m}^{n+1}(x) = p_1(x) p_2(x)$ where $p_1(x)$ and $p_2(x)$ are nontrivial factors. If α is a root of $p_1(x)$, then $2\gamma - \alpha$ is a root of $p_2(x)$ but not a root of $p_1(x)$.*

Proof. Let $G_{n+1} = \mathrm{Gal}(E_{n+1}/K)$, where E_{n+1} is the splitting field of $g_{\gamma,m}^{n+1}(x)$ over K. Because $g_{\gamma,m}^n(x)$ is irreducible over K, G_{n+1} acts transitively on the roots of $g_{\gamma,m}^n(x)$. Let α be a root of $p_1(x)$ and α' be a root of $g_{\gamma,m}^{n+1}$ but not a root of p_1. By the transitivity of the action of G_{n+1} on the roots of $g_{\gamma,m}^n$, we may take $\phi \in G_{n+1}$ such that $\phi(g_{\gamma,m}(\alpha)) = g_{\gamma,m}(\alpha')$. Hence

$$\phi((\alpha - \gamma)^2 + \gamma + m) = (\alpha' - \gamma)^2 + \gamma + m,$$

from which we deduce that $\phi(\alpha) - \gamma = \pm(\alpha' - \gamma)$. Indeed, we must have $\phi(\alpha) - \gamma = -(\alpha' - \gamma)$, for otherwise $\phi(\alpha) = \alpha'$, contradicting our assumption that α' is not a root of p_1. We thus obtain $\phi(\alpha) = 2\gamma - \alpha'$. In other words, $2\gamma - \alpha = \phi^{-1}(\alpha')$, and is therefore not a root of p_1. $\square$

Corollary 5. *Let $g_{\gamma,m}(x) = (x - \gamma)^2 + m + \gamma$ with $\gamma, m \in K$. Let $n \in \mathbb{Z}^+$, and assume $g_{\gamma,m}^n(x)$ is irreducible with $g_{\gamma,m}^{n+1}(x) = p_1(x) p_2(x)$, where $p_1(x)$ and $p_2(x)$ are nontrivial factors. Then, $\deg p_1(x) = \deg p_2(x) = 2^n$, and $p_1(x)$ and $p_2(x)$ are irreducible.*

Proof. Observe that $\deg g_{\gamma,m}^n(x) = 2^n$ and $\deg g_{\gamma,m}^{n+1}(x) = 2^{n+1}$. By Theorem 4, the roots of $p_1(x)$ are in bijection with the roots of $p_2(x)$, whence $\deg p_1(x) = \deg p_2(x) = 2^n$. If $\{\alpha_1, \ldots, \alpha_{2^n}\}$ are all the roots of $p_1(x)$, then by Theorem 4, $\{2\gamma - \alpha_1, \ldots, 2\gamma - \alpha_{2^n}\}$ are all the roots of $p_2(x)$. Thus the set

$$\{g_{\gamma,m}(\alpha_i) : i = 1, \ldots, 2^n\}$$

coincides with the set of all roots of $g_{\gamma,m}^n(x)$. Because $g_{\gamma,m}^n(x)$ is irreducible, the action of G_{n+1} on $\{g_{\gamma,m}(\alpha_i) : i = 1, \ldots, 2^n\}$ consists of a single orbit, and thus the action of G_{n+1} on $\{\alpha_1, \ldots, \alpha_{2^n}\}$ must consist of a single orbit. Hence $p_1(x)$ is irreducible. Similar reasoning gives that $p_2(x)$ is irreducible. $\qquad\square$

3.1. *Curves and Faltings' theorem.* We now use Theorem 4 to show that if $g_{\gamma,m}^{n+1}(x)$ is newly reducible, then there is a K-rational point, depending on m, on a certain curve.

Theorem 6. *If $g_{\gamma,m}^n(x)$ is irreducible and $g_{\gamma,m}^{n+1}(x)$ is reducible for some $n \geq 1$, then there exist $x, y \in K$ with $x = m$ such that*

$$y^2 = t_{n+1}(x),$$

where the polynomials $t_i(x)$ are defined by the recurrence relation $t_1(x) = x + \gamma$ and, for $i \geq 2$,

$$t_i(x) = (t_{i-1}(x) - \gamma)^2 + x + \gamma.$$

Remark. Note that $t_i(x) = (g_{\gamma,m}^i(\gamma))|_{m=x}$, as will be shown below (or can be easily seen by induction).

Proof. Assume $g_{\gamma,m}^n$ is irreducible and $g_{\gamma,m}^{n+1}(x) = p_1(x)p_2(x)$ for some $p_1(x), p_2(x) \in K[x]$ of positive degree. By Theorem 4, if $\{\alpha_1, \ldots, \alpha_{2^n}\}$ are all the roots of $p_1(x)$, then $\{2\gamma - \alpha_1, \ldots, 2\gamma - \alpha_{2^n}\}$ are all the roots of $p_2(x)$. Then,

$$p_1(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_{2^n}) \quad \text{and}$$
$$p_2(x) = (x - (2\gamma - \alpha_1))(x - (2\gamma - \alpha_2)) \cdots (x - (2\gamma - \alpha_{2^n}))$$
$$= (x - 2\gamma + \alpha_1)(x - 2\gamma + \alpha_2) \cdots (x - 2\gamma + \alpha_{2^n}).$$

So we have

$$p_1(\gamma) = (\gamma - \alpha_1)(\gamma - \alpha_2) \cdots (\gamma - \alpha_{2^n}) \quad \text{and}$$
$$p_2(\gamma) = (-\gamma + \alpha_1)(-\gamma + \alpha_2) \cdots (-\gamma + \alpha_{2^n})$$
$$= (-1)^{2^n} (\gamma - \alpha_1)(\gamma - \alpha_2) \cdots (\gamma - \alpha_{2^n}), \qquad (3\text{-}1)$$

and therefore $p_1(\gamma) = p_2(\gamma)$. Set $y = p_1(\gamma) = p_2(\gamma)$, so $g_{\gamma,m}^{n+1}(\gamma) = y^2$. We have

$$g_{\gamma,m}^{n+1}(\gamma) = g_{\gamma,m}(g_{\gamma,m}^n(\gamma)) = (g_{\gamma,m}^n(\gamma) - \gamma)^2 + m + \gamma.$$

Moreover, $g_{\gamma,m}(\gamma) = m + \gamma$, and thus $g^i_{\gamma,m}(\gamma)$ satisfies the same recurrence relation as $t_i(x)$, with x replaced by m.   □

The polynomials $t_i(x)$ play a critical role in our argument. The first few are

$$t_1(x) = x + \gamma, \quad t_2(x) = x^2 + x + \gamma, \quad t_3(x) = x^4 + 2x^3 + x^2 + x + \gamma,$$
$$t_4(x) = x^8 + 4x^7 + 6x^6 + 6x^5 + 5x^4 + 2x^3 + x^2 + x + \gamma. \tag{3-2}$$

Equations of the form $y^2 = t_i(x)$ may be interpreted geometrically as plane curves. A plane curve defined over a field F is the set of solutions $(x, y) \in F \times F$ of an equation of the form $h(x, y) = 0$, where $h(x, y) \in F[x, y]$. If K is a subfield of F, a K-rational point on the curve is one whose coordinates lie in K. For instance, $(1, -1)$ is a $\mathbb{Q}$-rational point on the curve $y^2 = x^3 + x - 1$, while $(-1, \sqrt{-3})$ is not (though it is K-rational for $K = \mathbb{Q}(\sqrt{-3})$).

The genus of a plane curve is a measure of its geometric complexity, and for curves of the form $y^2 = r(x)$, which is the case of interest to us in light of Theorem 6, there is a convenient way to calculate it — at least, when the roots of $r(x)$ in the algebraic closure of K are distinct.

Theorem 7 [Goldschmidt 2003]. *Consider the curve $C : y^2 = r(x)$. If $r(x)$ is separable and of degree d, then the genus g of C is given by*

$$g = \begin{cases} (d-1)/2 & \text{for } d \text{ odd,} \\ (d-2)/2 & \text{for } d \text{ even.} \end{cases}$$

Assume that $r(x)$ is separable. A curve of the form $y^2 = r(x)$ of genus at least two is called a *hyperelliptic curve*, while when such a curve has genus one it is known as a *elliptic curve*. The reason we care about the genus of a curve is that Faltings' theorem famously connects it to the number of K-rational points on the curve:

Theorem 8 (Faltings; see [Hindry and Silverman 2000, Theorem E.0.1]). *Let K be a number field, and let C be a curve defined over K of genus $g \geq 2$. Then the set of K-rational points on C is finite.*

Suppose for a moment that all of the polynomials $t_i(x)$ in Theorem 6 are separable. Clearly $\deg t_i(x) = 2^{i-1}$. By Theorem 7, the genus g_i of the curve $y^2 = t_i(x)$ then satisfies

$$g_i = \begin{cases} 0 & \text{for } i = 1, \\ 2^{i-2} - 1 & \text{for } i \geq 2. \end{cases} \tag{3-3}$$

Therefore, by Faltings' theorem, the curve $y^2 = t_{n+1}(x)$ has only finitely many K-rational points for $n \geq 3$. In particular, there are only finitely many $x \in K$ such that (x, y) is a K-rational point on $y^2 = t_{n+1}(x)$. Thus, by Theorem 6, when $n \geq 3$ there are only finitely many $m \in K$ with $g^n_{\gamma,m}(x)$ irreducible and $g^{n+1}_{\gamma,m}(x)$ reducible over K.

Hence the lone remaining obstacle to proving part (1) of Theorem 1 is to establish that the $t_i(x)$ in Theorem 6 are separable. Note that this is not true for all $\gamma \in K$. Indeed, if $\gamma = \frac{1}{4}$, then $t_2(x) = \left(x + \frac{1}{2}\right)^2$. The set

$$S := \{\gamma \in \overline{\mathbb{Q}} : t_i(x) \text{ is separable for all } i \geq 1\}$$

is the same as the set of $a \in \overline{\mathbb{Q}}$ such that the preimage curves $Y^{\mathrm{pre}}(N, -a)_{N \geq 1}$ defined in [Faber et al. 2009] are all nonsingular. In general, the set $\overline{\mathbb{Q}} \setminus S$ is poorly understood. One result [Faber et al. 2009, Proposition 4.8] gives a criterion for membership in S. Here we give an improvement on that result.

Proposition 9. *Let K be a number field with ring of integers $\mathbb{O}_K$, and let $t_i(x)$ be as in Theorem 6. Suppose there exists a prime $\mathfrak{p}$ of $\mathbb{O}_K$ with $v_{\mathfrak{p}}(2) = e \geq 1$ and $v_{\mathfrak{p}}(\gamma) = s$ with $s \neq -e2^j$ for all $j \geq 1$. Then $t_i(x)$ is separable over K for all $i \geq 1$.*

Remark. When $K = \mathbb{Q}$, Proposition 9 says that if $v_2(\gamma) \neq -2^j$ for all $j \geq 1$, then $t_i(x)$ is separable for all $i \geq 1$.

Proof. It suffices to establish that $t_i(x)$ and $t_i'(x)$ have no common roots in $\overline{K}$, which we do through the use of Newton polygons with respect to the valuation $v_{\mathfrak{p}}$ (we abbreviate these by NP). We assume the reader is familiar with the relationship between slopes of the Newton polygon of a polynomial and the $\mathfrak{p}$-adic valuation of the polynomial's roots (see, e.g., [Silverman 2007, Theorem 5.11]). The proposition is obvious for $i = 1$, so we take $i \geq 2$. We first claim that for each r with $0 \leq r \leq i - 2$, $t_i'(x)$ has 2^r roots in $\overline{K}$ with $\mathfrak{p}$-adic valuation $-e/2^r$. The statement is trivial for $i = 2$, so we assume inductively that it holds for given $i \geq 3$, and we consider the NP of $t_i'(x)$ with respect to the $\mathfrak{p}$-adic valuation. By the chain rule,

$$t_{i+1}'(x) = 2(t_i(x) - \gamma)t_i'(x) + 1.$$

Observe that $t_i(x) - \gamma$ is monic, has integer coefficients, and has linear coefficient 1 (and constant term 0). Thus its NP consists of a single horizontal line segment from $(1, 0)$ to $(2^{i-1}, 0)$. From our inductive hypothesis, it follows that the NP of $2(t_i(x) - \gamma)t_i'(x)$ consists of a horizontal line segment from $(1, e)$ to $(2^{i-1}, e)$, followed by a sequence of segments of slope $e/2^{i-2}, e/2^{i-3}, \ldots, e$ and respective lengths $2^{i-2}, 2^{i-3}, \ldots, 1$. Hence the NP of $2(t_i(x) - \gamma)t_i'(x) + 1$ consists of a line segment from $(0, 0)$ to $(2^{i-1}, e)$, having slope $e/2^{i-1}$, and otherwise is identical to the NP of $2(t_i(x) - \gamma)t_i'(x)$, since $e/2^{i-1} < e/2^c$ for $0 \leq c \leq i - 2$. This proves the claim.

For each $i \geq 1$, $t_i(x)$ is a monic polynomial with degree 2^{i-1} and constant term γ, whose nonconstant coefficients are all integers. If $v_{\mathfrak{p}}(\gamma) \geq 0$, then the NP of $t_i(x)$ consists of nonpositive slopes, and hence all its roots have nonnegative $\mathfrak{p}$-adic valuation, and therefore cannot coincide with roots of $t_i'(x)$ by the above claim. If $v_{\mathfrak{p}}(\gamma) = s < 0$, the NP for $t_i(x)$ consists of a single line segment from $(0, s)$ to $(2^{i-1}, 0)$, with length 2^{i-1} and slope $-s/2^{i-1}$. Hence if $t_i(x)$ and $t_i'(x)$ have a root

in common, then by the above claim, $-s/2^{i-1} = e/2^r$ with $0 \le r \le i - 2$. But this holds if and only if $s = -e2^{i-1-r}$, and since $i - 1 - r \ge 1$, the proof is complete. $\square$

Remark. To show that the genus of the curve $y^2 = t_i(x)$ is at least two, we can get by with a much weaker statement than Proposition 9. Indeed, the genus of $y^2 = t_i(x)$ depends on the degree of $t_i(x)/f(x)$, where $f(x)$ is the square polynomial of largest degree dividing $t_i(x)$. It suffices to show that the degree of $t_i(x)/f(x)$ is at least five, for each $i \ge 4$.

4. The case $n = 2$

Consider now the case where $n = 2$. From (3-3), we know that when $t_3(x)$ is separable, $g_3 = 1$, and so $y^2 = t_3(x)$ is an elliptic curve. (When $t_3(x)$ is not separable, $y^2 = t_3(x)$ gives a curve of genus 0.) Thus we cannot directly apply Faltings' theorem, and we must use a different approach to determine the set $S(2, 2, \gamma)$ of $m \in K$ such that $g_{\gamma,m}^2(x)$ is irreducible and $g_{\gamma,m}^3(x)$ is reducible over K.

Now for some number fields K and some $\gamma \in K$, it may still be the case that $y^2 = t_3(x)$ has only finitely many K-rational points, proving the finiteness of $S(2, 2, \gamma)$ over K. This is the case for $\gamma = 0$ and $K = \mathbb{Q}$, as we now show:

Proposition 10. *Let $\gamma = 0$ and C_3 be the curve given by*

$$y^2 = t_3(x) = x^4 - 2x^3 + x^2 - x.$$

The only $\mathbb{Q}$-rational points on C_3 are $(0, 0)$ and the point at infinity. In particular, there are no $m \in \mathbb{Q}$ such that $x^2 + m$ has a newly reducible third iterate.

Proof. Let $y = u/v^2$ and $x = -1/v$ define a birational map ϕ from

$$C_3' : u^2 = v^3 + v^2 + 2v + 1$$

to C_3. We compute the conductor of the elliptic curve C_3' to be 92, and locate it as curve 92A1 in [Cremona]. From the same reference, we know that it has rank zero over $\mathbb{Q}$ and torsion subgroup of order 3. Hence the obvious points $(0, \pm 1)$ together with the point at infinity give all $\mathbb{Q}$-rational points on C_3'. If (x, y) is an affine rational point on C_3 with $x \ne 0$, then $\phi^{-1}(x, y)$ is an affine rational point (v, u) on C_3' with $v \ne 0$. But there are no such points. $\square$

The strategy of Proposition 10, however, won't even work for all number fields K in the case $\gamma = 0$. Indeed, let $K = \mathbb{Q}(i)$ and let ϕ be the same transformation as in Proposition 10. One can check that $(-1, i)$ is a nontorsion point of C_3' in many ways. One of the more interesting, if not the simplest computationally, is to show that $(-1, i)$ has positive canonical height. Silverman [1990] gives upper and lower bounds for the difference between the canonical height $\hat{h}(P)$ and the Weil height $h(P)$ of a K-rational point P on an elliptic curve, computed in terms

of the discriminant and j-invariant of the curve. For C_3', we have $-1.5484 \le \hat{h}(P) - h(P) \le 1.4577$. In particular, $\hat{h}(P) \ge h(P) - 1.5484$, so $h(P) > 1.5484$ would imply that P is a nontorsion point. Using MAGMA [Bosma et al. 1997], we find that although $h(P) = 0$ for $P = (-1, i)$ on C_3', we have $h([2]P) = 1.6094$. Thus $\hat{h}(P) = \frac{1}{4}\hat{h}([2]P) > 0$, using algebraic properties of canonical height.

Since $(-1, i)$ is a nontorsion point on C_3', the curve C_3 has infinitely many K-rational points. However, when we check some corresponding x-values on C_3 as our choices for m in $x^2 + m$, we don't find a newly reducible third iterate over $\mathbb{Q}(i)$. Thus we must adopt a different approach to have any hope of proving the case $n = 2$ of Theorem 1, even for $\gamma = 0$.

Let K be a number field and $\gamma \in K$. Suppose that $g_{\gamma,m}^3(x)$ is newly reducible, so that by Corollary 5, $g_{\gamma,m}^3(x) = p_1(x)p_2(x)$ for irreducible polynomials $p_1(x)$, $p_2(x) \in K[x]$ with $\deg p_1(x) = \deg p_2(x) = 4$. Put

$$p_1(x) = (x - \gamma)^4 + a_3(x - \gamma)^3 + a_2(x - \gamma)^2 + a_1(x - \gamma) + a_0,$$
$$p_2(x) = (x - \gamma)^4 + b_3(x - \gamma)^3 + b_2(x - \gamma)^2 + b_1(x - \gamma) + b_0$$

with $a_i, b_i \in K$. We also have

$$g_{\gamma,m}^3(x) = (x - \gamma)^8 + 4m(x - \gamma)^6 + (6m^2 + 2m)(x - \gamma)^4$$
$$+ (4m^3 + 4m^2)(x - \gamma)^2 + m^4 + 2m^3 + m^2 + m + \gamma.$$

Multiplying $p_1(x)$ and $p_2(x)$ together, setting this product equal to $g_{\gamma,m}^3(x)$ and comparing coefficients, we obtain a system of eight equations. By simplifying this system using Theorem 6, and noting that $a_0 \ne 0$ by the irreducibility of $p_1(x)$, we get two cases:

Case I: $a_1 \ne 0$, which implies $b_1 = -a_1$, $b_2 = a_2$:

(1) $2a_2 - a_3^2 - 4m = 0$;

(2) $2a_0 + a_2^2 - 2a_1a_3 - 6m^2 - 2m = 0$;

(3) $2a_2a_0 - a_1^2 - 4m^3 - 4m^2 = 0$;

(4) $a_0^2 - m^4 - 2m^2 - m^2 - m - \gamma = 0$.

Case II: $a_1 = b_1 = 0$:

(1) $b_2 - a_3^2 + a_2 - 4m = 0$;

(2) $(b_2 - a_2)a_3 = 0$;

(3) $2a_0 + a_2b_2 - 6m^2 - 2m = 0$;

(4) $(a_2 + b_2)a_0 - 4m^3 - 4m^2 = 0$;

(5) $a_0^2 - m^4 - 2m^2 - m^2 - m - \gamma = 0$.

We use Gröbner bases to find the solutions to these systems of nonlinear equations. We dispense with Case II first, noting that it consists of five equations in five variables so we expect it will have only finitely many solutions in $\overline{K}$. We assign an ordering to the variables in which γ is last, and using MAGMA [Bosma et al. 1997] to compute a Gröbner basis for each system, we find that the system in Case II has one K-rational solution for each $m \in K$ with

$$
\begin{aligned}
0 = {} & m^{14} + m^{13}\gamma + \tfrac{13}{3}m^{13} + \tfrac{13}{3}m^{12}\gamma + \tfrac{22}{3}m^{12} + \tfrac{22}{3}m^{11}\gamma + \tfrac{57}{8}m^{11} + \tfrac{33}{4}m^{10}\gamma \\
& + 5m^{10} + \tfrac{9}{8}m^9\gamma^2 + \tfrac{23}{3}m^9\gamma + \tfrac{9}{4}m^9 + \tfrac{8}{3}m^8\gamma^2 + \tfrac{25}{6}m^8\gamma + \tfrac{7}{12}m^8 + \tfrac{23}{12}m^7\gamma^2 \\
& + \tfrac{17}{12}m^7\gamma - \tfrac{1}{24}m^7 + \tfrac{13}{12}m^6\gamma^2 - \tfrac{1}{12}m^6\gamma - \tfrac{1}{12}m^6 + \tfrac{1}{4}m^5\gamma^3 - \tfrac{1}{24}m^5\gamma^2 \\
& - \tfrac{1}{4}m^5\gamma - \tfrac{1}{24}m^5 - \tfrac{1}{4}m^4\gamma^2 - \tfrac{1}{6}m^4\gamma - \tfrac{1}{12}m^3\gamma^3 - \tfrac{1}{4}m^3\gamma^2 - \tfrac{1}{6}m^2\gamma^3 - \tfrac{1}{24}m\gamma^4.
\end{aligned}
$$

Clearly for any $\gamma \in K$, there are at most 14 such m, and so case II does not affect the finiteness of the number of m for which $g_{\gamma,m}(x)$ has a newly irreducible third iterate.

Case I proves more interesting. We compute that for fixed $\gamma \in K$, Case I has precisely one solution $(a_0, a_1, a_2, a_3, m) \in K^5$ for each K-rational point (a_3, m) on the curve

$$
\begin{aligned}
C_\gamma : 0 = {} & a_3^{16} + 32a_3^{14}m + 352a_3^{12}m^2 - 32a_3^{12}m + 1792a_3^{10}m^3 - 256a_3^{10}m^2 \\
& + 4352a_3^8m^4 - 1536a_3^8m^3 - 1792a_3^8m^2 - 2176a_3^8m - 2176a_3^8\gamma \\
& + 4096a_3^6m^5 - 8192a_3^6m^4 - 12288a_3^6m^3 - 10240a_3^6m^2 - 10240a_3^6m\gamma \\
& - 16384a_3^4m^5 - 32768a_3^4m^4 - 38912a_3^4m^3 - 22528a_3^4m^2\gamma - 14336a_3^4m^2 \\
& - 14336a_3^4m\gamma - 16384a_3^2m^4 - 16384a_3^2m^3\gamma - 16384a_3^2m^3 - 16384a_3^2m^2\gamma \\
& + 4096m^2 + 8192m\gamma + 4096\gamma^2.
\end{aligned}
$$

For instance, when $\gamma = \tfrac{1}{2}$, one checks that C_γ has the rational point $(1, -\tfrac{7}{4})$, which corresponds to the newly reducible example given in (1-3). The actual Gröbner basis is far too long to include here; however, we have included the Gröbner basis in the case $\gamma = 1$ in the Appendix to this article. Thus when C_γ has genus at least two, there can be only finitely many K-rational solutions to the system given in Case I, and hence only finitely many $m \in K$ such that $g_{\gamma,m}(x)$ has a newly irreducible third iterate. Part (2) of Theorem 1 is thus proved when the genus C_γ is at least two.

Using MAGMA again, we checked that C_γ has genus 11 for $\gamma = r/4$, $-200 \le r \le 200$, except for the cases $g(C_{-2}) = 9$, $g(C_0) = 9$, $g(C_{1/4}) = 7$, $g(C_1) = 10$. Note that we chose γ to have denominator 4 in order to include the case $\gamma = \tfrac{1}{4}$, where we strongly suspected degeneracies to occur. The map ψ sending C_γ to γ has fibers whose genus appears generally to be 11. Even the degenerate fibers seem to have genus greater than 1, and hence part (2) of Theorem 1 holds even in those cases. Interestingly, if we take a section of ψ by fixing a value of m and letting γ

vary, we appear always to get a curve of genus at most 1. This phenomenon was first noticed by Michael Zieve (personal correspondence). In other words, writing $C_{\gamma,m}$ instead of C_γ, and choosing ψ' to be the map sending $C_{\gamma,m}$ to m, the surface $C_{\gamma,m}$ is (birational to) an elliptic surface. This observation may pave the way for a full understanding of $C_{\gamma,m}$, and hence improvements to part (2) of Theorem 1.

Acknowledgements

Appendix

We report the Gröbner basis for Case I from page 491 with $\gamma = 1$ as calculated by MAGMA [Bosma et al. 1997]:

(1) $a_0 - a_1 a_3 + \frac{1}{8} a_3^4 - a_3^2 q - q^2 + q$

(2) $a_1^2 - a_1 a_3^3 + 4 a_1 a_3 q + \frac{1}{8} a_3^6 - \frac{3}{2} a_3^4 q + 3 a_3^2 q^2 + a_3^2 q$

(3) $a_1 a_3^5 + \frac{1920}{571} a_1 a_3 q^6 - \frac{35582}{1713} a_1 a_3 q^5 + \frac{641146}{15417} a_1 a_3 q^4 - \frac{173966}{5139} a_1 a_3 q^3$
$\quad + \frac{254212}{15417} a_1 a_3 q^2 - \frac{4322}{571} a_1 a_3 q + \frac{35}{30834} a_3^{14} q - \frac{1}{1152} a_3^{14} - \frac{4265}{123336} a_3^{12} q^2$
$\quad + \frac{200467}{7893504} a_3^{12} q + \frac{4199}{2631168} a_3^{12} + \frac{1775}{5139} a_3^{10} q^3 - \frac{191455}{986688} a_3^{10} q^2 - \frac{75881}{986688} a_3^{10} q$
$\quad - \frac{22705}{15417} a_3^8 q^4 + \frac{516139}{986688} a_3^8 q^3 + \frac{315853}{493344} a_3^8 q^2 + \frac{54587}{986688} a_3^8 q - \frac{7}{48} a_3^8$
$\quad + \frac{36880}{15417} a_3^6 q^5 + \frac{76901}{61668} a_3^6 q^4 - \frac{148475}{30834} a_3^6 q^3 + \frac{219505}{61668} a_3^6 q^2 - \frac{11}{18} a_3^6 q$
$\quad - \frac{240}{571} a_3^4 q^6 - \frac{429961}{61668} a_3^4 q^5 + \frac{677423}{61668} a_3^4 q^4 - \frac{402371}{61668} a_3^4 q^3 - \frac{75667}{123336} a_3^4 q^2$
$\quad + \frac{131047}{41112} a_3^4 q + \frac{1920}{571} a_3^2 q^7 - \frac{35582}{1713} a_3^2 q^6 + \frac{641146}{15417} a_3^2 q^5 - \frac{374378}{15417} a_3^2 q^4$
$\quad + \frac{152233}{15417} a_3^2 q^3 - \frac{189763}{15417} a_3^2 q^2 + \frac{960}{571} q^5 - \frac{14911}{1713} q^4 + \frac{186374}{15417} q^3 - \frac{104975}{15417} q^2 + \frac{4}{3} q$

(4) $a_1 a_3^2 q + \frac{720}{571} a_1 q^6 - \frac{17791}{2284} a_1 q^5 + \frac{320573}{20556} a_1 q^4 - \frac{86983}{6852} a_1 q^3 + \frac{53275}{10278} a_1 q^2 - \frac{4199}{2284} a_1 q$
$\quad - \frac{45}{292352} a_3^{15} q^3 + \frac{14911}{18710528} a_3^{15} q^2 - \frac{93187}{84197376} a_3^{15} q + \frac{104975}{168394752} a_3^{15}$
$\quad + \frac{45}{9136} a_3^{13} q^4 - \frac{14911}{584704} a_3^{13} q^3 + \frac{93187}{2631168} a_3^{13} q^2 - \frac{11415}{584704} a_3^{13} q - \frac{1}{3072} a_3^{13}$
$\quad - \frac{495}{9136} a_3^{11} q^5 + \frac{161141}{584704} a_3^{11} q^4 - \frac{1915915}{5262336} a_3^{11} q^3 + \frac{300037}{1754112} a_3^{11} q^2 + \frac{206789}{7016448} a_3^{11} q$
$\quad + \frac{4199}{7016448} a_3^{11} + \frac{315}{1142} a_3^9 q^6 - \frac{101497}{73088} a_3^9 q^5 + \frac{1170419}{657792} a_3^9 q^4 - \frac{154417}{219264} a_3^9 q^3$
$\quad - \frac{203785}{877056} a_3^9 q^2 - \frac{75881}{2631168} a_3^9 q - \frac{765}{1142} a_3^7 q^7 + \frac{236207}{73088} a_3^7 q^6 - \frac{545431}{164448} a_3^7 q^5$
$\quad - \frac{142777}{109632} a_3^7 q^4 + \frac{4272259}{877056} a_3^7 q^3 - \frac{4322155}{1315584} a_3^7 q^2 + \frac{3623737}{2631168} a_3^7 q + \frac{360}{571} a_3^5 q^8$
$\quad - \frac{9151}{4568} a_3^5 q^7 - \frac{19973}{5139} a_3^5 q^6 + \frac{128675}{6852} a_3^5 q^5 - \frac{4341377}{164448} a_3^5 q^4 + \frac{1413245}{82224} a_3^5 q^3$
$\quad - \frac{830245}{164448} a_3^5 q^2 - \frac{41}{48} a_3^5 q - \frac{1440}{571} a_3^3 q^8 + \frac{20671}{1142} a_3^3 q^7 - \frac{258976}{5139} a_3^3 q^6$
$\quad + \frac{12676049}{164448} a_3^3 q^5 - \frac{3880925}{54816} a_3^3 q^4 + \frac{688435}{18272} a_3^3 q^3 - \frac{881653}{109632} a_3^3 q^2 + \frac{172159}{109632} a_3^3 q$
$\quad + \frac{2160}{571} a_3 q^7 - \frac{53373}{2284} a_3 q^6 + \frac{320573}{6852} a_3 q^5 - \frac{85543}{2284} a_3 q^4 + \frac{177007}{13704} a_3 q^3 - \frac{148651}{41112} a_3 q^2$

(5) $a_1q^7 - \frac{68}{9}a_1q^6 + \frac{1606}{81}a_1q^5 - \frac{578}{27}a_1q^4 + \frac{853}{81}a_1q^3 - \frac{50}{9}a_1q^2 + a_1q - \frac{1}{8192}a_3^{15}q^4$

$\quad + \frac{59}{73728}a_3^{15}q^3 - \frac{1075}{663552}a_3^{15}q^2 + \frac{377}{331776}a_3^{15}q - \frac{25}{73728}a_3^{15} + \frac{1}{256}a_3^{13}q^5$

$\quad - \frac{59}{2304}a_3^{13}q^4 + \frac{1075}{20736}a_3^{13}q^3 - \frac{83}{2304}a_3^{13}q^2 + \frac{35}{3456}a_3^{13}q - \frac{11}{256}a_3^{11}q^6$

$\quad + \frac{5}{18}a_3^{11}q^5 - \frac{5647}{10368}a_3^{11}q^4 + \frac{9341}{27648}a_3^{11}q^3 - \frac{847}{13824}a_3^{11}q^2 - \frac{275}{27648}a_3^{11}q$

$\quad - \frac{1}{3072}a_3^{11} + \frac{7}{32}a_3^9q^7 - \frac{101}{72}a_3^9q^6 + \frac{3497}{1296}a_3^9q^5 - \frac{5249}{3456}a_3^9q^4 + \frac{203}{1728}a_3^9q^3$

$\quad + \frac{365}{10368}a_3^9q^2 + \frac{11}{1152}a_3^9q - \frac{17}{32}a_3^7q^8 + \frac{949}{288}a_3^7q^7 - \frac{7261}{1296}a_3^7q^6 + \frac{1103}{3456}a_3^7q^5$

$\quad + \frac{19607}{3456}a_3^7q^4 - \frac{58525}{10368}a_3^7q^3 + \frac{15673}{5184}a_3^7q^2 - \frac{863}{1152}a_3^7q + \frac{1}{2}a_3^5q^9 - \frac{41}{18}a_3^5q^8$

$\quad - \frac{115}{81}a_3^5q^7 + \frac{4409}{216}a_3^5q^6 - \frac{23737}{648}a_3^5q^5 + \frac{19853}{648}a_3^5q^4 - \frac{2225}{162}a_3^5q^3 + \frac{427}{216}a_3^5q^2$

$\quad - 2a_3^3q^9 + \frac{154}{9}a_3^3q^8 - \frac{37351}{648}a_3^3q^7 + \frac{8318}{81}a_3^3q^6 - \frac{11993}{108}a_3^3q^5 + \frac{3571}{48}a_3^3q^4$

$\quad - \frac{776}{27}a_3^3q^3 + \frac{3539}{432}a_3^3q^2 - \frac{41}{48}a_3^3q + 3a_3q^8 - \frac{68}{3}a_3q^7 + \frac{1606}{27}a_3q^6 - \frac{1147}{18}a_3q^5$

$\quad + \frac{778}{27}a_3q^4 - \frac{2075}{162}a_3q^3 + \frac{49}{18}a_3q^2$

(6) $a_2 - \frac{1}{2}a_3^2 + 2q$

(7) $a_3^{16} - 32a_3^{14}q + 352a_3^{12}q^2 + 32a_3^{12}q - 1792a_3^{10}q^3 - 256a_3^{10}q^2$

$\quad + 4352a_3^8q^4 + 1536a_3^8q^3 - 1792a_3^8q^2 + 2176a_3^8q - 4096a_3^6q^5$

$\quad - 8192a_3^6q^4 + 12288a_3^6q^3 - 10240a_3^6q^2 + 16384a_3^4q^5 - 32768a_3^4q^4$

$\quad + 38912a_3^4q^3 - 14336a_3^4q^2 - 16384a_3^2q^4 + 16384a_3^2q^3 + 4096q^2$

References

[Beardon 1991] A. F. Beardon, *Iteration of rational functions: Complex analytic dynamical systems*, Graduate Texts in Mathematics **132**, Springer, New York, 1991. MR 92j:30026 Zbl 0742.30002

[Bosma et al. 1997] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system, I: The user language", *J. Symbolic Comput.* **24**:3–4 (1997), 235–265. MR 1484478 Zbl 0898.68039

[Boston and Jones 2007] N. Boston and R. Jones, "Arboreal Galois representations", *Geom. Dedicata* **124** (2007), 27–35. MR 2009e:11103 Zbl 1206.11069

[Cremona] J. E. Cremona, "Elliptic curve data", online tables, University of Warwick, Available at http://homepages.warwick.ac.uk/staff/J.E.Cremona//ftp/data/INDEX.html.

[Danielson and Fein 2002] L. Danielson and B. Fein, "On the irreducibility of the iterates of $x^n - b$", *Proc. Amer. Math. Soc.* **130**:6 (2002), 1589–1596. MR 2002m:12001 Zbl 1007.12001

[Faber et al. 2009] X. Faber, B. Hutz, P. Ingram, R. Jones, M. Manes, T. J. Tucker, and M. E. Zieve, "Uniform bounds on pre-images under quadratic dynamical systems", *Math. Res. Lett.* **16**:1 (2009), 87–101. MR 2009m:11095 Zbl 1222.11086

[Fein and Schacher 1996] B. Fein and M. Schacher, "Properties of iterates and composites of polynomials", *J. London Math. Soc.* (2) **54**:3 (1996), 489–497. MR 97h:12007 Zbl 0865.12003

[Goldschmidt 2003] D. M. Goldschmidt, *Algebraic functions and projective curves*, Graduate Texts in Mathematics **215**, Springer, New York, 2003. MR 2003j:14001 Zbl 1034.14011

[Hamblen et al. 2013] S. Hamblen, R. Jones, and K. Madhu, "The density of primes in orbits of $z^d + c$", preprint, 2013. arXiv 1303.6513

[Hindry and Silverman 2000] M. Hindry and J. H. Silverman, *Diophantine geometry: An introduction*, Graduate Texts in Mathematics **201**, Springer, New York, 2000. MR 2001e:11058 Zbl 0948.11023

[Jones 2008] R. Jones, "The density of prime divisors in the arithmetic dynamics of quadratic polynomials", *J. Lond. Math. Soc.* (2) **78**:2 (2008), 523–544. MR 2010b:37239 Zbl 1193.37144

[Jones 2012] R. Jones, "An iterative construction of irreducible polynomials reducible modulo every prime", *J. Algebra* **369** (2012), 114–128. MR 2959789

[Jones and Boston 2012] R. Jones and N. Boston, "Settled polynomials over finite fields", *Proc. Amer. Math. Soc.* **140**:6 (2012), 1849–1863. MR 2012m:37142 Zbl 1243.11115

[Jones and Manes 2011] R. Jones and M. Manes, "Galois theory of quadratic rational functions", preprint, 2011. To appear in *Comment. Math. Helv.* arXiv 1101.4339

[Odoni 1985] R. W. K. Odoni, "On the prime divisors of the sequence $w_{n+1} = 1 + w_1 \cdots w_n$", *J. London Math. Soc.* (2) **32**:1 (1985), 1–11. MR 87b:11094 Zbl 0574.10020

[Silverman 1990] J. H. Silverman, "The difference between the Weil height and the canonical height on elliptic curves", *Math. Comp.* **55**:192 (1990), 723–743. MR 91d:11063 Zbl 0729.14026

[Silverman 2007] J. H. Silverman, *The arithmetic of dynamical systems*, Graduate Texts in Mathematics **241**, Springer, New York, 2007. MR 2008c:11002 Zbl 1130.37001

[Stoll 1992] M. Stoll, "Galois groups over $\mathbb{Q}$ of some iterated polynomials", *Arch. Math.* (*Basel*) **59**:3 (1992), 239–244. MR 93h:12004 Zbl 0758.11045

kacham12@g.holycross.edu   *Department of Mathematics and Computer Science, College of the Holy Cross, One College Street, Worcester, MA 10610, United States*

ercolb13@g.holycross.edu   *Department of Mathematics and Computer Science, College of the Holy Cross, One College Street, Worcester, MA 01610, United States*

sfrechet@holycross.edu   *Department of Mathematics and Computer Science, College of the Holy Cross, One College Street, Worcester, MA 01610, United States*

peheff13@g.holycross.edu   *Department of Mathematics and Computer Science, College of the Holy Cross, One College Street, Worcester, MA 01610, United States*

rfjones@carleton.edu   *Department of Mathematics, Carleton College, One North College Street, Northfield, MN 55057, United States*

seorch13@g.holycross.edu   *Department of Mathematics and Computer Science, College of the Holy Cross, One College Street, Worcester, MA 01610, United States*

Positive symmetric solutions of a second-order difference equation

Jeffrey T. Neugebauer and Charley L. Seelbach

(Communicated by Johnny Henderson)

Using an extension of the Leggett–Williams fixed-point theorem due to Avery, Henderson, and Anderson, we prove the existence of solutions for a class of second-order difference equations with Dirichlet boundary conditions, and discuss a specific example.

1. Introduction

Many fixed-point theorems have applications in proving the existence of positive solutions of boundary value problems. One class of such theorems, originating with [Krasnoselskii 1964], involves an operator defined on a "wedge" — a portion of a Banach space bounded by level surfaces of positive functionals — and satisfying certain criteria. In Krasnoselskii's original theorem, the functional was the norm; that is, the wedge conditions where $a \leq \|x\|$ and $\|x\| \leq b$, for $0 < a < b$. A later variant, in [Leggett and Williams 1979], replaced the lower wedge condition by $a \leq \alpha(x)$, where α is a concave positive functional with $\alpha(x) \leq \|x\|$; this allows more flexibility in the choice of the wedge in applications. The Leggett–Williams theorem was extended by Avery, Henderson, and Anderson [Avery et al. 2009] to allow flexibility also in the upper wedge condition, which gets replaced by $\beta(x) \leq b$, where β is a convex positive functional. This is the result of primary interest to this paper; other related results can be found in [Guo 1984; Avery and Henderson 2001; Anderson et al. 2010; Mavridis 2010].

Applications of such fixed-point theorems have been seen in works dealing with ordinary differential equations [Avery et al. 2000; 2010; Erbe and Wang 1994] and dynamic equations on time scales [Erbe et al. 2005; Liu et al. 2012; Prasad and Sreedhar 2011]. Most relevant to this paper, these theorems have been utilized for results that involve finite difference equations [Anderson et al. 2011; Cai and Yu 2006; Henderson et al. 2010].

Here we give an application of the fixed-point theorem of [Avery et al. 2009], stated below as Theorem 2.1, to obtain at least one positive solution of the difference equation

$$\Delta^2 u(k) + f(u(k)), \quad k \in \{0, \ldots, N-2\}, \tag{1-1}$$

with boundary conditions

$$u(0) = u(N) = 0. \tag{1-2}$$

Here $f : [0, \infty) \to [0, \infty)$ is any continuous function and Δ^2 is the second-difference operator, defined by $(\Delta^2 u)(k) = u(k) - 2u(k+1) + u(k+2)$. In fact we will obtain a *symmetric* solution, in the sense that $u(k) = u(N-k)$ for each k.

In Section 2 we present the fixed-point theorem of Avery et al. Section 3 contains preliminaries needed for our result on the difference equation (1-1), (1-2). That result is stated and proved in Section 4, and applied to a particular case in Section 5.

2. Statement of the fixed-point theorem

Let E be a real Banach space. A nonempty closed convex set $\mathcal{P} \subset E$ is called a *cone* if it contains the origin, is closed under multiplication by positive scalars, and has no overlap with its negative (apart from the origin). In symbols,

$$u \in \mathcal{P}, \, \lambda \geq 0 \implies \lambda u \in \mathcal{P} \quad \text{and} \quad u \in \mathcal{P}, \, -u \in \mathcal{P} \implies u = 0.$$

Let $\mathcal{P}$ be a cone in E. A map $\alpha : \mathcal{P} \to [0, \infty)$ is said to be a *nonnegative continuous concave functional* on $\mathcal{P}$ if it is continuous and satisfies

$$\alpha(tu + (1-t)v) \geq t\alpha(u) + (1-t)\alpha(v),$$

for all $u, v \in \mathcal{P}$ and $t \in [0, 1]$. Replacing $\geq$ by $\leq$ we obtain the definition of a *nonnegative continuous convex functional* on $\mathcal{P}$.

In the statement of the theorem, there appear two concave functionals, α and ϕ, and two convex ones, β and γ. The functionals α and β delimit the wedge where the operator is defined; the other two ensure additional flexibility in applications, in comparison with the Leggett–Williams theorem.

Theorem 2.1 [Avery et al. 2009]. *Let $\mathcal{P}$ be a cone in a real Banach space E. Suppose that α and ψ are nonnegative continuous concave functionals on $\mathcal{P}$ and that β and δ are nonnegative continuous convex functionals on $\mathcal{P}$. For nonnegative real numbers $a, b, c,$ and d, define*

$$A := A(\alpha, \beta, a, d) = \{x \in P : a \leq \alpha(x) \text{ and } \beta(x) \leq d\}, \tag{2-1}$$

and suppose that A is a bounded subset of P. Let $T : A \to \mathcal{P}$ be a completely continuous operator (that is, it is continuous and maps bounded sets into precompact sets). Then T has a fixed point in A provided that the following conditions hold:

A0. $\{x \in P : \alpha(x) < a \text{ and } d < \beta(x)\} = \emptyset.$

A1. $\{x \in A : c < \psi(x) \text{ and } \delta(x) < b\} \neq \emptyset.$

A2. $\alpha(Tx) \geq a$ for all $x \in A$ with $\delta(x) \leq b$.

A3. $\alpha(Tx) \geq a$ for all $x \in A$ with $b < \delta(Tx)$.

A4. $\beta(Tx) \leq d$ for all $x \in A$ with $c \leq \psi(x)$.

A5. $\beta(Tx) \leq d$ for all $x \in A$ with $\psi(Tx) < c$.

3. Application of the theorem to a difference equation

In this section we return to the system (1-1), (1-2), stating in Theorem 3.2 sufficient conditions for the existence of a solution. This result is proved in the next section using Theorem 2.1. First, however, we set up some of the objects that appear in the statement of Theorem 2.1. Throughout the discussion we use the abbreviations

$$\underline{N} = \left\lfloor \frac{N}{2} \right\rfloor \quad \text{and} \quad \overline{N} = \left\lceil \frac{N}{2} \right\rceil.$$

Define the Banach space E to be the space of functions $u : \{0, \ldots, N\} \to \mathbb{R}$ with the norm

$$\|u\| = \max_{k \in \{0,1,\ldots,N\}} |u(k)|.$$

Within E, consider the cone $\mathcal{P}$ consisting of all u that are nonnegative, symmetric, nondecreasing on $\{0, 1, \ldots, \underline{N}\}$, and satisfy $wu(y) \geq yu(w)$ for $w \geq y$, where $y, w \in \{0, 1, \ldots, \underline{N}\}$.

Set

$$H(k, l) = \frac{1}{N} \begin{cases} k(N-l), & k \in \{0, \ldots, l\}, \\ l(N-k), & k \in \{l+1, \ldots, N\}. \end{cases}$$

(This is the Green's function for $-\Delta^2$ satisfying the boundary conditions (1-2).) Define the operator T by

$$(Tu)(k) := \sum_{l=1}^{N-1} H(k, l) f(u(l)).$$

By direct checking one sees that the condition $Tu = u$ is equivalent to (1-1) and (1-2). Thus any fixed point of T is a solution of our problem.

Lemma 3.1. *The operator T maps A into $\mathcal{P}$.*

Proof. Let $u \in A$. We first need to show that $Tu(N-k) = Tu(k)$. Notice that $H(N-k, N-l) = H(k, l)$. Now

$$Tu(N-k) = \sum_{l=1}^{N-1} H(N-k, l) f(u(l)).$$

Applying the substitution $r = N - l$, we can write

$$
Tu(N-k) = \sum_{r=1}^{N-1} H(N-k, N-r) f(u(N-r))
$$

$$
= \sum_{r=1}^{N-1} H(k, r) f(u(r)) = Tu(k).
$$

Next we need to show $Tu(k)$ is nonnegative and nondecreasing on $\{0, 1, \ldots, \underline{N}\}$. Since $H(k, l) \geq 0$ for $k, l \in \{0, \ldots, N\}$ and f only takes nonnegative values, $Tu(k)$ is nonnegative for all $k \in \{0, \ldots, N\}$.

To prove that $Tu(k)$ is nondecreasing on $\{0, 1, \ldots, \underline{N}\}$, we show that $\Delta Tu(k) := Tu(k-1) - Tu(k)$ is nonnegative on $\{0, 1, \ldots, \underline{N}\}$. Now

$$
H(k+1, l) - H(k, l) = \frac{1}{N} \times \begin{cases} N-l & \text{if } k \in \{0, \ldots, l\}, \\ -l & \text{if } k \in \{l, \ldots, N-1\}. \end{cases}
$$

So

$$
\Delta Tu(k) = \sum_{l=1}^{N-1} \bigl(H(k+1, l) - H(k, l) \bigr) f(u(l))
$$

$$
= \sum_{l=1}^{k-1} \frac{-l}{N} f(u(l)) + \sum_{l=k}^{N-1} \frac{N-l}{N} f(u(l))
$$

$$
= \sum_{l=1}^{k-1} \frac{-l}{N} f(u(l)) + \sum_{l=k}^{N-1} \frac{N-l}{N} f(u(N-l))
$$

$$
= \sum_{l=1}^{k-1} \frac{-l}{N} f(u(l)) + \sum_{r=1}^{N-k} \frac{r}{N} f(u(r))
$$

$$
= \sum_{l=1}^{k-1} \frac{-l}{N} f(u(l)) + \sum_{l=1}^{N-k} \frac{l}{N} f(u(l)).
$$

Since $k \in \{0, 1, \ldots, \underline{N}\}$,

$$
\Delta Tu(k) = \sum_{l=1}^{k-1} \frac{-l}{N} f(u(l)) + \sum_{l=1}^{N-k} \frac{l}{N} f(u(l)) = \sum_{l=k}^{N-k} \frac{l}{N} f(u(l)) \geq 0,
$$

as needed.

Lastly, we have $wTu(y) \geq yTu(w)$, since $H(k, l)$ satisfies $\dfrac{H(y, l)}{H(w, l)} \geq \dfrac{y}{w}$ for all l and all $w \geq y$. Thus T maps A into $\mathscr{P}$. $\qquad \square$

Theorem 3.2. *Assume that $\tau, \mu, \nu \in \{1, \ldots, \underline{N}\}$ are fixed with $\tau \leq \mu < \nu$, that d and m are positive real numbers with $0 < m < d\mu/\underline{N}$, and that $f : [0, \infty) \to [0, \infty)$ is a continuous function such that*

(i) $f(w) \geq \dfrac{2Nd}{(\nu - \tau)(2N - 1 - \tau - \nu)\underline{N}}$ *for $w \in [\tau d/\underline{N}, \nu d/\underline{N}]$,*

(ii) *$f(w)$ is decreasing for $w \in [0, m]$ and $f(m) \geq f(w)$ for $w \in [m, d]$, and*

(iii) $2 \displaystyle\sum_{l=1}^{\mu} \dfrac{l\overline{N}}{N} f\left(\dfrac{ml}{\mu}\right) \leq d - f(m)\dfrac{1}{N}(\overline{N})(\underline{N} - \mu)(\mu + 1 + \underline{N})$.

Set $a = \tau d/\underline{N}$. Then (1-1), (1-2) has at least one positive symmetric solution $u^* \in A$, where A is given by (2-1).

4. Proof of Theorem 3.2

Let $a = \tau d/\underline{N}$, $b = \nu d/\underline{N}$, $c = \mu d/\underline{N}$. By Lemma 3.1, T maps A into $\mathcal{P}$. Let $u \in A$. Then $\beta(u) = u(\underline{N}) \leq d$. But u achieves its maximum at $\underline{N}$, so A is bounded. By the Arzelà–Ascoli theorem, T is a completely continuous operator.

Now define the functionals appearing in the theorem as follows, where $u \in \mathcal{P}$:

$$\alpha(u) = \min_{k \in \{\tau, \ldots, \underline{N}\}} u(k) = u(\tau), \quad \psi(u) = \min_{k \in \{\mu, \ldots, \underline{N}\}} u(k) = u(\mu),$$

$$\delta(u) = \max_{k \in \{0, \ldots, \nu\}} u(k) = u(\nu), \quad \beta(u) = \max_{k \in \{0, \ldots, \underline{N}\}} u(k) = u(\underline{N}).$$

It is easy to check that α and ψ are concave and β and δ are convex.

We check conditions A0–A5 in turn. Let $u \in P$ and let $\beta(u) > d$. Then

$$\alpha(u) = u(\tau) \geq \frac{\tau}{N} u(\underline{N}) = \frac{\tau}{N} \beta(u) > \frac{\tau d}{\underline{N}} = a.$$

So $\{u \in P : \alpha(u) < a \text{ and } d < \beta(u)\} = \varnothing$, which is A0.

Now let $K \in \left(\dfrac{2d}{\underline{N}(3N - 4 - \mu)}, \dfrac{2d}{\underline{N}(3N - 4 - \nu)}\right)$. Define

$$u_K(k) = K \sum_{l=1}^{N-1} H(k, l) = \frac{Kk}{2}(3N - 4 - k).$$

Then

$$\alpha(u_k) = u_k(\tau) = \frac{K\tau}{2}(3N - 4 - \tau) > \frac{2d\tau(3N - 4 - \tau)}{2\underline{N}(3N - 4 - \mu)} \geq \frac{\tau d}{\underline{N}} = a$$

and

$$\beta(u_k) = u_k(\underline{N}) = \frac{K\underline{N}}{2}(3N - 4 - \underline{N}) < \frac{2\underline{N}d(3N - 4 - \underline{N})}{2\underline{N}(3N - 4 - \nu)} \leq \frac{Nd}{\underline{N}} = d.$$

So $u_k \in A$.

Since

$$\psi(u_k) = u_k(\mu) = \frac{K\mu}{2}(3N-4-\mu) > \frac{2d\mu(3N-4-\mu)}{2\underline{N}(3N-4-\mu)} = \frac{\mu d}{\underline{N}} = c$$

and

$$\delta(u_k) = u_k(\nu) = \frac{K\nu}{2}(3N-4-\nu) < \frac{2d\nu(3N-4-\nu)}{2\underline{N}(3N-4-\nu)} = \frac{\nu d}{\underline{N}} = b,$$

we have $\{u \in A : c < \psi(u) \text{ and } \delta(u) < b\} \neq \varnothing$. Therefore A1 holds.

To show that A2 holds, take $u \in A$ with $\delta(u) < b$. By (i),

$$\alpha(Tu) = \sum_{l=1}^{N-1} H(\tau, l) f(u(l)) \geq \sum_{l=\tau+1}^{\nu} H(\tau, l) f(u(l))$$

$$\geq \frac{2Nd}{(\nu-\tau)(2N-1-\tau-\nu)\underline{N}} \cdot \frac{\tau(\nu-\tau)(2N-1-\tau-\nu)}{2N} \geq \frac{\tau d}{\underline{N}} = a.$$

To show that A3 holds, let $u \in A$ with $\delta(Tu) > b$. Then

$$\alpha(Tu) = Tu(\tau) = \sum_{l=1}^{N-1} H(\tau, l) f(u(l)) \geq \frac{\tau}{\nu} \sum_{l=1}^{N-1} H(\nu, l) f(u(l))$$

$$= \frac{\tau}{\nu}\delta(Tu) > \frac{\tau}{\nu} b = \frac{d\tau}{\underline{N}} = a.$$

Now we show that A4 holds. Let $u \in A$ satisfy $c \leq \phi(x)$. By the concavity of u and since $c = \mu d/\underline{N}$, for all $k \in \{0, 1, \ldots, \mu\}$, we have

$$u(k) \geq \frac{ck}{\mu} \geq \frac{mk}{\mu}.$$

So, by (ii) and (iii), we have

$$\beta(Tu) = \sum_{l=1}^{N-1} H(\underline{N}, l) f(u(l)) \leq 2 \sum_{l=1}^{\overline{N}} \frac{l(N-\underline{N})}{N} f(u(l))$$

$$= 2 \sum_{l=1}^{\mu} \frac{l(\overline{N})}{N} f(u(l)) + 2 \sum_{l=\mu+1}^{\overline{N}} \frac{l(\overline{N})}{N} f(u(l))$$

$$\leq 2 \sum_{l=1}^{\mu} \frac{l(\overline{N})}{N} f(u(ml/\mu)) + 2 \sum_{l=\mu+1}^{\overline{N}} \frac{l(\overline{N})}{N} f(m)$$

$$\leq d - f(m)\frac{\overline{N}}{N}(\underline{N}-\mu)(\mu+1+\underline{N}) + f(m)\frac{\overline{N}}{N}(\underline{N}-\mu)(\mu+1+\underline{N}) = d.$$

Thus A4 is satisfied.

Last, we show that A5 is satisfied. Let $u \in A$ with $\psi(Tu) < c$. Then

$$\beta(Tu) = \sum_{l=1}^{N-1} H(\underline{N}, l) f(u(l)) \leq \frac{N}{\mu} \sum_{l=1}^{N-1} H(\mu, l) f(u(l)) \leq \frac{N}{\mu} \psi(Tu) < \frac{cN}{\mu} = d.$$

Therefore T has a fixed point and (1-1), (1-2) has at least one positive symmetric solution $u^* \in A$.

5. Example

Example 1. Let $N = 20$, $\tau = 1$, $\mu = 9$, $\nu = 10$, $d = 5$, and $m = 4.4$. Notice that $0 < \tau \leq \mu < \nu \leq 10 = \underline{N}$, and $0 < m = 4.4 \leq 4.5 = d\mu/\underline{N}$. Define a continuous function $f : [0, \infty) \to [0, \infty)$ by

$$f(w) = \begin{cases} \frac{1}{500}(45 - w) & \text{if } 0 \leq w \leq 40, \\ \frac{1}{100} & \text{if } w \geq 40. \end{cases}$$

Then,

(i) for $w \in [\frac{1}{2}, 5]$, $f(w) \geq f(5) = \dfrac{2}{25} > \dfrac{5}{63} = \dfrac{2 \cdot 20 \cdot 5}{(10-1) \cdot (3 + 2 \cdot 18 - 1 - 10)(10)}$,

(ii) $f(w)$ is decreasing for $w \in [0, 4.4]$ and $f(m) \geq f(w)$ for $w \in [4.4, 5]$, and

(iii) $2 \sum_{l=1}^{9} \dfrac{10l}{20} f\left(\dfrac{4.4l}{9}\right) = \dfrac{5657}{1500} < \dfrac{1047}{250} = 5 - f(4.4)\left(\dfrac{1}{20}\right)(10)(10-9)(9+1+10).$

So the hypotheses of Theorem 3.2 are satisfied. Therefore, the difference equation

$$\Delta^2 u(k) + f(u(k)), \quad k \in \{0, 1, \dots, 18\},$$

with boundary conditions

$$u(0) = u(20) = 0,$$

has a positive symmetric solution u^* with $u(1) \geq \frac{1}{2}$ and $u(10) \leq 5$.

References

[Anderson et al. 2010] D. R. Anderson, R. I. Avery, and J. Henderson, "Functional expansion: compression fixed point theorem of Leggett–Williams type", *Electron. J. Differ. Equations* **2010** (2010), Article ID #63. MR 2011f:47091 Zbl 1226.47054

[Anderson et al. 2011] D. R. Anderson, R. I. Avery, J. Henderson, X. Liu, and J. W. Lyons, "Existence of a positive solution for a right focal discrete boundary value problem", *J. Differ. Equ. Appl.* **17**:11 (2011), 1635–1642. MR 2012h:39003 Zbl 1234.39002

[Avery and Henderson 2001] R. I. Avery and J. Henderson, "Two positive fixed points of nonlinear operators on ordered Banach spaces", *Comm. Appl. Nonlinear Anal.* **8**:1 (2001), 27–36. MR 2001k:47079 Zbl 1014.47025

[Avery et al. 2000] R. I. Avery, J. M. Davis, and J. Henderson, "Three symmetric positive solutions for Lidstone problems by a generalization of the Leggett–Williams theorem", *Electron. J. Differ. Equations* **2000** (2000), Article ID #40. MR 2001c:34048 Zbl 0958.34020

[Avery et al. 2009] R. I. Avery, J. Henderson, and D. R. Anderson, "A topological proof and extension of the Leggett–Williams fixed point theorem", *Comm. Appl. Nonlinear Anal.* **16**:4 (2009), 39–44. MR 2011a:47115 Zbl 1184.47024

[Avery et al. 2010] R. I. Avery, J. Henderson, and D. R. Anderson, "Existence of a positive solution to a right focal boundary value problem", *Electron. J. Qual. Theory Differ. Equ.* **2010** (2010), Article ID #5. MR 2011a:34051 Zbl 1202.34054

[Cai and Yu 2006] X. Cai and J. Yu, "Existence theorems for second-order discrete boundary value problems", *J. Math. Anal. Appl.* **320**:2 (2006), 649–661. MR 2007b:39013 Zbl 1113.39019

[Erbe and Wang 1994] L. H. Erbe and H. Wang, "On the existence of positive solutions of ordinary differential equations", *Proc. Amer. Math. Soc.* **120**:3 (1994), 743–748. MR 94e:34025 Zbl 0802.34018

[Erbe et al. 2005] L. Erbe, A. Peterson, and C. Tisdell, "Existence of solutions to second-order BVPs on time scales", *Appl. Anal.* **84**:10 (2005), 1069–1078. MR 2007e:39002 Zbl 1088.39014

[Guo 1984] D. J. Guo, "Some fixed point theorems on cone maps", *Kexue Tongbao* (*English Ed.*) **29**:5 (1984), 575–578. MR 87f:47088 Zbl 0553.47022

[Henderson et al. 2010] J. Henderson, X. Liu, J. W. Lyons, and J. T. Neugebauer, "Right focal boundary value problems for difference equations", *Opuscula Math.* **30**:4 (2010), 447–456. MR 2012a:39003 Zbl 1227.39004

[Krasnoselskii 1964] M. A. Krasnoselskii, *Positive solutions of operator equations*, P. Noordhoff, Groningen, 1964. MR 31 #6107 Zbl 0121.10603

[Leggett and Williams 1979] R. W. Leggett and L. R. Williams, "Multiple positive fixed points of nonlinear operators on ordered Banach spaces", *Indiana Univ. Math. J.* **28**:4 (1979), 673–688. MR 80i:47073 Zbl 0421.47033

[Liu et al. 2012] X. Liu, J. T. Neugebauer, and S. Sutherland, "Application of a functional type compression expansion fixed point theorem for a right focal boundary value problem on a time scale", *Comm. Appl. Nonlinear Anal.* **19**:2 (2012), 25–39. MR 2953282 Zbl 06077348

[Mavridis 2010] K. G. Mavridis, "Two modifications of the Leggett–Williams fixed point theorem and their applications", *Electron. J. Differential Equations* (2010), No. 53, 11. MR 2011a:47130 Zbl 1226.47060

[Prasad and Sreedhar 2011] K. R. Prasad and N. Sreedhar, "Even number of positive solutions for 3nth order three-point boundary value problems on time scales", *Electron. J. Qual. Theory Differ. Equ.* **2011** (2011), Article ID #98. MR 2012m:34168

jeffrey.neugebauer@eku.edu   *Department of Mathematics and Statistics,*
Eastern Kentucky University, 521 Lancaster Avenue,
313 Wallace Building, Richmond, KY 40475, United States

charley_seelbach@mymail.eku.edu
Department of Mathematics and Statistics,
Eastern Kentucky University, 521 Lancaster Avenue,
313 Wallace Building, Richmond, KY 40475, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2012   vol. 5   no. 4