The Effect of Links and Excerpts on Internet News Consumption: Online Appendices

Online Appendix A: Simulation Procedure

Here we document the simulation procedure reported in model section of the manuscript and provide further detailed results.

Procedure

We simulate browsing for two types of consumers—1) myopic (with $\delta = 0$), and 2) forward-looking (with $\delta = 1$)—under three types of excerpting—1) no links, 2) links are noisy signals ($\tau_s = .2$), and 3) links are informative signals ($\tau_s = 2$). We simulate 30,000 browsing sessions under each of the six conditions. The consumer's cost is set to $\gamma = 2$, match preference to $v = 2$, and both sites are located at $z = 0$ (hence they provide the same average horizontal match utility). Site L always links to site R, but the reverse is not true: $\omega_{\vec{LR}} = 1$ and $\omega_{\vec{RL}} = 0$.

Results

Initiating a browsing session. Figure OA1 shows the probability of initiating a browsing session (visiting at least one site on a given day) under the six conditions. Forward-looking consumers are increasingly likely to initiate browsing sessions as links become more informative. When links are especially informative, the anticipated future benefits are even higher because consumers can choose to visit the linked site only when it provides very high match. Myopic consumers, on the other hand, are insensitive to the precision of link signals, since they cannot anticipate the future benefits from seeing excerpts.

Share of sessions starting at the linking site. Because the two sites offer identical match utility in expectation, myopic consumers are equally likely to start their sessions at both sites, as seen in Figure OA2. Forward-looking consumers behave the same when there are no links, but as links become more informative, they are increasingly likely to start their sessions at the linking site (L) given the anticipated future benefits from seeing excerpts from site R.

Number of sites visited per session. Figure OA3 shows that as links convey more information, both myopic and forward-looking consumers visit more sites (conditional on having initiated a session—i.e., the denominator in this average is the number of sessions in each condition). The increase in session length is due to the consumer being more likely to visit site R after seeing an excerpt at site L. The even greater increase among forward-looking consumers is due to their greater likelihood of initiating their session at site L when links are informative.

Share of sessions visiting the linked site. Figure OA4 shows that when links are informative, total traffic at the linked site is higher. The increase in traffic going to site R is highest if consumers are myopic, however, because forward-looking consumers *delay* their visits to the linked site, and sometimes choose to end their session before visiting R.

Effect of signal valence. Figure OA5 shows the asymmetric effect of match signals on visit probabilities by considering only sessions that begin at site L. When the excerpt at site L signals higher than average match, then the probability of subsequently visiting site R increases. (The amount of the increase is the same for forward-looking and myopic consumers.) Similarly, when the excerpt at L signals lower than average match, then the probability of subsequently visiting site R decreases. The magnitude of the decrease, however, is smaller than the magnitude of the increase because the probability of subsequently visiting R is already low to start with. That is, there is a floor effect limiting the damage that low match signals can inflict on the excerpted site.

Figure OA1: Probability of Initiating a Browsing Session

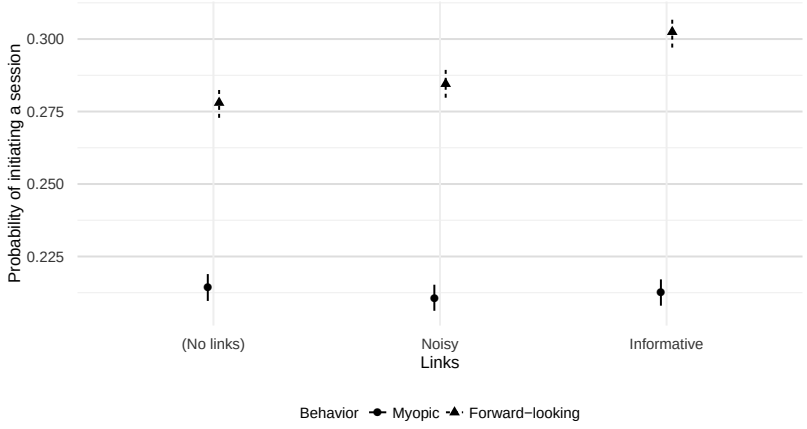
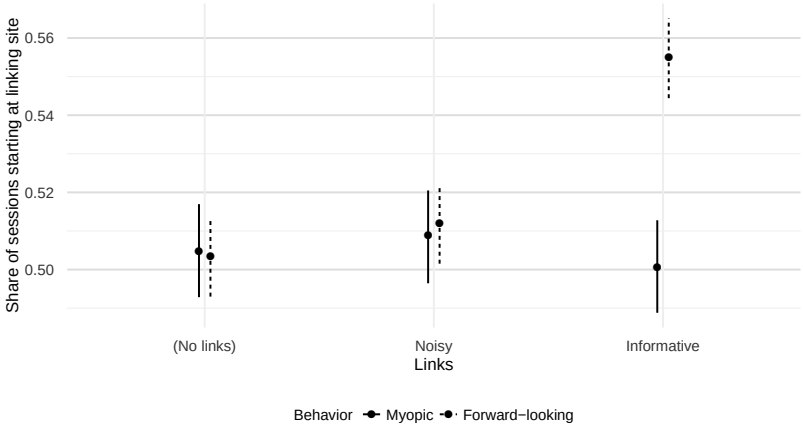


Figure OA2: Share of Sessions Starting at Linking Site (L)



Online Appendix B: Site Parameter Estimates for Alternative Models

Posterior estimates for the site parameters for horizontal (z_j) and vertical (α_j) quality are shown in Table OA1 for four model specifications. The first three (*no signals*, *myopic*, and *full*), which are presented in the main text, differ in the number of restricted parameters that are estimated. The no signals model assumes consumers are myopic, and that excerpts do not provide information about future horizontal match utility. The myopic model differs from the no signals model by estimating how much links affect consumers' choices of where to browse next. The improvement in fit is significant (see Table 6 of the main text). The full model differs from the myopic by estimating the discount parameter, thus allowing consumers to anticipate the value of links in their browsing choices.

Figure OA3: Number of Sites Visited Conditional on Browsing

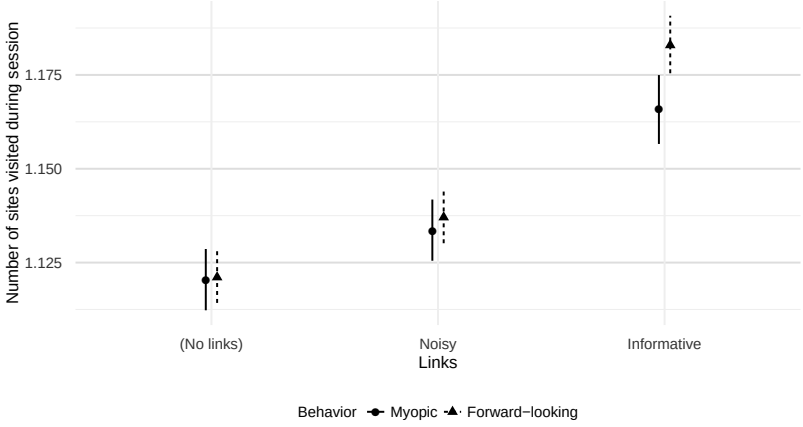
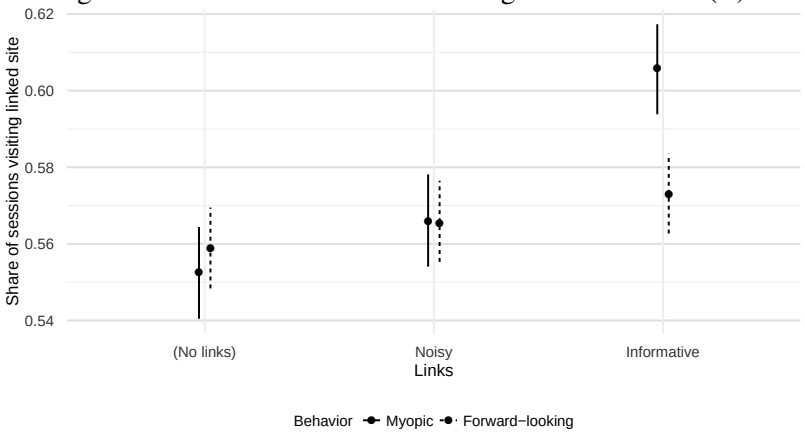


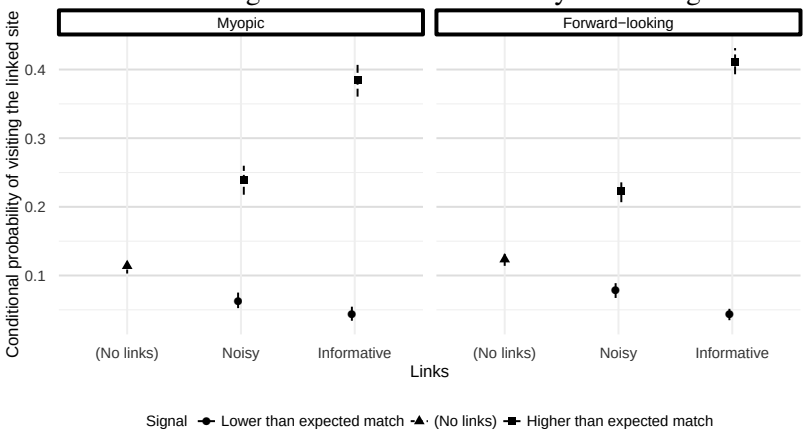
Figure OA4: Share of Sessions Visiting the Linked Site (R)



The fourth specification shown in Table OA1 (*larger sample*) is based on the full model, but estimated using a sample for which the inclusion criteria are less restrictive than those described in the manuscript. Specifically, the minimums for the criteria "number of sessions in each month visiting any of our 5 sites" are lowered from 5 to 4; for "total number of sessions visiting any of our 5 sites" from 16 to 12; and for "average number of sessions per month (visiting any site)" from 4 to 3. Using these less restrictive criteria increases the number of panelists from 127 to 155. Although the number of panelists increases by 22%, the number of site visits (i.e., non-zero observations) increases by only 9%.

Table OA1 shows that the site parameters related to vertical quality due to the volume of news published (α_j) are highly consistent across specifications, whereas the horizontal match locations (z_j) change

Figure OA5: Effect of Signal Valence on Probability of Visiting Linked Site



NOTES. Probabilities are calculated conditional on having chosen to visit the linking site (L) first in the session.

Table OA1: Site Parameter Estimates for Alternative Models

		Model			
	Parameter	No signals	Myopic	Full	Larger sample
Model feature					
Links and excerpts			x	x	x
Forward-looking consumers				x	x
Less restrictive inclusion criteria					x
Site					
celebuzz	z_j	0.38	0.57	0.95	0.94
		(0.28, 0.49)	(0.49, 0.64)	(0.85, 1.04)	(0.85, 1.02)
	α_j	0.034	0.032	0.03	0.031
		(0.026, 0.042)	(0.028, 0.036)	(0.025, 0.036)	(0.026, 0.038)
dlisted	z_j	-0.65	-0.63	-0.89	-0.8
		(-0.71, -0.58)	(-0.72, -0.56)	(-1.01, -0.69)	(-0.93, -0.70)
	α_j	0.22	0.2	0.22	0.23
		(0.21, 0.23)	(0.20, 0.21)	(0.21, 0.23)	(0.21, 0.26)
egotastic	z_j	-1.83	-1.95	-2.72	-2.69
		(-1.91, -1.77)	(-2.02, -1.88)	(-2.85, -2.60)	(-2.76, -2.62)
	α_j	0.0085	0.0099	0.0098	0.01
		(0.0029, 0.014)	(0.007, 0.013)	(0.0058, 0.015)	(0.0048, 0.017)
perezhilton	z_j	2.31	1.87	2.2	2.13
		(2.26, 2.37)	(1.84, 1.91)	(2.15, 2.25)	(2.09, 2.17)
	α_j	0.077	0.077	0.082	0.084
		(0.07, 0.084)	(0.072, 0.082)	(0.076, 0.088)	(0.078, 0.09)
thesuperficial	z_j	-0.22	0.14	0.45	0.43
		(-0.35, -0.10)	(0.045, 0.23)	(0.31, 0.59)	(0.29, 0.56)
	α_j	0.014	0.014	0.014	0.014
		(0.0082, 0.02)	(0.012, 0.017)	(0.01, 0.018)	(0.0098, 0.02)

more, while retaining the same general spatial configuration.

Online Appendix C: Selection of Consumers for the Estimation Sample

Here we compare the 127 panelists used for estimation to the full set of panelists who visited any of the five sites in Q4 of 2009. Figure OA6 summarizes and compares demographic variables for these groups. Compared to the full panel, the estimation sample has a higher proportion of consumers who are female, age 25–55, and have higher income. In terms of browsing behavior at the five sites, there are differences due to the selection criteria. In the estimation panel, sessions are more prevalent (averaging 45.3 versus 3.7 in Q4 2009) and longer (averaging 1.27 sites visited versus 1.04) compared to the full panel. 75% of consumers in the full panel have only 1 or 2 site visits in Q4 2009, hence the 127 individuals in the estimation panel account for 13% of site visits. The estimation sample is by construction more frequent and consistent in their site visits. But apart from this, we see no difference between the samples on the other observable variables.

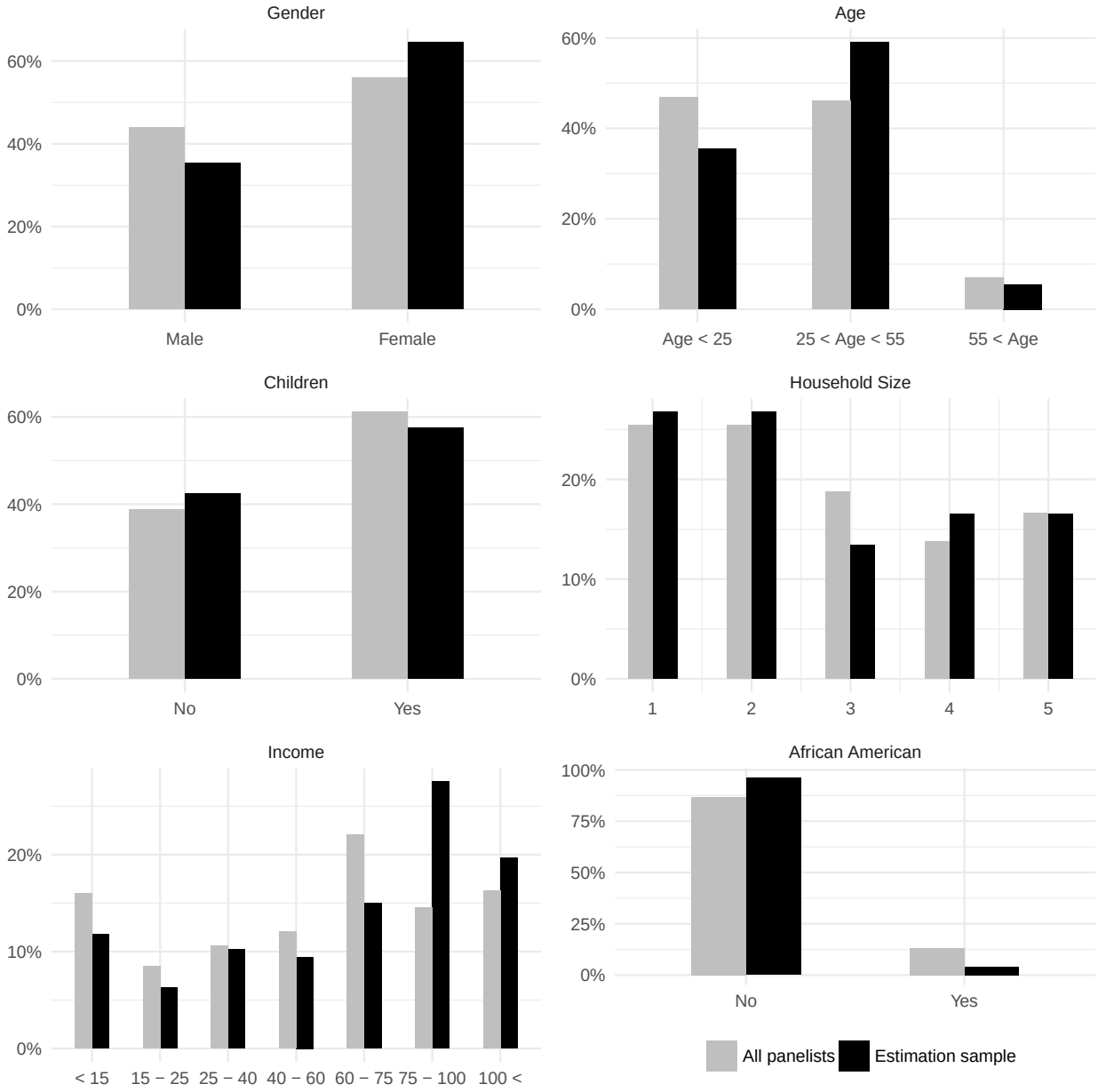
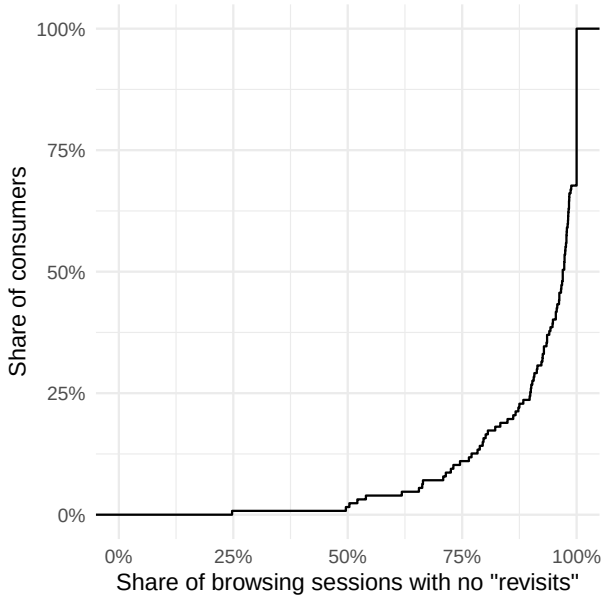


Figure OA6: Comparison of Full Panel with Estimation Sample

Online Appendix D: Revisits in the Browsing Data

Figure OA7 shows the cumulative distribution of browsing sessions with no "revisits" in the raw data. A browsing session is said to have a revisit in the raw data when: 1) there is one or more page requests for a given site, 2) these page requests are followed temporally by one or more page requests at a different site, and 3) this second set of page requests is followed temporally by one or more page requests for the original site. As noted in the main text, these patterns may reflect decisions to revisit earlier sites (true positives), or

Figure OA7: Distribution of Browsing Sessions without Revisits in the Raw Data



NOTES. The distribution depicts the share of consumers (y-axis) for whom a given share of browsing sessions did not contain a revisit (x-axis).

Table OA2: Comparison of Step 1 Traffic by Number of Outbound and Inbound Links

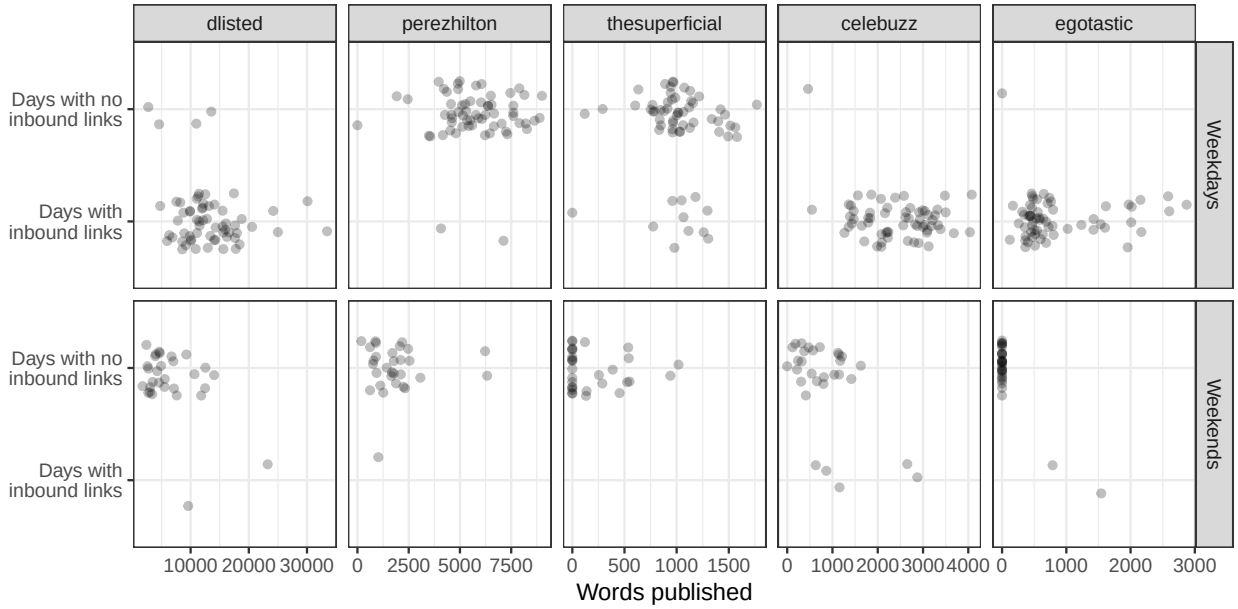
Outbound Links at First Site	Visits at Step 1	Inbound Links at First Site	Visits at Step 1
0	.76	0	.71
	(.34)		(.34)
1	.75	1	.69
	(.18)		(.18)
2-3	.81	2-3	1.00
	(.33)		(.28)

browsing software refreshing pages in open tabs (false positives). Revisits do not appear to be common in our data.

Online Appendix E: Links and Browsing Sessions

The model assumes that consumers do not know whether sites have made outbound or received inbound links from other sites before they actually encounter those links. If this assumption were to be violated, we would expect to observe that on days when sites make more outbound or receive more inbound links, the number of browsing sessions starting at those sites is higher (or lower). To see if the data contradict this assumption, we performed the following analysis. First, we consider browsing sessions initiated on weekdays, as the amount of browsing is systematically different on weekends and holidays, and due to the small number of days, statistics calculated from weekend and holiday browsing are highly variable.

Figure OA8: Distribution of daily word counts by the presence or absence of inbound links



Second, for each day, we calculate the number of sessions that began at each site, the number of outbound links that site made, and the number of inbound links it received. Third, to account for differences in site popularity, we subtract the median daily sessions starting at each site. And fourth, for different numbers of outbound and inbound links, we calculate the average number of sessions starting at each site on days with that many links. Table OA2 shows the results of this analysis. There does not appear to be any systematic relationship between the amount of outbound or inbound links at each site, and the number of visitors those sites receive at the start of the browsing session.

Online Appendix F: Links and Word Counts

We do not model a direct effect of links on beliefs about the quantity of news information at the linked site. In our empirical setting, this direct effect would imply a meaningful correlation between 1) whether a site receives inbound links, and 2) the number of words it publishes. Figure OA8 shows the distribution of daily word counts, conditional on whether the site received inbound links, and controlling for site and weekends, as in the model. Within each panel, there does not appear to be a discernible pattern that would allow consumers to infer the amount of information at a site based on the mere existence of a link.

Algorithm OA1: MCMC sampling procedure. At each iteration, parameters are sampled in blocks. The value function is then iterated and the result either replaces the oldest saved iteration or is appended to the set of saved iterations.

initialize saved MCMC samples: Θ
saved value function iterations: $\mathcal{W}$
foreach *MCMC iteration* t **do**
  foreach *parameter block* $\theta \subseteq \theta^{(t-1)}$ **do**
    Propose new θ using mMALA proposal distribution:
      calculate marginal posterior probability: $p(\theta|\mathcal{W})$
        derivatives of log posterior probability: $\mathcal{D}_\theta$ // [1]
      set $(m, S) \leftarrow \mathcal{M}(\theta, \mathcal{D}_\theta)$ // [2]
      draw $\theta^c \sim N(m, S)$
    Maintain detailed balance:
      calculate $p(\theta^c|\mathcal{W})$ and $\mathcal{D}_{\theta^c}$ // [1]
      set $(m^\circ, S^\circ) \leftarrow \mathcal{M}(\theta^c, \mathcal{D}_{\theta^c})$ // [3]
    Accept or reject proposal:
      set $\alpha \leftarrow \frac{p(\theta^c)N(\theta|m^\circ,S^\circ)}{p(\theta)N(\theta^c|m,S)}$
      draw $u \sim U(0, 1)$
        if $u < \alpha$ **then set** $\theta_b^{(t)} \leftarrow \theta^c$
      else set $\theta_b^{(t)} \leftarrow \theta$
  Iterate value function using IJC:
    draw $I \sim f(I|\theta^{(t)})$
    calculate $\widehat{W} \leftarrow \mathfrak{f}(I, \theta^{(t)})$ using IJC // [4]
  Save parameters and value function:
    append $\mathcal{W} \leftarrow \{\widehat{W}, I, \theta^{(t)}\}$
    append $\Theta \leftarrow \theta^{(t)}$

Online Appendix G: Sampling Algorithm

The general sampling procedure is outlined in Algorithm OA1. This algorithm is an application of IJC (Imai, Jain, and Ching 2009), with a few differences. At the lines marked [1] in Algorithm OA1, a single procedure calculates both $p(\theta|\mathcal{W})$ and $\mathcal{D}_\theta$ using automatic differentiation. In the MMALA procedure (Girolami and Calderhead 2011), the value of $\mathcal{D}_\theta$ would typically contain derivatives of the log posterior density function. In our setting, however, $\mathcal{D}_\theta$ contains the derivatives of the log posterior function while ignoring the contributions to these derivatives from the IJC value function approximation subroutine. The loss in precision in calculating $\mathcal{D}_\theta$ is compensated for by lower computational burden and better numerical stability.

At the lines marked [2] and [3] in Algorithm OA1, the function $\mathcal{M}(\cdot,\cdot)$ indicates the MMALA proposal distribution described in Girolami and Calderhead (2011). At line [2], the proposal distribution is created conditional on the current parameter vector θ and the derivatives of the log posterior density function evaluated at the point θ, $\mathcal{D}_\theta$. At line [3], the proposal distribution is created conditional on the proposed parameter vector θ^c and the derivatives of the log posterior density function evaluated at the point θ^c, $\mathcal{D}_{\theta^c}$. The proposal distributions are not symmetric, and therefore do not cancel out of the Metropolis-Hastings accept/reject ratio.

Finally, the line marked [4] indicates calculation of a new value function iteration for step t of the IJC sample, as described in (Imai, Jain, and Ching 2009). IJC recommend increasing the efficiency of the sampler by using θ^c to calculate the next approximation of the value function. Because θ^c has greater distance from θ compared to a random walk sampler (owing to the MMALA proposal distribution), we found θ to be more efficient than θ^c for approximating the value function.

References

Girolami, Mark and Ben Calderhead (2011), "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 73 (2), 123–214.

Imai, Susumu, Neelam Jain, and Andrew T. Ching (2009), "Bayesian estimation of dynamic discrete choice models," *Econometrica,* 77 (6), 1865–1899.