Generalized Stochastic Gradient Learning[*]

George W. Evans
University of Oregon

Seppo Honkapohja
Bank of Finland and University of Cambridge

Noah Williams
Princeton University and NBER

May 18, 2008

Abstract

We study the properties of the generalized stochastic gradient (GSG) learning in forward-looking models. GSG algorithms are a natural and convenient way to model learning when agents allow for parameter drift or robustness to parameter uncertainty in their beliefs. The conditions for convergence of GSG learning to a rational expectations equilibrium are distinct from but related to the well-known stability conditions for least squares learning.

Key words: adaptive learning, E-stability, Bayesian learning, robust estimation

JEL classification: C62, C65, D83, E10, E17

1 Introduction

Over the past decade or two there has been a significant amount of macroeconomic research studying the implications of adaptive learning. This literature replaces rational expectations with the assumption that economic agents are boundedly rational but employ a learning scheme such as recursive least squares (RLS) algorithms to estimate and update the parameters of their forecasting model. A central issue in this literature is to obtain the conditions under which the economy with learning converges to a rational expectations equilibrium (REE) in the long run.

The basic learning setting presumes that the agents' perceptions take the form of a forecasting model with fixed unknown parameters, which they update over time as new data becomes available. Such a setting does not explicitly allow for parameter drift and regime

[*] We thank Chryssi Giannitsarou, Larry Samuelson, Felix Kubler, and two anonymous referees for useful comments. Financial support from National Science Foundation Grant No. SES-0617859 and ESRC grant RES-000-23-1152 is gratefully acknowledged.

switching or for model uncertainty and robustness. Both of these topics have received a lot of attention in the recent macroeconomics literature; for example, see the recent papers Sims and Zha (2006), Cogley and Sargent (2005), Primiceri (2006) and Sargent, Williams, and Zha (2006).

A standard way of modeling parameter drift makes use of the Kalman filter, which is known to provide the Bayesian estimator when the coefficients evolve according to Gaussian random walk. We develop an approximation to the Kalman filter, which is given by the following algorithm

$$\varphi_{t+1} = \varphi_t + \gamma \Gamma z_t (y_t - \varphi_t' z_t), \tag{1}$$

where φ_t are the estimates of the vectors of drifting parameters, y_t is the endogenous variable, and z_t is the vector of regressors. $\gamma\Gamma$ is the perceived covariance matrix of the parameter drift with Γ controlling the direction and γ the speed of the drift. As we will see, algorithm (1) also emerges as the maximally robust estimator in a setting where there is uncertainty about the true data generating process and one wants to employ an estimator that performs well across a number of alternative models. Additionally, this estimator is optimal under "risk-sensitivity" since it minimizes the expected exponential of the sum of squared errors.

In fact, the algorithm (1) with $\Gamma = I$ has been previously studied in the statistics and learning literature, where it is known as the constant-gain Stochastic Gradient (SG) algorithm.[1] We will, therefore, call (1) the Generalized Stochastic Gradient (GSG) algorithm. An advantage of the GSG over the classic SG algorithm is (as we will see) that it can preserve scale invariance.

The main focus of this paper is to consider stability of equilibrium under GSG learning in the context of standard macroeconomic models. In contrast to the classical statistical framework, macroeconomic models with expectations and learning are self-referential, i.e. the evolution of the endogenous variables is influenced by the learning process itself. This has the consequence that it is not a foregone conclusion that estimators will be consistent, as the feedback from the learning process to the evolution of the state variables may lead the overall system to fail to converge to an equilibrium.

It is well-known that an REE is locally stable under RLS learning if what are known as expectational stability (E-stability) conditions hold. Some recent papers have examined the relationship between E-stability and convergence of classic SG learning in specific models.[2] We clarify and extend the existing results by considering the conditions for convergence of GSG learning for a general class of self-referential linear models. We develop sufficient conditions that strengthen E-stability and guarantee GSG-stability for all weighting matrices, Γ. We also show that E-stability and GSG-stability are equivalent in "diagonal" environments. Finally, convergence of Bayesian learning in self-referential models is demonstrated when the

[1] In decreasing-gain versions of the SG algorithm γ is replaced by a decreasing gain sequence such as $1/t$.

[2] The two conditions are identical in cobweb-type models; see Evans and Honkapohja (1998). In models with dependence on expectations of future values of the endogenous variables, the correspondence between E-stability and convergence of SG learning no longer holds. See Barucci and Landi (1997) and Heinemann (2000). Giannitsarou (2005) provides an economic example with lagged endogenous variables in which E-stability of the fundamental REE does not imply convergence of SG learning.

GSG stability holds.

2 Bayesian and Robust Justifications for the GSG Algorithm

2.1 Bayesian Interpretation of GSG

We show here that for any prior of the form of Gaussian random walk parameter drift there is a corresponding GSG algorithm that approximates the Bayesian estimator. We generalize Sargent and Williams (2005), who consider Kalman filter estimation when parameter drift is modeled according to a random walk hypermodel with a specific prior that is associated with RLS estimation.

We suppose that an agent believes that the data are generated by the drifting coefficients model:

$$\begin{aligned} y_t &= \beta_{t-1}' z_t + \eta_t && (2) \\ \beta_t &= \beta_{t-1} + \Lambda_t && (3) \end{aligned}$$

where η and Λ are viewed as mean zero Gaussian shocks with $E\eta_t^2 = \sigma^2$ and $\text{cov}(\Lambda_t) = V << \sigma^2 I$. We assume that y_t is a scalar.

The agent's estimator is $\varphi_t \equiv \hat{\beta}_{t|t-1}$, the optimal estimate of β_t conditional on information up to date $t-1$. It is well known that the (Bayes) optimal estimates in this linear model are provided by the Kalman filter. The Kalman filtering equations here are:

$$\begin{aligned} \varphi_{t+1} &= \varphi_t + \frac{P_t}{1 + z_t' P_t z_t} z_t (y_t - \varphi_t' z_t) && (4) \\ P_{t+1} &= P_t - \frac{P_t z_t z_t' P_t}{1 + z_t' P_t z_t} + \sigma^{-2} V. && (5) \end{aligned}$$

Here $\text{cov}(\varphi_t - \beta_t) \equiv \sigma^2 P_t$.

While one could in principle work directly with the Kalman filtering equations, and indeed we do so below in Proposition 8, it is convenient in practice to consider some approximations. This will allow us to more readily study the analytic properties of the learning rules, compare different specifications, and link our results to a broad previous literature. Benveniste, Metivier, and Priouret (1990) note that for large t (5) is well approximated by:

$$P_{t+1} = P_t - P_t M_z P_t + \sigma^{-2} V,$$

where $M_z = E z_t z_t'$.[3] Using this approximation (and assuming $1/(1 + z_t' P_t z_t) \approx 1$), the Kalman filter equations simplify to:

$$\begin{aligned} \varphi_{t+1} &= \varphi_t + P_t z_t (y_t - \varphi_t' z_t) && (6) \\ P_{t+1} &= P_t - P_t M_z P_t + \sigma^{-2} V. && (7) \end{aligned}$$

[3] It would be possible to extend the analysis to models with lagged endogenous variables. The extension is discussed in the 2006 working paper version.

The approximations which we've used are valid for slowly varying models, in which V is "small" in a sense made clear below. In what follows we use the fact that (7) is independent of φ_t to first analyze the limit of $\{P_t\}$. Then we use this limiting evolution as an approximation in (6). In particular, we have:

Lemma: *Under (7), P_t locally converges to the unique positive definite matrix P that solves the equation:*

$$PM_zP = \sigma^{-2}V. \tag{8}$$

Proofs of this and other results are in Appendix A.

The prior belief V on the form of the parameter drift influences the estimator, as it influences the speed and direction along which parameters should be updated. In particular, suppose we normalize V by writing:

$$V = \gamma^2\sigma^2\Omega,$$

where γ controls the overall speed of the parameter drift and Ω specifies the direction of the drift (e.g. we might normalize by setting $\det(\Omega) = 1$). Since the P_t recursion (7) converges, the limit P satisfies:

$$PM_zP = \gamma^2\Omega \text{ or } (\gamma^{-1}P)M_z(\gamma^{-1}P) = \Omega.$$

Therefore, letting $\Gamma = \gamma^{-1}P$ we have:

$$\Gamma M_z\Gamma = \Omega = \gamma^{-2}\sigma^{-2}V. \tag{9}$$

Substituting P for P_t and using this relation, we see that asymptotically the parameter estimates satisfy the GSG algorithm (1):

Proposition 1 *The GSG algorithm (1) asymptotically approximates the Bayesian optimal estimation (4)-(5) for model (2)-(3).*

Note that while (1) has the same asymptotic behavior as (4)-(5) under the assumed form of V (since P_t converges to $\gamma\Gamma$), the transient responses from arbitrary initial conditions in general differ.

Thus, the choice of the gain matrix Γ in the GSG rule is closely tied to the prior V on the parameter drift. Just as a prior is specified in advance and fixed throughout the sample (although of course the posterior is updated), the gain matrix Γ is specified based on *a priori* assumptions or knowledge about the regressors. Sargent and Williams (2005) apply results from Benveniste, Metivier, and Priouret (1990) to show that if $V = \gamma^2\sigma^2M_z^{-1}$, then the Kalman filter is closely related to a constant gain RLS algorithm, as $\Gamma = M_z^{-1}$.[4] Alternatively, suppose that instead of being proportional to the *ratio* of the observation

[4] The two rules have the same limits, but the transient phases differ. In the application of Sargent and Williams (2005), the Kalman filter converges faster.

noise variance to the covariance matrix of the regressors (M_z), the parameter innovation covariance matrix is proportional to the *product* of the two: $V = \gamma^2\sigma^2 M_z$. In this case $\Gamma = I$ and the classic SG rule results. More generally, the prior on V will lead through (9) to a particular choice of the optimal matrix Γ in the GSG rule.

2.2 GSG Algorithm and Robustness

In the previous section, we showed that the GSG rule is an approximate optimal predictor for a particular model of parameter variation. However, while the random walk form of parameter variation is commonly used, it is quite particular, and its appropriateness in any given application is an open issue. In this section, we provide a motivation that is in some ways more general in that it encompasses a range of different model specifications. More particularly, if the correct specification of the model is not known, then one may want to choose an estimator which is robust in that it performs well across a range of alternatives. Here we show that the GSG algorithm is such a robust optimal estimator. Our results here follow Hassibi, Sayed, and Kailath (1996). Robust control methods have recently been applied to a range of economic problems (see Hansen and Sargent (2007) for an overview), and it is interesting to see that the GSG algorithm has a robust interpretation.

In particular, we suppose that the true coefficients are believed to be constant over time, but we are uncertain about the data generating process. We represent this by a variation on (2)-(3) which now takes the form:

$$\begin{aligned} y_t &= \beta_{t-1}' z_t + \eta_t \\ \beta_t &= \beta_{t-1}, \quad \beta_0 = \beta. \end{aligned} \tag{10}$$

But now, instead of the η_t shocks having a Gaussian distribution, η_t is treated as an approximation error without a specified probability distribution. The η_t shocks are introduced as a means of capturing the possible misspecification of the model, and they may be both autocorrelated and correlated with the state z_t.

The key assumption in (10) is that there is no (unique) prior probability distribution over the shocks. A sequence of predictors φ_t is chosen to minimize the prediction errors:

$$e_t = \beta' z_t - \varphi_{t-1}' z_t, \tag{11}$$

but acknowledging the potential misspecification error. In particular, we treat (10) with $\eta_t \equiv 0$ as a benchmark model, but consider a set of perturbations in a neighborhood of this model. As we cannot evaluate the likelihood of potential perturbations, we guard against the worst case in the set of possibilities.

More specifically, instead of minimizing the expected squared errors as in the Kalman filter case, we now solve a minimax problem. At date zero we have an initial estimate φ_{-1} of the true value β, with prior precision $(\gamma\Gamma)^{-1}$, where Γ is a symmetric, positive definite, nonsingular matrix and γ a positive constant. Our use of the same notation as above is not coincidental, as again γ is a scale parameter while Γ governs the "shape" of the prior precision. Then the estimation problem is:

$$\min_{\{\varphi_s\}} \max_{\{\eta_s\},\beta} \sum_{s=0}^{t} |e_s|^2$$

subject to (10), (11), and:

$$\sum_{s=0}^{t} |\eta_s|^2 + \frac{1}{\gamma}(\beta - \varphi_{-1})'\Gamma^{-1}(\beta - \varphi_{-1}) \leq \mu. \tag{12}$$

Here $\mu > 0$ measures the size of the set of the alternative models, which are represented by different values of the parameter β and the shocks η_s satisfying the bound (12). As is standard, we can convert the problem from a constrained to a penalized one by putting a Lagrange multiplier $\theta > 0$ on the constraint (12). Then we can re-write the problem as:

$$\min_{\{\varphi_s\}} \max_{\{\eta_s\},\beta} \sum_{s=0}^{t} \left(|e_s|^2 \ - \theta|\eta_s|^2\right) - \frac{\theta}{\gamma}\ (\beta - \varphi_{-1})'\Gamma^{-1}(\beta - \varphi_{-1}), \tag{13}$$

subject to (10) and (11), where we leave off the inessential term in μ. Notice that θ and μ are inversely related, so we can use θ as a measure of the size of the set of alternatives, which hence is a measure of robustness.

As θ increases to infinity, perturbations are penalized more, and the size of the set of alternatives shrinks ($\mu \rightarrow 0$) to just the baseline model. There is also a lower bound $\underline{\theta}$ for θ which makes the problem well-posed, and this "maximally robust" critical value is the square of the so-called H_∞ norm of the system, see Hansen and Sargent (2007). This is the largest set of uncertainty μ that the problem can tolerate, and also has an interpretation as what is known as an induced norm. Loosely speaking, the H_∞ norm of a system represents the maximum factor by which errors in inputs get translated into errors in outputs.

The robust estimation problem (13) is a special type of a robust control problem, and in turn is equivalent to a H_∞ estimation problem, see Hassibi, Sayed, and Kailath (1996). The solution is known to have the following form:

$$\begin{aligned}
\varphi_{t+1} &= \varphi_t + K_t z_t (y_t - z_t'\varphi_t) && (14)\\
K_t &= \frac{\left(P_t^{-1} - \theta^{-1} z_t z_t'\right)^{-1}}{1 + z_t'\left(P_t^{-1} - \theta^{-1} z_t z_t'\right)^{-1} z_t} \\
P_{t+1}^{-1} &= P_t^{-1} + (1 - \theta^{-1}) z_t z_t', && (15)
\end{aligned}$$

with $P_{-1} = \gamma\Gamma$. Note the similarities between these equations and the Kalman filter algorithm in (4)-(5) with $V = 0$. In particular, as $\theta \rightarrow +\infty$ we see that they coincide.

While the robust rule collapses to the Kalman filter as the level of robustness decreases, it is more interesting in this case to consider the maximally robust learning rule with $\theta = \underline{\theta}$.

Hassibi, Sayed, and Kailath (1996) show that if $\lim_{T\to\infty}\sum_{t=0}^{T} z_t' z_t = +\infty$ and $(\gamma\Gamma)^{-1} > \sup_t z_t z_t'$ (i.e. the difference is a positive definite matrix), then $\underline{\theta} = 1$.[5] Recalling our discussion above, if $\underline{\theta}$ were greater than one, then the learning rule would magnify the effect of modeling errors on estimation errors. But here the maximally robust learning rule allows for no such magnification, and hence performs well in the face of misspecification.

Under these conditions, setting $\theta = \underline{\theta} = 1$, we see from (15) that $P_t = \gamma\Gamma$ for all t. This in turn implies that:

$$\begin{aligned} K_t &= \frac{\left(\gamma^{-1}\Gamma^{-1} - z_t z_t'\right)^{-1}}{1 + z_t'\left(\gamma^{-1}\Gamma^{-1} - z_t z_t'\right)^{-1} z_t} \\ &= \gamma\Gamma, \end{aligned}$$

where the last equality follows from the matrix inversion lemma. Thus the "gain matrix" K_t in the maximally robust learning rule (14) is constant over time, and thus this rule is the constant gain GSG rule (1) from above. The following proposition summarizes this discussion. A more formal proof is given in the appendix.

Proposition 2 *Given prior precision $(\gamma\Gamma)^{-1}$ on β, the GSG algorithm (1) is the maximally robust learning rule.*

2.3 Constant Gain GSG Learning and Risk Sensitivity

The previous section showed that the constant gain GSG learning rule was the (maximally) robust optimal predictor. This derivation was completely deterministic and relied on minimizing the worst case performance of the predictor over a certain class of alternative models. In this section we briefly discuss a different interpretation of these results in a stochastic setting with enhanced risk aversion, known as risk-sensitivity.[6] Once again we follow Hassibi, Sayed, and Kailath (1996).

Consider again the state space model (10), where now β and η are Gaussian random variables with means φ_{-1} and 0 and variances $\gamma\Gamma$ and I respectively. Then instead of minimizing the expected sum of squared errors as in the Kalman filter case, suppose that we solve the following:

$$\min_{\{\varphi_s\}} 2\theta \log E \exp\left(\frac{1}{2\theta}\sum_{s=0}^{t} |e_s|^2\right) \tag{16}$$

subject to (10) and (11). This exponential adjustment of the objective function increases risk aversion, and hence (16) is known as a risk-sensitive optimization problem (see Whittle (1990) for a monograph on problems of this type). This can also be thought of as a particular choice of an undiscounted recursive utility objective as in Epstein and Zin (1989). This

[5] Hassibi, Sayed, and Kailath (1996) set $\Gamma = I$, but allowing for more general weighting matrices Γ is straightforward.

[6] Applications of risk-sensitivity in economics include Tallarini (2000) and Anderson (2005). See Hansen and Sargent (2007) for further discussion.

specification shares some features of the Bayesian Kalman filter setup we began with, in that it exploits the linear Gaussian nature of the model. However the interpretation of Γ is closer to the robust filtering approach above, as it captures the prior uncertainty about a fixed parameters. The degree of risk sensitivity is captured by $1/\theta$, so as θ decreases the loss function is more sensitive to extreme events. There is a maximal degree of risk sensitivity, which Whittle (1990) calls "the point of the onset of neurotic breakdown," beyond which the objective is not defined.

While being motivated as an enhanced adjustment to risk instead of robustness against unknown disturbances, there are well-established results linking the solutions of robust and risk-sensitive control problems.[7] In particular, as shown by Hassibi, Sayed, and Kailath (1996), the risk sensitive optimal filter solving (16) is identical to the robust optimal filter (14)-(15) above. Loosely speaking, the risk sensitivity adjustment skews the loss and gives more weight to the tails of the distribution. As θ decreases, the weight on the tails increases. Finally, as θ approaches its lower bound, the loss function weights the maximal prediction error. Thus for the maximally risk sensitive level (with the smallest θ), minimizing the exponentially tilted expected squared prediction errors is the same as minimizing the worst case prediction error. So for the maximally robust level of $\theta = \underline{\theta} = 1$, which is also the maximally risk sensitive value, risk-sensitive optimal predictor is again the constant gain GSG rule (1).

3 GSG Algorithms in Self-Referential Models

We now take up self-referential models and the stability of REE when agents employ GSG learning because they are either (approximate) Bayesians concerned with parameter drift or because they are concerned about model uncertainty and robustness (or have extreme risk sensitivity).

3.1 The Basic Framework

We study GSG learning within the multivariate linear forward-looking model

$$\begin{aligned} y_t &= \alpha + AE_t^* y_{t+1} + Bw_t + \eta_t, \\ w_t &= Fw_{t-1} + e_t, \end{aligned} \tag{17}$$

where y_t is $n \times 1$, the $k \times 1$ observed exogenous variables w_t are assumed to follow a known vector autoregression (VAR), and the unobserved shock η_t is white noise. The innovation e_t has zero mean and covariance matrix Σ_e. F is assumed to be invertible with roots inside the unit circle. The asymptotic covariance matrix $\lim_{t\to\infty} Ew_t w_t' = M_w$ is positive definite and

[7] As noted above, these connections go back to Jacobson (1973), but the explicit formulation given here was established by Glover and Doyle (1988). See Whittle (1990), Hassibi, Sayed, and Kailath (1996), and Hansen and Sargent (2007) for further discussions.

solves the Lyapunov equation:

$$M_w = FM_wF' + \Sigma_e.$$

$E_t^* y_{t+1}$ denotes the expectations held by private agents, which under learning can differ from rational expectations (RE). This model has a unique RE solution of the form $y_t = \bar{a} + \bar{b}w_t$. This solution is often called the "fundamentals" or minimal state variable (MSV) solution.[8]

Under learning agents have a "perceived law of motion" (PLM) of the form $y_t = a + bw_t$ and estimate the parameters a and b econometrically. Thus at time t agents have the estimated PLM:

$$E_t^* y_t = a_t + b_t w_t,$$

which implies the forecast function

$$E_t^* y_{t+1} = a_t + b_t F w_t.$$

To simplify the analysis we have assumed that F is known, but it would be straightforward to allow F also to be estimated, and our results would be in essence unaffected. Any given PLM induces an "actual law of motion" (ALM) that gives the temporary equilibrium value of y_t. This is obtained by substituting $E_t^* y_{t+1}$ into (17). For PLM estimates a_t, b_t we obtain

$$y_t = \alpha + Aa_t + (Ab_tF + B)w_t + \eta_t.$$

Introducing the notation $z_t' = (1, w_t')$ for the state variables and $\varphi_t' = (a_t, b_t)$ for the parameters, we can summarize the PLM at t as $y_t = \varphi_t' z_t$ and the ALM at t as $y_t = T(\varphi_t)' z_t + \eta_t$, where

$$T(\varphi)' = (\alpha + Aa, AbF + B). \tag{18}$$

The MSV RE solution is given by the fixed point of T, i.e. $\bar{\varphi}' = (\bar{a}, \bar{b})$, where $\bar{a} = (I - A)^{-1}\alpha$ and $\bar{b} = A\bar{b}F + B$.

In the current setting the generalized stochastic gradient algorithm for estimating and updating a_t, b_t is given by:

$$\varphi_t = \varphi_{t-1} + \gamma \Gamma z_{t-1}(y_{t-1} - \varphi_{t-1}' z_{t-1})', \tag{19}$$

we $\gamma > 0$ is a small scalar gain parameter. Substituting in the ALM we can write:

$$\varphi_t = \varphi_{t-1} + \gamma \Gamma z_{t-1} \left[z_{t-1}'(T(\varphi_{t-1}) - \varphi_{t-1}) + \eta_t' \right], \tag{20}$$

which is formally a (constant-gain) stochastic approximation or stochastic recursive algorithm. Provided a suitable stability condition is satisfied, with sufficiently small γ the time paths of (20) converge to a stochastic process near the REE. Here convergence is in the sense of weak convergence, as discussed e.g. in Chapters 7 and 14 of Evans and Honkapohja (2001) and Cho, Williams, and Sargent (2002). In this case we say that the REE is locally

[8] For simplicity, our main points are made within the purely forward-looking model (17). Our earlier working paper Evans, Honkapohja, and Williams (2006) shows how to extend the analysis of convergence of GSG learning to models with lagged endogenous variables.

stable for sufficiently small $\gamma > 0$. We remark that in the multivariate case we are assuming that the same weighting matrix $\gamma\Gamma$ is used for each of the endogenous variables. Generalization to heterogeneous weighting matrices would be possible but would take us outside GSG framework.

A well-known method for obtaining the convergence conditions is based on a study of stability of an ordinary differential equation that is associated with the recursive algorithm. In the current case of constant-gain learning the trajectories of differential equation give the mean dynamics of the stochastic process.[9] Convergence of φ_t depends, in particular, on the properties of the mapping $T(\varphi)$. For the system (20) the mean dynamics will converge locally to a fixed point $\bar{\varphi}$ of $T(\varphi)$ if $\bar{\varphi}$ is a locally stable equilibrium of the associated differential equation

$$\frac{d\varphi}{d\tau} = \Gamma M_z(T(\varphi) - \varphi), \tag{21}$$

where τ is notional or virtual time.[10] Since both Γ and M_z are positive definite, their product is nonsingular, which implies that the only equilibrium of the differential equation is the REE $\bar{\varphi}$.

Local stability conditions for $\varphi(\tau) \to \bar{\varphi}$ are given by the linearization of the matrix differential equation (21), giving

$$\frac{d \operatorname{vec} \varphi'}{d\tau} = (\Gamma M_z \otimes I)(DT' - I) \operatorname{vec} \varphi',$$

where "vec" refers to the vectorization of a matrix and DT' is the $n(k+1) \times n(k+1)$ Jacobian matrix, of the vectorized T' map given in (18), evaluated at the fixed point $\bar{\varphi}$.[11] Local stability of the differential equation requires that all eigenvalues of

$$(\Gamma M_z \otimes I)(DT' - I) \tag{22}$$

have negative real parts. In what follows we say that a matrix is *stable* if all of its eigenvalues have negative real parts. Equivalently, the *GSG stability condition* is that the matrix

$$(\Gamma M_z \otimes I) \begin{pmatrix} I \otimes A - I & 0 \\ 0 & F' \otimes A - I \end{pmatrix} \tag{23}$$

is stable. When $n = 1$ the matrix (23) can be simplified to

$$\Gamma M_z \begin{pmatrix} A - 1 & 0 \\ 0 & AF' - I \end{pmatrix}. \tag{24}$$

[9] See Evans and Honkapohja (2001), especially Chapters 6 and 7 for a discussion of stochastic recursive algorithms and the study of their convergence properties.

[10] Global convergence applies because equation (21) is linear here. Thus $\bar{\varphi}$ is in fact globally asymptotically stable if it is locally so.

[11] For an $m \times n$ matrix X, $\operatorname{vec} X$ is the $mn \times 1$ vector that stacks in order the columns of X. For the vectorization and matrix differential results see the summary in Section 5.7 of Evans and Honkapohja (2001). For a full discussion, see Magnus and Neudecker (1988).

3.2 The Classic SG Special Case

We remark that since $M_z = \text{diag}(1, M_w)$, for the classic SG algorithm the stability conditions are that $A - I$ and $(M_w \otimes I)((F' \otimes A) - I)$ are stable matrices. Here "diag" denotes a block diagonal matrix. It is also worth noting that setting $\Gamma = M_z^{-1}$, where $M_z = \lim_{t\to\infty} Ez_t z_t' = \text{diag}(1, M_w)$ delivers an algorithm that is asymptotically equivalent to RLS. In this case, one obtains the stability condition that the matrix $DT' - I$ is stable. Equivalently, $A - I$ and $F' \otimes A - I$ are stable matrices,[12] which are standard *E-stability conditions*. Note that in contrast to SG-stability, the E-stability conditions that govern convergence of RLS learning do not depend on M_w.

Remark: *SG- and E-stability conditions are not always the same, i.e. in general neither implies the other.*

Appendix B gives numerical examples showing that stability differences can arise in purely forward-looking models with a single endogenous variable and two exogenous variables. These are the simplest examples that can be provided, since with a single exogenous variable it is immediate that E-stability and SG-stability are equivalent.

Next, we comment more on SG- and E-stability. The fact that SG-stability depends on M_w suggests that the stability conditions depend on how the exogenous variables are measured. We now show that this is the case and with a suitable change of variables the two sets of stability conditions coincide.

We begin with the Cholesky decomposition:

$$M_w = QQ'.$$

This is always possible for a positive definite matrix, resulting in a matrix Q which is triangular and nonsingular.[13] Letting $L = Q^{-1}$ we have $LM_wL' = I$. Transforming independent variables to

$$\tilde{w}_t = Lw_t,$$

the RE solution becomes $y_t = \bar{a} + \hat{b}\tilde{w}_t$ where $\hat{b} = \bar{b}L^{-1}$.

Under SG learning with the transformed independent variables $\tilde{w}_t$ the PLM becomes

$$\begin{aligned} y_t &= a + \tilde{b}\tilde{w}_t + \eta_t, \text{ where} \\ \tilde{w}_t &= \tilde{F}\tilde{w}_{t-1} + \tilde{e}_t, \text{ with } \tilde{F} = LFL^{-1} \text{ and } \tilde{e}_t = Le_t. \end{aligned}$$

Note that $E\tilde{w}_t\tilde{w}_t' = I$. We have transformed the independent explanatory variables to orthogonal variables with unit variances. Clearly $\tilde{w}_t$ has the same information content as w_t, and thus they are equally good for forecasting. Furthermore, note that $\tilde{F}$ has the same eigenvalues as F because F and $\tilde{F}$ are similar matrices.

[12] We remark that the eigenvalues of $F' \otimes A - I$ are $f_k\lambda_i - 1$, where f_k and λ_i are eigenvalues of F and A, respectively. This follows since the eigenvalues of the Kronecker product of two matrices consist of the products of the eigenvalues of each matrix and, of course, F and F' have the same eigenvalues.

[13] For the Cholesky decomposition see, e.g., Hamilton (1994) pp. 91-2.

The SG-stability conditions for the transformed specification are that $A - I$ and $(M_{\tilde{w}} \otimes I)((\tilde{F}' \otimes A) - I)$ are stable matrices. But $M_{\tilde{w}} \otimes I = I$ and hence the SG-stability conditions are that $A - I$ and $\tilde{F}' \otimes A - I$ are stable matrices. Since $\tilde{F}$ and F have the same eigenvalues it follows that the SG-stability conditions of the transformed model reduce precisely to the E-stability conditions. We have therefore shown:

Proposition 3 *There exists a transformation of variables $\tilde{w}_t = Lw_t$, with L positive definite, under which SG-stability is equivalent to E-stability.*

There is a further aspect of the classic SG algorithms that has not received attention, namely that the SG algorithm is not scale invariant and thus the resulting estimates are affected by the choice of units. This is demonstrated in the Appendix, where we also show that the SG algorithm with a change of units is equivalent to a GSG algorithm with a non-trivial weighting matrix. In contrast, RLS is invariant to a change in units.

4 Stability Results

4.1 Stability for GSG Algorithms

We return to our main focus of GSG learning. The preceding result suggests that there may be useful results for GSG stability that do not depend on M_w. Suppose we transform the exogenous variables using the Cholesky transformation, so that $\tilde{w}_t = \tilde{F}\tilde{w}_{t-1} + \tilde{e}_t$ and $M_{\tilde{w}} = I$. The GSG-algorithm is given by (19) with z_t replaced by $\tilde{z}_t' = (1, \tilde{w}_t')$, and where we allow for a possibly different weighting matrix $\tilde{\Gamma}$. The GSG-stability condition would then be that the matrix

$$(\tilde{\Gamma} \otimes I) \begin{pmatrix} I \otimes A - I & 0 \\ 0 & \tilde{F}' \otimes A - I \end{pmatrix} \tag{25}$$

is stable.

To obtain further results we introduce the definitions:[14]

Definition: *A matrix C is H-stable if all the eigenvalues of HC have negative real parts whenever H is a positive definite matrix.*

Definition: *A matrix C is D-stable if all the eigenvalues of DC have negative real parts whenever D is a positive diagonal matrix.*

A sufficient condition for convergence of GSG learning is:

Proposition 4 *Consider model (17) and a GSG algorithm (19). Suppose that the matrix*

$$\begin{pmatrix} I \otimes A - I & 0 \\ 0 & \tilde{F}' \otimes A - I \end{pmatrix} \tag{26}$$

[14] Honkapohja and Mitra (2006) use H-stability to obtain sufficient stability conditions in the context of structural and learning heterogeneity.

is H-stable, where $\tilde{F} = LFL^{-1}$. Then GSG-stability holds for all Γ and the REE is locally stable under GSG learning.

This is a strengthening of the E-stability condition, which is that the matrix (26) is stable. The property of H-stability is quite restrictive. A sufficient condition for H-stability of a matrix C is that C is negative quasi-definite, i.e. that $C + C'$ is negative definite, i.e. has negative eigenvalues.[15] Note that if Γ is diagonal, then the set of sufficient conditions is that (i) $A - I$ is a stable matrix and (ii) $(\tilde{F}' \otimes A) - I$ is D-stable. There exist various necessary or sufficient condition for D-stability, but a full characterization is apparently not available (this is in contrast to H-stability).[16]

Corollary: *Assume E-stability. If in addition $\tilde{F}$ is symmetric with positive eigenvalues and $A - I$ is negative quasi-definite then GSG-stability holds for all Γ.*

The conditions in the Corollary can be convenient to apply, but they are much stronger than E-stability of the MSV REE.

In the next case of diagonal Γ and uncorrelated exogenous variables, it is unnecessary to transform variables into Cholesky form in order to obtain stability results.

Proposition 5 *When Γ, F and Σ_e are diagonal, E-stability and GSG-stability are equivalent.*

This case arises in applications, as illustrated below.

There are some further special cases in which E-stability guarantees convergence of GSG learning. In many cases it is natural to assume zero off-diagonal elements of the first row of Γ. Assuming again scale-invariance so that the Cholesky transformation can be used, we have the following result for models with a scalar endogenous variable:

Proposition 6 *Assume that $n = 1$ with $|A| < 1$ and that $\Gamma = \operatorname{diag}(\tilde{\Gamma}, \hat{\Gamma})$ for $\tilde{\Gamma} > 0$. If the largest singular value of $\tilde{F}$ is not greater than one, then E-stability implies GSG-stability for all Γ.*

We recall that the largest singular value of $\tilde{F}$ is equal to the largest eigenvalue of $\tilde{F}\tilde{F}'$.

4.2 Convergence of Bayesian Learning

Returning to the economic model (17) in the scalar case $n = 1$, it follows that if agents are updating their estimates according to the approximate Bayesian learning rule (6)-(7) then for small γ local stability is determined by (24). We have the following result.

[15] See, for example, Arrow and McManus (1958). They refer to H-stability as S-stability. Necessary and sufficient conditions for H-stability are given in Carlson (1968).

[16] See Arrow (1974) and Johnson (1974).

Proposition 7 *(i) The REE $\bar{\varphi}$ is locally stable for sufficiently small $\gamma > 0$ under the approximate Bayesian learning rule (6)-(7) if the matrix (24), with Γ defined in (9), is stable. (ii) The approximate Bayesian learning rule (6)-(7) is invariant to a change of variables to $\tilde{z}_t = Dz_t$ for any positive definite matrix D.*

We now ask whether there are restrictions on the economic model that guarantee local stability of Bayesian learning for all priors on the parameter drift that are sufficiently small. Combining the Propositions 4 and 7 we obtain the following result.

Corollary: *If the matrix*

$$\begin{pmatrix} A-1 & 0 \\ 0 & A\tilde{F}'-I \end{pmatrix} \tag{27}$$

is H-stable, then under the approximate Bayesian learning rule (6)-(7), the REE $\bar{\varphi}$ is locally stable for all priors V on the parameter drift, where V is sufficiently small (in the sense of small γ in the normalization above).

Note that the Corollary gives a condition for local stability, for all sufficiently small priors V, that holds regardless of how the variables w_t are measured. We also remark that the condition given is a strengthening of E-stability, since the latter is equivalent to $A < 1$ and stability of the matrix $A\tilde{F}' - I$.

The convergence results just described can be extended to the system under exact Bayesian learning.

Proposition 8 *The REE $\bar{\varphi}$ is locally stable for sufficiently small $\gamma > 0$ under the Bayesian learning rule (4)-(5) if the matrix (24), with Γ defined in (9), is stable.*

Remarks on stability of robust estimation: For the self-referential economic model (17), the issue of local stability of the REE also arises in the context of robust estimation. Since, according to Proposition 2, the maximally robust learning rule takes the form of the GSG algorithm (1), the earlier stability results apply directly. For a given Γ the REE is locally stable if the matrix (23) is stable. If the matrix (27) is H-stable then the REE is locally stable for all Γ and all γ sufficiently small. We also remark that using an argument along the lines of part (ii) of Proposition 7, it can be shown that the maximally robust learning rule is invariant to a change of variables $\tilde{z}_t = Dz_t$ since the corresponding initial estimate of $\tilde{\beta}$ becomes $\tilde{\varphi}_{-1} = D^{-1}\varphi_{-1}$ with prior precision $(\gamma D^{-1}\Gamma D^{-1})^{-1} = D(\gamma\Gamma)^{-1}D$.

This setup is particularly interesting in the context of the self-referential model with learning, since under robust estimation agents explicitly allow for possible misspecification. This contrasts, for example, with standard least-squares learning formulations in which agents ignore a transitory misspecification. There agents assume that the true regression parameters are constant over time, while in reality, under learning, the parameters are time-varying, although they converge over time to the (constant) REE values.

5 Economic Examples

5.1 A Scalar Model: The Performance of Different Rules

The linearized overlapping generations model with money and preference and endowment shocks leads to the equations (17), where the coefficient A is a scalar and the vector of exogenous variables w_t is 2×1. Different cases for values of A can be generated by different elasticity of substitution parameters. In particular, any value $A < 1$ is possible, including cases where $A < -1$. For convenience we assume that $M_z = I$, i.e. the normalization of variables has been done.

Our propositions imply GSG stability for a number of cases. First, if F is symmetric with positive eigenvalues and $A < 1$, then there is GSG stability for all Γ. This follows from the Corollary to Proposition 4. Second, if $|A| < 1$ and the largest eigenvalue of FF' is less than one, then by Proposition 6 GSG stability holds for all $\Gamma = \text{diag}(\tilde{\Gamma}, \hat{\Gamma})$. Finally, if $A < 1$ and the eigenvalues of $A(F + F')/2$ are less than one, then the matrix

$$\begin{pmatrix} A - 1 & 0 \\ 0 & AF' - I \end{pmatrix}$$

is H-stable and therefore GSG stability holds for all Γ. There are, however, cases in which E-stability holds but GSG stability fails. This was already seen in the numerical examples in connection with the classic SG special case.

We now analyze the performance of different learning rules in variations on this simple model. While we've shown above that GSG rules have justifications as estimators, the results above do not necessarily transfer to self-referential settings where agents' beliefs determine actual outcomes. However, here we see that the GSG rules do tend to outperform a version of the more commonly used constant-gain RLS learning rule. We also analyze two different Kalman filter specifications that also perform well. Moreover, our simulations show that for small gain the asymptotic approximations we applied above hold up fairly well, as the GSG rules track their Kalman counterparts relatively closely.

We take as the baseline model (17) as above with $n = 1$ and $k = 2$. We set the parameters as follows: $\alpha = 0$, $A = 0.9$, $B = [1, 1]$, $F = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}$, and $\Sigma_e = \begin{bmatrix} 0.15^2 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_\eta = 1$. Thus the only asymmetry between the two regressors is that the first is much less volatile.

Although our stability analysis has focused on the case where the parameters are fixed over time and the model is correctly specified, our motivations for considering GSG rules were that agents may allow for parameter drift or model misspecification. Thus we look at four different data generating processes which reflect these varying specifications. We consider:

- **M0:** Constant coefficients.
- **M1:** Drifting coefficients. The two elements of B follow independent random walks with Gaussian innovations as in (3) with $V = V1 \equiv \gamma^2\sigma^2 M_z^{-1}$, the prior consistent with RLS and GSG with $\Gamma = M_z^{-1}$.

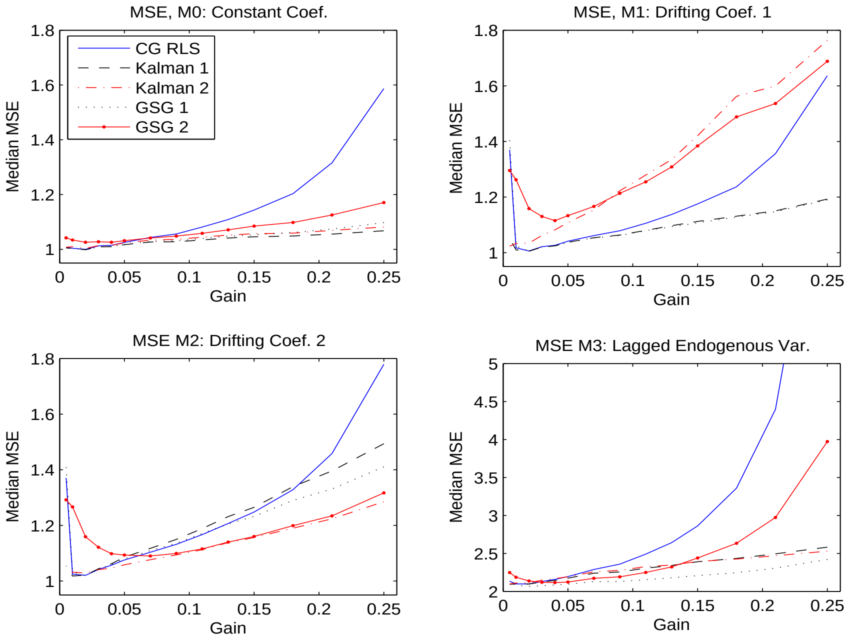

Figure 1: Mean squared forecast error under CG RLS, two Kalman filter specifications, and their corresponding GSG approximations. Data generated by constant coefficients (NK0), two different specifications of random coefficients (M1, M2) and an alternative model (NK4). Medians of 500 simulations of 1000 periods, showing gain vs. MSE.

- **M2:** Drifting coefficients. The two elements of B follow independent random walks with Gaussian innovations as in (3) with $V = V2 \equiv \gamma^2 \sigma^2 M_z$, the prior consistent with classic SG with $\Gamma = I$.

- **M3:** A misspecification. The model generating the data replaces the first equation of (17) with:
$$y_t = \rho_y y_{t-1} + \alpha + A E_t^* y_{t+1} + B w_t + \eta_t$$
with $\rho_y = -0.5$ and the other coefficients as in M0.

For each model we consider five different learning rules, constant-gain RLS, the Kalman filter with $V = V1$ (Kalman 1), the Kalman filter with $V = V2$ (Kalman 2), the GSG rule where $\Gamma = M_z^{-1}$ that approximates Kalman 1 (GSG 1), and the classic SG rule where $\Gamma = I$ that approximates Kalman 2 (GSG 2). In each case the perceived law of motion corresponds to the benchmark model in M0. Thus M1 and M2 explore different forms of parameter drift. M3 studies a particular type of misspecification, where the perceived law of motion neglects some dynamics. Here the dynamics come from partial adjustment of the type often used in empirical analysis of model of this class. We then run a small simulation study, whose results are shown in Figure 1. In each simulation run, we initialize agents' beliefs at a perturbation of the REE values with constant coefficients.[17] We then run each simulation for 2000 periods and discard the first 1000 observations, in order allow for convergence to a stationary distribution. We repeat this 500 times for a variety of gain settings. The figures plot the median mean squared forecast error (MSE) for the endogenous variable y_t from the 1000 simulations for different values of the gain γ.

Turning now to the relative performance of the different learning rules, note that for larger gain setting the CG RLS rule performed rather poorly across all of the specifications. By contrast, the GSG and Kalman filter rules all did well across all specifications. We expected that in M0 all the rules would do reasonably well, and they all did although CG RLS had some problems with relatively large gain settings. For gains of about 0.05 the all the rules were quite similar. For M1, Kalman 1 is the theoretically optimal choice, GSG 1 approximates it, and CG RLS is very closely related. Kalman 1 and GSG 1 did indeed fare quite well, with their similarity showing the validity of the approximations here, and the CG RLS rule nearly matched them for small gains. The other Kalman and GSG specifications imputed the wrong amount of parameter drift, and thus were not able to predict as well. A similar story holds for M2, where Kalman 2 is optimal and GSG 2 approximates it. These generally outperform the other rules, although for reasons that are not entirely clear, the performance of GSG 2 worsens for very small gains. Finally, for M3 none of the rules are optimal in any sense, but the robustness of the GSG rules shows through. For gains below 0.1 both GSG specifications outperform the other rules. CG RLS does very poorly in this model, particularly for larger gain settings.

In summary, these results largely confirm our theoretical predictions. We find that the approximations we use to justify GSG rules in a Bayesian setting are generally valid here, and that the performance and robustness analysis is mostly borne out in this application. Although we have shown above that the generalized stochastic gradient learning rules are optimal in situations of parameter drift, and provide robustness to model misspecification, that of course does not mean that all such rules are always preferable to other learning rules. In this example we've seen that the Kalman filter specifications, which the GSG rules approximate, perform relatively well. However, the GSG rules are nearly as good, and hold up better when the model is misspecified. Thus the stochastic gradient rules may sacrifice a bit of median performance in order to achieve greater robustness.

In interpreting the simulation results, it is important to recognize that the comparison across different learning rules assesses the performance when all agents use the same rule. In economic self-referential models, the choice of learning rule directly affects the stochastic process followed by the endogenous variables. A distinct question, which we do not examine in the current paper, is the performance of alternative estimation rules for an individual agent, given the rules used by the other agents in the economy. This would require a separate study.[18]

[17] In particular, each element of the belief vector is drawn from a normal distribution whose mean is the REE values and variance is $\Sigma_\eta M_z^{-1}$.

[18] For specific models and simple learning rules this question has been analyzed in Evans and Honkapohja (1993), Ch. 6 of Sargent (1999), Ch. 14 of Evans and Honkapohja (2001), Marcet and Nicolini (2003) and

5.2 New Keynesian Model

As a second example we consider the bivariate New Keynesian model of monetary policy, which is widely used in current discussions of monetary policy.[19] The key equations of the model take the form:

$$\begin{aligned} x_t &= c_x + E_t^* x_{t+1} - \sigma^{-1}(r_t - E_t^* \pi_{t+1}) + g_t, & (28) \\ \pi_t &= c_\pi + \kappa x_t + \mathcal{B} E_t^* \pi_{t+1} + u_t. & (29) \end{aligned}$$

Here x_t is the output gap, π_t is the inflation rate and r_t is the nominal interest rate. The parameters $\sigma, \kappa > 0$ and $0 < \mathcal{B} < 1$. c_x and c_π are intercepts, which are from the log-linearization of the exact model. These are usually suppressed by writing the model in terms of deviations. $w_t' = (g_t, u_t)$ consists of observable shocks to the output gap and inflation, respectively. The stochastic process for w_t has the form given in (17). The first equation is the IS curve that comes from the Euler equation for consumer optimality and the second equation is the forward-looking Phillips curve based on Calvo price stickiness.

The model is completed by specification of an interest rate rule. A wide variety of different rules have been studied in the literature.[20] One possibility is the standard Taylor rule:

$$r_t = c_r + \phi_\pi \pi_t + \phi_x x_t, \tag{30}$$

where c_r denotes an intercept. The parameters satisfy $\phi_\pi, \phi_x > 0$. Bullard and Mitra (2002) show that the E-stability condition under the standard Taylor rule is

$$\kappa(\phi_\pi - 1) + (1 - \mathcal{B})\phi_x > 0. \tag{31}$$

Alternatively, Evans and Honkapohja (2003b) consider optimal discretionary policy and show that the expectations-based interest rate rule

$$r_t = c_r + \left(1 + \frac{\sigma\kappa\mathcal{B}}{\alpha + \kappa^2}\right) E_t^* \pi_{t+1} + \sigma E_t^* x_{t+1} + \sigma g_t + \frac{\sigma\kappa}{\alpha + \kappa^2} u_t, \tag{32}$$

where α is the weight on output gap in a quadratic loss function of the policy-maker and c_r is an intercept, always leads to E-stability of the REE. If the two shocks g_t and u_t in model (28)-(29) are uncorrelated, Proposition 5 applies:

Proposition 9 *Assume that g_t and u_t are independent stationary $AR(1)$ processes and that Γ is diagonal.*
(i) Under the Taylor rule (30) the REE is GSG-stable if condition (31) holds, and
(ii) The REE is GSG-stable when optimal discretionary policy employs the expectations-based rule (32).

Evans and Ramey (2006).

[19] See e.g. Clarida, Gali, and Gertler (1999), Svensson (2003), and Woodford (2003) for details and analysis.

[20] The issue of stability under learning has been examined by Bullard and Mitra (2002) and Evans and Honkapohja (2003b) among others. Evans and Honkapohja (2003a) review the literature.

In particular the Proposition holds when agents learn using the classic SG algorithm. We next consider GSG-stability further under more general assumptions about the shocks g_t and u_t. For brevity, we restrict attention to the case where the policy-maker employs the Taylor rule (30). Introducing the notation $y_t = (x_t, \pi_t)'$, equations (28), (29) and (30) can be combined to yield the bivariate system of the form (17) with:

$$A = \frac{1}{\sigma + \phi_x + \kappa\phi_\pi} \begin{pmatrix} \sigma & 1 - \mathcal{B}\phi_\pi \\ \kappa\sigma & \kappa + \mathcal{B}(\sigma + \phi_x) \end{pmatrix}.$$

We omit the explicit form of B as it does not affect the stability conditions.

We give numerical examples of the above results using the calibration of the model due to Rotemberg and Woodford (1997) and widely employed in Woodford (2003).

Calibration: $\mathcal{B} = 0.99$, $\sigma = 0.157$, $\kappa = 0.024$.

Suppose first that the policy parameters take on values $\phi_\pi = 1.05$ and $\phi_x = 0.2$. The E-stability conditions on $A - I$ hold since both eigenvalues of $A - I$ are in the interval $(-1, 0)$. Furthermore, $A - I$ is negative quasi-definite as the eigenvalues of $(A - I) + (A - I)'$ are -1.186 and -0.0174. Assuming that the coefficient matrix $\tilde{F}$ of the vector of shocks in (17) is symmetric with positive eigenvalues, the sufficient conditions given in the Corollary to Proposition 4 are met and therefore GSG-learning is convergent under the specified policy parameter values.

As a second numerical example, set policy parameters at $\phi_\pi = 1.1$ and $\phi_x = 0.1$. E-stability on $A - I$ continues to hold, as the eigenvalues of $A - I$ are in the interval $(-1, 0)$, but $A - I$ fails to be negative quasi-definite since the eigenvalues of $(A - I) + (A - I)'$ are -0.987 and 0.0599. Since E-stability holds we also have GSG-stability for Γ close to I provided the regressors are put into Cholesky form, but GSG-stability may hold much more generally. For example, suppose $\tilde{F} = \text{diag}(\rho_1, \rho_2)$ with $|\rho_i| < 1$ and let $\Gamma = \text{diag}(\tilde{\Gamma}, \hat{\Gamma})$ for $\tilde{\Gamma} > 0$. We have conducted numerical searches over positive definite $\hat{\Gamma}$ and over $\tilde{F}$, and it appears that GSG-stability holds generally. Thus GSG learning appears to be convergent even though the sufficient condition of the Corollary to Proposition 4 fails. This suggests that Proposition 6 established above for the case where $n = 1$ may hold more broadly: apart from the fact that $n = 2$ here, all of the other conditions of the proposition hold. We conjecture that a version of the result extends to higher dimensions.

These numerical examples illustrate both the applicability and limitations of the preceding stability conditions.[21]

6 Conclusions

We have proposed GSG algorithms as a generalization of the classic SG algorithm and studied their properties in self-referential macroeconomic models. The constant-gain GSG

[21] The theoretical stability results and applications have analogues for the classic SG algorithm using untransformed exogenous variables, see the earlier working paper Evans, Honkapohja, and Williams (2006).

algorithm can be viewed as an approximate Bayesian learning scheme when agents allow for parameter drift in their beliefs. The GSG algorithm also has interpretations as the maximally robust optimal prediction rule and as the risk-sensitive optimal filter when there is parameter uncertainty.

The conditions for stability of REE under GSG learning differ from but are related to the E-stability conditions that govern stability under least squares learning. We developed several sufficient conditions under which E-stability of the REE implies convergence of GSG learning. We have also provided examples in which the two condition differ. We have also demonstrated that when the GSG stability conditions hold, the exact Bayesian learning converges to the REE.

Recent macroeconomic research has emphasized the role of (actual or perceived) parameter drift and robust decision-making for understanding macrodynamics. The current paper has shown how to extend the analysis of econometric learning by economic agents to settings where agents allow for these features.

Appendix

A Proofs of Results

Proof of Lemma of Section 2.1: Equation (8) is an algebraic Riccati equation. Existence of a unique positive definite solution P follows from Theorem 3.7 of Kwakernaak and Sivan (1972). The time invariant regulator associated with (8) takes the form $\dot{x}(t) = u(t)$ and $z(t) = x(t)$. Thus in their general set-up we are setting $A = 0$, $B = D = I$, $R_3 = \sigma^{-2}V$ and $R_2^{-1} = M_z$. The result requires that the associated regulator is stabilizable and detectable. Stabilizability is in turn implied by complete controllability (see their Theorem 1.27), which follows from their Theorem 1.23. Detectability is implied by complete reconstructability (see their Theorem 1.36), which follows from their Theorem 1.32. This establishes the existence of a unique P.

To show local convergence we linearize and vectorize (7), yielding

$$\operatorname{vec} dP_{t+1} = (I - PM \otimes I - I \otimes PM) \operatorname{vec} dP_t,$$

where dP_t refers to the deviation from the steady state P. The eigenvalues of the coefficient matrix are given by $1 - 2\lambda$ where the λ are the eigenvalues of PM. This follows from Theorem 4.4.5 of Horn and Johnson (1991) concerning the eigenvalues of the Kronecker sum of two matrices. Next, note that Theorem 7.6.3 of Horn and Johnson (1985) implies that the eigenvalues of PM are positive. Finally, for V sufficiently small, P is small and the eigenvalues of PM can be made small. Thus V sufficiently small implies that all eigenvalues of the coefficient matrix have modulus strictly less than one. Local stability follows.

Proof of Proposition 2: The result is a slight generalization of Theorem 6 of Hassibi, Sayed, and Kailath (1996). The proof here follows their Appendix A very closely.

First, we introduce the notation: $\|e\|^2 = \sum_{s=0}^{\infty} |e_s|^2$. Then as discussed above, it is well-known (see Hansen and Sargent (2007) among others) that minimal value $\underline{\theta}$ of the multiplier is the square of the H_∞ norm. In particular, we can rearrange the infinite horizon version of (13) to now have:

$$\underline{\theta} = \inf_{\{\varphi_s\}} \max_{\{\eta_s\},\beta} \frac{\|e\|^2}{\|\eta\|^2 + (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})}. \tag{33}$$

It is simple to show that $\underline{\theta} \geq 1$. For example, suppose that an initial estimate φ_{-1} is given and choose $\{\eta_s\}$ so that the observation agrees with the expected output, i.e. so that $y_t = \beta' z_t + \eta_t = \varphi_{-1}' z_t$. In this case $\varphi_t = \varphi_{-1}$ for all t, and thus the right side of (33) becomes:

$$\frac{\|\eta\|^2}{\|\eta\|^2 + (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})} = \frac{\|(\beta - \varphi_{-1})'z\|^2}{\|(\beta - \varphi_{-1})'z\|^2 + (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})}$$

Now with $\lim_T \sum_{t=0}^T z_t' z_t = +\infty$, for any $\varepsilon > 0$ we can find a β and an integer N such that $\sum_{s=0}^N |\beta - \varphi_{-1})' z_s|^2 \geq (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})$. Thus for these have:

$$\frac{\sum_{s=0}^N |(\beta - \varphi_{-1})' z_s|^2}{\sum_{s=0}^N |(\beta - \varphi_{-1})' z_s|^2 + (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})} \geq 1 - \varepsilon.$$

Since this was just one particular choice of β and $\{\eta_s\}$ we conclude that $\underline{\theta} \geq 1$.

We now show that the GSG learning rule attains $\underline{\theta} = 1$. First, note that by defining $\widehat{\varphi}_t = \beta - \varphi_t$, we can write the GSG updating equation as:

$$(\gamma\Gamma)^{-1/2}\widehat{\varphi}_t = (\gamma\Gamma)^{-1/2}\left[\widehat{\varphi}_{t-1} - \gamma\Gamma z_t(y_t - z_t'\varphi_t)\right].$$

Also, we can write:

$$\eta_t = y_t - \varphi_{t-1}' z_t - \widehat{\varphi}_{t-1}' z_t.$$

Then square both sides of these last two equations and subtract the results to obtain:

$$\widehat{\varphi}_t'(\gamma\Gamma)^{-1}\widehat{\varphi}_t - |\eta_t|^2 = \widehat{\varphi}_{t-1}'(\gamma\Gamma)^{-1}\widehat{\varphi}_{t-1} - \widehat{\varphi}_{t-1}' z_t - (I - \gamma\Gamma z_t z_t')(y_t - z_t'\varphi_t)^2.$$

Now since we've assumed that $(\gamma\Gamma)^{-1} \geq z_t z_t'$, the third term on the right is negative and thus we can rearrange this to get:

$$\widehat{\varphi}_{t-1}'(\gamma\Gamma)^{-1}\widehat{\varphi}_{t-1} + |\eta_t|^2 \geq \widehat{\varphi}_t'(\gamma\Gamma)^{-1}\widehat{\varphi}_t + |e_t|^2,$$

where we recall the definition of e_t. Then adding the inequalities of this form for each date from $s = 0, \ldots, t$ we have:

$$(\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1}) + \sum_{s=0}^t |\eta_s|^2 \geq \widehat{\varphi}_t'(\gamma\Gamma)^{-1}\widehat{\varphi}_t + \sum_{s=0}^t |e_s|^2 \geq \sum_{s=0}^t |e_s|^2$$

This in turn implies:

$$\frac{\sum_{s=0}^t |e_s|^2}{\sum_{s=0}^t |\eta_s|^2 + (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})} \leq 1.$$

Then taking the limit as $t \to \infty$ we thus have that for the GSG learning rule:

$$\frac{\|e\|^2}{\|\eta\|^2 + (\beta - \varphi_{-1})'(\gamma\Gamma)^{-1}(\beta - \varphi_{-1})} = 1.$$

Thus the GSG attains the minimal value of 1 and so is the maximally robust learning rule.

Proof of Proposition 4: We consider the matrix (23)

$$\begin{aligned} & (\Gamma M_z \otimes I)\begin{pmatrix} I \otimes A - I & 0 \\ 0 & F' \otimes A - I \end{pmatrix} \\ = \ & (\Gamma M_z \otimes I)\left(\begin{pmatrix} I & 0 \\ 0 & F' \end{pmatrix} \otimes A\right) - (\Gamma M_z \otimes I)\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \end{aligned}$$

Also, we have

$$M_z = \begin{pmatrix} 1 & 0 \\ 0 & L^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix}$$

and define

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \tilde{\Gamma} \begin{pmatrix} 1 & 0 \\ 0 & L \end{pmatrix} \text{ so that } \tilde{\Gamma} = \begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix} \Gamma \begin{pmatrix} 1 & 0 \\ 0 & L^{-1} \end{pmatrix}. \tag{34}$$

Then

$$\Gamma M_z = \begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \tilde{\Gamma} \begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix}$$

and

$$\begin{pmatrix} 1 & 0 \\ 0 & F' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & L'\tilde{F}'(L^{-1})' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{F}' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix}.$$

We get

$$\begin{aligned} & (\Gamma M_z \otimes I) \left(\begin{pmatrix} 1 & 0 \\ 0 & F' \end{pmatrix} \otimes A \right) \\ = & \begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \tilde{\Gamma} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{F}' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix} \otimes A \\ = & \left(\begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \otimes I \right) \left(\tilde{\Gamma} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{F}' \end{pmatrix} \otimes A \right) \left(\begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix} \otimes I \right) \end{aligned}$$

Also

$$\begin{aligned} (\Gamma M_z \otimes I) \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \tilde{\Gamma} \begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix} \otimes I \\ &= \left(\begin{pmatrix} 1 & 0 \\ 0 & L' \end{pmatrix} \otimes I \right) (\tilde{\Gamma} \otimes I) \left(\begin{pmatrix} 1 & 0 \\ 0 & (L^{-1})' \end{pmatrix} \otimes I \right). \end{aligned}$$

The last two expressions show that the matrix (23) is similar to the matrix (25), i.e.

$$(\tilde{\Gamma} \otimes I) \begin{pmatrix} I \otimes A - I & 0 \\ 0 & \tilde{F}' \otimes A - I \end{pmatrix}.$$

Therefore, stability of (23) is equivalent to (25) when $\tilde{\Gamma}$ is specified as in (34). To complete the proof we simply note that if the matrix (26) is H-stable, then (25) and hence (23) are stable matrices.

Proof of the Corollary to Proposition 4: Clearly $I \otimes A - I$ is negative quasi-definite. Since $A - I$ is negative quasi-definite, $(A + A')/2 - I$ has negative eigenvalues and $(A + A')/2$ has roots less than one. It follows that $\tilde{F} \otimes ((A + A')/2)$ has roots less than one and thus

$$\begin{aligned} \tilde{F} \otimes (A + A') - 2I &= (\tilde{F} \otimes A - I) + (\tilde{F} \otimes A' - I) \\ &= (\tilde{F} \otimes A - I) + (\tilde{F} \otimes A - I)' \end{aligned}$$

has negative roots, i.e. $\tilde{F} \otimes A - I$ is negative quasi-definite. Thus the matrix (26) is negative quasi-definite and hence H-stable. The result follows.

Proof of Proposition 5: We start with the differential equation (21). This can be written explicitly as

$$\frac{d\varphi'}{d\tau} = (\alpha + \tilde{\Gamma}(A - I)a, (AbF - b + B)M_w\hat{\Gamma}),$$

where $\Gamma = \text{diag}(\tilde{\Gamma}, \hat{\Gamma})$ and $\tilde{\Gamma} > 0$ is a scalar. The first equation gives as usual the condition that $A - I$ should be a stable matrix. For the b components, we write the differential equation associated with GSG learning as

$$\dot{b} = (AbF - b)M_w\hat{\Gamma},$$

where we have dropped the inessential constant term involving B. Under our assumptions $M_w\hat{\Gamma}$ is diagonal. We first show that E-stability implies GSG-stability. Next, we write the coefficient matrix A in real Jordan canonical form: $A = S\Lambda S^{-1}$, where Λ is an upper block triangular matrix. The diagonal blocks are either 1×1 blocks, consisting of real eigenvalues, or 2×2 blocks of the form $\begin{pmatrix} \mu & -\nu \\ \nu & \mu \end{pmatrix}$ for nonreal eigenvalues of the form $\mu \pm \nu i$. Multiplying we get

$$S^{-1}\dot{b} = (\Lambda S^{-1}bF - S^{-1}b)M_w\hat{\Gamma},$$

which, defining $q = S^{-1}b$, is

$$\dot{q} = (\Lambda qF - q)M_w\hat{\Gamma}.$$

We then vectorize to get

$$\begin{aligned} vec(\dot{q}) &= (\hat{\Gamma}M_wF \otimes \Lambda - \hat{\Gamma}M_w \otimes I)vec(q) \\ &= (\hat{\Gamma}M_w \otimes I)(F \otimes \Lambda - I)vec(q). \end{aligned}$$

Now the matrix $F \otimes \Lambda - I$ is block diagonal, i.e.

$$\begin{pmatrix} f_1\Lambda - I & 0 & \cdots & 0 \\ 0 & f_2\Lambda - I & \cdots & 0 \\ \vdots \cdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_K\Lambda - I \end{pmatrix},$$

where moreover $f_i\Lambda - I$ are upper triangular matrices. Also $\hat{\Gamma}M_w \otimes I$ is a diagonal matrix, so that we get

$$(\hat{\Gamma}M_w \otimes I)(F \otimes \Lambda - I) = \begin{pmatrix} \hat{m}_1(f_1\Lambda - I) & 0 & \cdots & 0 \\ 0 & \hat{m}_2(f_2\Lambda - I) & \cdots & 0 \\ \vdots \cdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{m}_K(f_K\Lambda - I) \end{pmatrix},$$

where $\hat{m}_i$ is the i-th diagonal element of $\hat{\Gamma}M_w$. Each of the matrices $\hat{m}_k(f_k\Lambda - I)$ is upper block triangular with either diagonal elements of the form

$$\hat{m}_k(f_k\lambda_i - 1),$$

where $\hat{m}_k > 0$ and where $f_k\lambda_i - 1$ is negative by E-stability, or 2×2 blocks of the form

$$\hat{m}_k \left(f_k \begin{pmatrix} \mu & -\nu \\ \nu & \mu \end{pmatrix} - I_2 \right),$$

which again has eigenvalues with negative real parts by E-stability.

To prove that GSG-stability implies E-stability, we note that $\hat{m}_k(f_k\lambda_i - 1) < 0$ for real eigenvalues of A and negativity of eigenvalues of $\hat{m}_k \left(f_k \begin{pmatrix} \mu & -\nu \\ \nu & \mu \end{pmatrix} - I_2 \right)$ for a complex pair of eigenvalues of A clearly also imply $f_k\lambda_i - 1 < 0$ and negativity of eigenvalues of $f_k \begin{pmatrix} \mu & -\nu \\ \nu & \mu \end{pmatrix} - I_2$, respectively as $\hat{m}_k > 0$ for all k. The latter are just the E-stability conditions.

Proof of Proposition 6: It is sufficient to show that the eigenvalues of $\hat{\Gamma}(A\tilde{F}' - I)$ have negative real parts. We use results on the field of values of matrices given in Horn and Johnson (1991). Let $\mathcal{F}(N)$ denote the field of values of a matrix N, which is defined as $\mathcal{F}(N) = \{z^*Nz \mid z \in C^n \text{ with } |z| = 1\}$. Here z^* denotes the complex conjugate of z'. By assumption $\tilde{F}'$ is a "contraction" in the sense used by Horn and Johnson. By p. 155 of Horn and Johnson (1991) we can write $\tilde{F}'$ as a finite sum $\tilde{F}' = \sum c_i U_i$ where $0 < c_i < 1$, with $\sum c_i = 1$, and where U_i are unitary matrices. Since unitary matrices are normal, $\mathcal{F}(U_i)$ is equal to the convex hull of the spectrum of U_i, which we denote $\sigma(U_i)$, see p. 11 of Horn and Johnson (1991). Thus $\mathcal{F}(U_i)$ is a subset of the unit disk since the eigenvalues of U_i lie exactly on the unit circle (see p. 71 of Horn and Johnson (1985)). By the properties of fields of values given on pp. 9-10 of Horn and Johnson (1991) we have

$$\mathcal{F}(\tilde{F}') \subset \sum c_i \mathcal{F}(U_i) \subset \text{unit disk.}$$

Hence $\mathcal{F}(A\tilde{F}' - I) = A\mathcal{F}(\tilde{F}') - 1$ lies in the left half-plane of the complex plane. Next we note that $\mathcal{F}(\hat{\Gamma})$ is a subset of the positive reals since $\hat{\Gamma}$ is symmetric positive definite. (This can be verified by direct computation). Finally, we use the result that $\sigma(CD) \subset \mathcal{F}(C)\mathcal{F}(D)$ if D is positive semidefinite (p. 67 of Horn and Johnson (1991)). Thus $\sigma(\hat{\Gamma}(A\tilde{F}' - I))$ lies in the negative half-plane.

Proof of Proposition 7: (i) Given the positive definite matrix Ω, Lemma of Section 2.1 shows that for $\gamma > 0$ sufficiently small, and hence for V sufficiently small, the difference equation (7) has a unique positive definite fixed point P and P is locally stable. It follows that the evolution of φ_t can be approximated by (1). Finally, as noted in Section 4, local stability of $\bar{\varphi}$ under (1) is determined by (23).

(ii) Under the transformation $\tilde{z}_t = Dz_t$ the model being estimated becomes

$$\begin{aligned} y_t &= \tilde{\beta}'_{t-1}\tilde{z}_t + \eta_t \\ \tilde{\beta}_t &= \tilde{\beta}_{t-1} + \tilde{\Lambda}_t, \end{aligned}$$

where $\tilde{\beta}_t = D^{-1}\beta_t$ and $\tilde{\Lambda}_t = D^{-1}\Lambda_t$ with $\text{cov}(\tilde{\Lambda}_t) = D^{-1}VD^{-1}$. The corresponding estimator of $\tilde{\beta}_t$ is given by

$$\begin{aligned} \tilde{\varphi}_{t+1} &= \tilde{\varphi}_t + \tilde{P}_t\tilde{z}_t(y_t - \tilde{\varphi}'_t\tilde{z}_t) \\ \tilde{P}_{t+1} &= \tilde{P}_t - \tilde{P}_t M_{\tilde{z}}\tilde{P}_t + \sigma^{-2}\tilde{V}, \end{aligned}$$

where $M_{\tilde{z}} = \lim_{t\to\infty} E\tilde{z}_t\tilde{z}'_t = DM_zD$ and $\tilde{V} = D^{-1}VD^{-1}$. The initial priors will also be related by $\tilde{\varphi}_0 = D^{-1}\varphi_0$ and $\tilde{P}_0 = D^{-1}P_0D^{-1}$. It is easily seen that the $\tilde{\varphi}_t, \tilde{P}_t$ system is equivalent to (6)-(7) with $\varphi_t = D\tilde{\varphi}_t$ and $P_t = D\tilde{P}_tD$.

Proof of Corollary to Proposition 7: By (ii) of Proposition 7, the Bayesian learning rule is invariant to a transformation of variables. Letting $\tilde{w}_t = Lw_t$ we have $\tilde{w}_t = \tilde{F}\tilde{w}_{t-1}$ and $M_{\tilde{z}} = I$. Since (27) is H-stable then (23) is stable for all Γ and hence for all V. Stability of the REE then follows by (i) of Proposition 7.

Proof of Proposition 8: (Outline) We consider the algorithm (4)-(5). Since $V = \gamma^2\sigma^2\Omega$, equation (5) can be written as

$$P_{t+1} = P_t - \frac{P_t z_t z'_t P_t}{1 + z'_t P_t z_t} + \gamma^2\Omega.$$

Defining $\tilde{P}_t = \gamma^{-1}P_t$, we get the constant-gain algorithm

$$\begin{aligned} \varphi_{t+1} &= \varphi_t + \gamma\left[\frac{\tilde{P}_t}{1 + \gamma z'_t\tilde{P}_t z_t} z_t(y_t - \varphi'_t z_t)\right] \\ \tilde{P}_{t+1} &= \tilde{P}_t + \gamma\left[\Omega - \frac{\tilde{P}_t z_t z'_t \tilde{P}_t}{1 + \gamma z'_t\tilde{P}_t z_t}\right], \end{aligned}$$

where under learning $y_t = T(\varphi_t)'z_t + \eta_t$. Using stochastic approximation techniques we compute the associated differential equation

$$\frac{d\varphi}{d\tau} = h_\varphi(\varphi, \tilde{P}), \tag{35}$$

$$\frac{d\tilde{P}}{d\tau} = h_{\tilde{P}}(\varphi, \tilde{P}), \tag{36}$$

where using Lebesgue's dominated convergence theorem

$$\begin{aligned} h_\varphi(\varphi, \tilde{P}) &= \lim_{\gamma\to 0} E\left[\frac{\tilde{P}}{1 + \gamma z'_t\tilde{P}z_t}(z_t z'_t(T(\varphi) - \varphi) + z_t\eta_t)\right] \\ &= E\left[\tilde{P}(z_t z'_t(T(\varphi) - \varphi) + z_t\eta_t)\right] \\ &= \tilde{P}M_z(T(\varphi) - \varphi) \end{aligned}$$

and

$$\begin{aligned} h_{\tilde{P}}(\varphi, \tilde{P}) &= \lim_{\gamma \to 0} E\left[\Omega - \frac{\tilde{P} z_t z_t' \tilde{P}}{1 + \gamma z_t' \tilde{P} z_t}\right] \\ &= \Omega - \tilde{P} M_z \tilde{P}. \end{aligned}$$

The system (35)-(36) has a fixed point at $(\bar{\varphi}, \Gamma)$, where Γ is a positive definite matrix as in equation (9) and $T(\bar{\varphi}) = \bar{\varphi}$. The system is block recursive and clearly Γ is locally stable under (36). Evaluating the linearization of (35) at the fixed point, we get the stability condition the eigenvalues of the matrix $\Gamma M_z(DT - I)$ should have negative real parts.

B Numerical Examples

Here we provide numerical examples that justify the Remark at the end of Section 3.

Example 1: (E-stability does not imply SG-stability) For $n = 1$ and $k = 2$ select

$$A = -1.8800,\ F = \begin{pmatrix} -0.9390 & -0.6979 \\ 0.8722 & 0.0828 \end{pmatrix},\ \Sigma_e = \begin{pmatrix} 1.0520 & -0.5164 \\ -0.5164 & 0.2581 \end{pmatrix}.$$

These yield

$$(M_w \otimes I)((F' \otimes A) - I) = \begin{pmatrix} -0.2503 & -1.3829 \\ 0.9012 & 0.3256 \end{pmatrix}.$$

This solution is E-stable but it is not convergent under SG learning since the eigenvalues of $(M_w \otimes I)((F' \otimes A) - I)$ are $0.0377 \pm 1.7086i$.

Example 2: (SG-stability does not imply E-stability) For $n = 1$ and $k = 2$ select

$$A = -1.9022,\ F = \begin{pmatrix} -1.1281 & 0.7252 \\ -0.4944 & 0.0117 \end{pmatrix},\ \Sigma_e = \begin{pmatrix} 0.5361 & 0.5760 \\ 0.5760 & 1.1807 \end{pmatrix}.$$

These yield eigenvalues $-0.0838 \pm 0.3493i$ for $(M_w \otimes I)((F' \otimes A) - I)$ and eigenvalues $0.0618 \pm 0.3493i$ for $F' \otimes A - I$.

To find the counterexamples we simply conducted a random search over A, F and Σ_e under the required constraints. The Matlab routine is available on request.

C Scaling Invariance

One of the reasons for our investigation of GSG algorithms is that the classic SG algorithm suffers from a disadvantage relative to RLS, which has not received attention. The SG algorithm is not scale invariant, and thus the resulting estimates are affected by the choice of units. This is demonstrated as follows.

For simplicity, we consider a univariate case. Suppose we are estimating the regression model:

$$y_t = \beta' z_t + \eta_t$$

by least squares, where β and z_t are $p \times 1$ column vectors. Our discussion here will initially be in terms of the standard (non-self-referential) regression model, since the point holds generally, but it also applies to the model (17) with learning. The RLS estimate using data through $t-1$ is given by

$$\begin{aligned} \varphi_t &= \varphi_{t-1} + \gamma_t R_t^{-1} z_{t-1}(y_{t-1} - b'_{t-1} z_{t-1})' \\ R_t &= R_{t-1} + \gamma_t (z_{t-1} z'_{t-1} - R_{t-1}), \end{aligned} \tag{37}$$

where the standard decreasing gain assumption is that $\gamma_t = 1/t$. Suppose we now change units of the regressors so that $\tilde{z}_t = D z_t$, where $D = \mathrm{diag}(k_1, \ldots, k_p)$. Here diag denotes a diagonal matrix and we assume $k_i > 0$. Then:

$$y_t = \tilde{\beta}' \tilde{z}_t + \eta_t,$$

where $\tilde{\beta} = D^{-1}\beta$. Let $\tilde{\varphi}_t$ be the RLS estimate of $\tilde{\beta}$ based on a regression of y_t on $\tilde{z}_t$. Then RLS is *scale invariant* in the sense that:

$$\tilde{\varphi}_t = D^{-1}\varphi_t.$$

To see this, pre-multiply the RLS equation for φ_t by D^{-1} and for the R_t equation, premultiply by D and postmultiply by $D' = D$. Defining $\tilde{R}_t = D R_t D'$ we get:

$$\begin{aligned} \tilde{\varphi}_t &= \tilde{\varphi}_{t-1} + \gamma_t \tilde{R}_t^{-1} \tilde{z}_{t-1}(y_{t-1} - \tilde{\varphi}'_{t-1} \tilde{z}_{t-1}) \\ \tilde{R}_t &= \tilde{R}_{t-1} + \gamma_t (\tilde{z}_{t-1} \tilde{z}'_{t-1} - \tilde{R}_{t-1}). \end{aligned}$$

But this is exactly RLS applied to a regression of y_t on $\tilde{z}_t$.

In contrast, SG estimation is not scale invariant. The classic SG algorithm for a regression of y_t on z_t is:

$$\varphi_t = \varphi_{t-1} + \gamma_t z_{t-1}(y_{t-1} - \varphi'_{t-1} z_{t-1}).$$

Multiplying through by D^{-1} we get that $\tilde{\varphi}_t = D^{-1}\varphi_t$ satisfies:

$$\tilde{\varphi}_t = \tilde{\varphi}_{t-1} + \gamma_t D^{-2} \tilde{z}_{t-1}(y_{t-1} - \tilde{\varphi}'_{t-1} \tilde{z}_{t-1}). \tag{38}$$

But SG estimation based on a regression of y_t on $\tilde{z}_t$ is instead:

$$\hat{\varphi}_t = \hat{\varphi}_{t-1} + \gamma_t \tilde{z}_{t-1}(y_{t-1} - \hat{\varphi}'_{t-1} \tilde{z}_{t-1}),$$

and clearly $\tilde{\varphi}_t \neq \hat{\varphi}_t$.

Note that the same argument applies to transformations of variable $\tilde{z}_t = D z_t$ for D positive definite: RLS is invariant to such transformations while SG is not.

References

Anderson, E. W. (2005): "The Dynamics of Risk-Sensitive Allocations," *Journal of Economic Theory*, 125, 93–150.

Arrow, K. J. (1974): "Stability Independent of Adjustment Speed," in Horwich and Samuelson (1974), pp. 181–201.

Arrow, K. J., and M. McManus (1958): "A Note on Dynamic Stability," *Econometrica*, 26, 448–454.

Barucci, E., and L. Landi (1997): "Least Mean Squares Learning in Self-Referential Stochastic Models," *Economics Letters*, 57, 313–317.

Benveniste, A., M. Metivier, and P. Priouret (1990): *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin.

Bullard, J., and K. Mitra (2002): "Learning About Monetary Policy Rules," *Journal of Monetary Economics*, 49, 1105–1129.

Carlson, D. (1968): "A New Criterion for H-Stability of Complex Matrices," *Linear Algebra and Applications*, 1, 59–64.

Cho, I.-K., N. Williams, and T. J. Sargent (2002): "Escaping Nash Inflation," *Review of Economic Studies*, 69, 1–40.

Clarida, R., J. Gali, and M. Gertler (1999): "The Science of Monetary Policy: A New Keynesian Perspective," *Journal of Economic Literature*, 37, 1661–1707.

Cogley, T., and T. J. Sargent (2005): "The Conquest of US Inflation: Learning and Robustness to Model Uncertainty," *Review of Economic Dynamics*, 8, 528–563.

Epstein, L. G., and S. E. Zin (1989): "Substitution, Risk Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica*, 57, 937–969.

Evans, G. W., and S. Honkapohja (1993): "Adaptive Forecasts, Hysteresis and Endogenous Fluctuations," *Federal Reserve Bank of San Francisco Economic Review*, 1993(1), 3–13.

——— (1998): "Stochastic Gradient Learning in the Cobweb Model," *Economics Letters*, 61, 333–337.

——— (2001): *Learning and Expectations in Macroeconomics*. Princeton University Press, Princeton, New Jersey.

——— (2003a): "Adaptive Learning and Monetary Policy Design," *Journal of Money, Credit and Banking*, 35, 1045–1072.

——— (2003b): "Expectations and the Stability Problem for Optimal Monetary Policies," *Review of Economic Studies*, 70, 807–824.

Evans, G. W., S. Honkapohja, and N. Williams (2006): "Generalized Stochastic Gradient Learning," mimeo.

Evans, G. W., and G. Ramey (2006): "Adaptive Expectations, Underparameterization and the Lucas Critique," *Journal of Monetary Economics*, 53, 249–264.

Giannitsarou, C. (2005): "E-Stability Does Not Imply Learnability," *Macroeconomic Dynamics*, 9, 276–287.

Glover, K., and J. Doyle (1988): "State-Space Formulae for All Stabilizing Controllers that Satisfy an H-infinity Norm Bound and Relations to Risk-Sensitivity," *Systems & Control Letters*, 11, 167–172.

Hamilton, J. D. (1994): *Time Series Analysis.* Princeton University Press, Princeton, NJ.

Hansen, L. P., and T. J. Sargent (2007): *Robustness.* Princeton University Press, Princeton, NJ.

Hassibi, B., A. H. Sayed, and T. Kailath (1996): "H-infinity Optimality of the LMS Algorithm," *IEEE Transactions on Signal Processing*, 44, 267–280.

Heinemann, M. (2000): "Convergence of Adaptive Learning and Expectational Stability: The Case of Multiple Rational Expectations Equilibria," *Macroeconomic Dynamics*, 4, 263–288.

Honkapohja, S., and K. Mitra (2006): "Learning Stability in Economies with Heterogeneous Agents," *Review of Economic Dynamics*, 9, 284–309.

Horn, R., and C. Johnson (1985): *Matrix Analysis.* Cambridge University Press, Cambridge.

——— (1991): *Topics in Matrix Analysis.* Cambridge University Press, Cambridge.

Horwich, G., and P. A. Samuelson (eds.) (1974): *Trade, Stability and Macroeconomics.* Academic Press, New York.

Jacobson, D. (1973): "Optimal Stochastic Linear Systems with Exponential Performance Criteria and Their Relation to Deterministic Games," *IEEE Transactions on Automatic Control*, AC-18, 124–131.

Johnson, C. R. (1974): "Sufficient Conditions for D-Stability," *Journal of Economic Theory*, 9, 53–62.

Kwakernaak, H., and R. Sivan (1972): *Linear Optimal Control Systems.* Wiley-Interscience, New York.

Magnus, J., and H. Neudecker (1988): *Matrix Differential Calculus.* Wiley, New York.

Marcet, A., and J. P. Nicolini (2003): "Recurrent Hyperinflations and Learning," *American Economic Review*, 93, 1476–1498.

Primiceri, G. E. (2006): "Why Inflation Rose and Fell: Policy-Makers' Beliefs and U. S. Postwar Stabilization Policy," *Quarterly Journal of Economics*, 121, 867–901.

Rotemberg, J. J., and M. Woodford (1997): "An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy," *NBER Macroeconomics Annual*, 12, 297–346.

Sargent, T. J. (1999): *The Conquest of American Inflation.* Princeton University Press, Princeton NJ.

Sargent, T. J., and N. Williams (2005): "Impacts of Priors on Convergence and Escapes from Nash Inflation," *Review of Economic Dynamics*, 8, 360–391.

Sargent, T. J., N. Williams, and T. Zha (2006): "Shocks and Government Beliefs: The Rise and Fall of American Inflation," *American Economic Review*, 96, 1193–1224.

Sims, C., and T. Zha (2006): "Were There Regime Switches in US Monetary Policy?," *American Economic Review*, 96, 54–81.

Svensson, L. E. (2003): "What is Wrong with Taylor Rules? Using Judgement in Monetary Policy through Targeting Rules," *Journal of Economic Literature*, 41, 426–477.

Tallarini, T. D. (2000): "Risk-Sensitive Real Business Cycles," *Journal of Monetary Economics*, 45, 507–32.

Whittle, P. (1990): *Risk Sensitive Optimal Control.* Wiley, New York.

Woodford, M. (2003): *Interest and Prices: Foundations of a Theory of Monetary Policy.* Princeton University Press, Princeton, NJ.