A probabilistic interpretation of the constant gain learning algorithm

Michele Berardi

The University of Manchester, Manchester, England

Abstract
This paper proposes a novel interpretation of the constant gain learning algorithm through a probabilistic setting with Bayesian updating. The underlying process for the variable being estimated is not specified a priori through a parametric model, and only its probabilistic structure is defined. Such framework allows to understand the gain coefficient in the learning algorithm in terms of the probability of changes in the estimated variable. On the basis of this framework, I assess the range of values commonly used in the macroeconomic empirical literature in terms of the implied probabilities of changes in the estimated variables.

KEYWORDS
adaptive learning, Bayesian learning, constant gain

JEL CLASSIFICATION
C63, D83, D84, D90, E70

1 | INTRODUCTION

Adaptive learning has been used extensively in the macroeconomic literature to depart from the assumption of rationality. An extensive treatise of adaptive learning in macroeconomics is provided by Evans and Honkapohja (2001). In this line of work, learning algorithms are used to model how agents recursively estimate relationships between variables of interest, i.e., parameters in economic models. An important element of any learning algorithm is its gain parameter, which
determines the weighting structure on past observations. In this regard, learning algorithms can be divided into two classes: decreasing gain (DG) and constant gain (CG) algorithms.

The most common instance of a DG algorithm is the recursive least squares, where the gain coefficient is set equal to $1/t$, $t$ being the time period of the estimate, and all observations are thus weighted equally: this is suitable for estimating quantities that are believed to be constant over time. With a CG algorithm, instead, more recent observations receive a higher weight, and the weights decrease geometrically with time: this is usually employed when the estimated parameters are believed to change over time, as it allows for better tracking.

While the value of the gain coefficient in the DG case is obvious and has a clear interpretation, much less straightforward is the interpretation of the gain in the CG case, and there seems to be no agreement in the literature on an appropriate value for it. The aim of this paper is to provide a novel interpretation of the CG coefficient and thus help researchers in need to select a value for such parameter.

A growing literature in applied macroeconomics has used CG learning to explain a range of features, from the rise and fall of U.S. inflation in the 70s and 80s (in particular, the seminal works of Sargent (1999) and Sargent, Williams, & Zha (2006)) to the causes of business cycles (e.g., Milani (2011) and Eusepi & Preston (2011)). Though there is no direct evidence of the appropriate value for the gain parameter, Berardi and Galimberti (2017) provide a thorough discussion of the role and estimate bands for the gain parameter in macroeconomic applications. In general, higher gains imply faster reaction to changes, but more volatile estimates.

The CG algorithm is a reduced form learning model, which could be derived as an optimal solution of inference in a number of underlying frameworks. For example, Muth (1960) has shown how adaptive expectations can be optimal under certain assumptions about the structure of the variable being forecasted. A CG algorithm for estimating the (mean) value of a variable, in fact, implements adaptive expectations, and as such it provides optimal forecasts under conditions specified in Muth (1960). Those conditions are quite restrictive on the underlying process for the variable being forecasted, which must be representable as an infinite sum of current and past exogenous disturbances, with appropriate weights related to the gain parameter.

A CG algorithm can also be obtained through a Kalman filter model, which implements Bayesian updating in a state-space framework, with appropriate initial conditions. It is well known that with a time-invariant state-space model, the Kalman gain converges to a constant: choosing such constant as initial value for the gain, the Kalman filter gives rise to a CG algorithm. The natural interpretation of such gain coefficient is usually in terms of the variances of disturbances in the measurement and transition equations. An analysis of the differences between CG least squares and the Kalman filter is provided in Sargent and Williams (2005), who highlight the effect of different priors on the convergence properties of Bayesian learning. Similarly, Evans, Honkapohja, and Williams (2010) propose a constant gain generalized stochastic gradient algorithm, which can be viewed as an approximate Bayesian learning scheme when agents allow for parameter drift in their beliefs.

I propose here instead an interpretation of the CG learning algorithm through a probabilistic setting where Bayesian learners estimate recursively the value of an unobservable variable through a signal. The underlying process for the variable being estimated is not specified a priori through a parametric model, and only its probabilistic structure is defined. This framework allows for a novel interpretation of the gain coefficient in terms of the probability of changes in the estimated variables. I then assess the values of various gain coefficients used in empirical studies against this background, deriving some implications on the implied frequency of changes in the estimated variables.
In general, Bayesian learning in a state-space model requires knowledge (or estimation) of the variance covariance structure of state and measurement noise. Adaptive learning relies instead usually on much less information, though in the case of a constant gain, it introduces a free parameter to be determined. Such determination can be arbitrary, and hence likely to be inefficient in terms of the implied prediction error, or based on some additional information on the dynamics of the system. My proposed interpretation of the gain parameter allows one to understand (or select) its value on the basis of the probability of changes in the estimated variables: such probability thus represents the additional information required to pin down an otherwise arbitrary parameter in an adaptive learning scheme.

2 | CONSTANT GAIN ALGORITHM AND BAYESIAN LEARNING

Recursive learning algorithms can represent optimal learning behaviour under certain assumptions about the underlying quantities being learned and the observations available. The simplest example is the case where a constant needs to be estimated, and the series of observations include i.i.d. noise with constant variance. In such case, a decreasing gain algorithm with gain equal to $1/t$ is optimal, as it provides the least squares estimate of the sample mean. If the underlying variable to be estimated is instead time-varying, the literature suggests that, in general, a constant gain algorithm should be used, as it puts more weight on more recent observations and thus allows for better tracking.

More sophisticated approaches to the determination of the gain coefficient have also been proposed in the literature, endogenizing the gain to the volatility of the system. For example, Marcet and Nicolini (2003) in a contest of hyperinflationary threats, propose a learning algorithm where the gain coefficient switches from decreasing to constant (and vice versa) depending on the level of instability detected in the data: as instability increases, agents switch from a decreasing to a constant gain learning mechanism, in order to enhance tracking relative to smoothing.

In the same hyperinflationary framework, Kostyshyna (2012) proposes an adaptive step-size algorithm which allows the gain to evolve in response to changes in the environment and results in an increasing gain during hyperinflationary episodes and a decreasing gain after the hyperinflation ends.

In Milani (2014) the gain coefficient is also endogenous and is adjusted according to past forecast errors: when the average of the past forecast errors is below a certain threshold, agents use a decreasing gain, while if average past errors is above the threshold, agents use instead a constant gain, for fear that the economy may be experiencing a structural break.

2.1 | Constant gain algorithm

Suppose agents need to estimate the (time-varying) mean of a random variable $x_t$ over time. Denoting $\bar{x}_t$ such estimate, the CG algorithm takes the form

$$\bar{x}_t = \bar{x}_{t-1} + g(x_t - \bar{x}_{t-1}) = (1 - g)\bar{x}_{t-1} + gx_t$$

(1)

$$= g \sum_{j=2}^{t} (1 - g)^{t-j} x_j + (1 - g)^{t-1} x_1,$$

(2)
where I have used the assumption \( \bar{x}_1 = x_1 \), since no previous information is available to agents. The constant gain \( g \) determines the weight put on past observations, as

\[
b'_1 = (1 - g)^{t-1}
\]

for \( j = 2, \ldots, t \), where \( b'_j \) denotes the weight put at time \( t \) on time \( j \leq t \) observation. Since \( g < 1 \), it can be seen that weights on past observations decrease exponentially with time.

The same relationship between gains and weights holds also for a multivariate model where agents estimate a vector of time-varying coefficients through a linear regression model. See Berardi and Galimberti (2013).

### 2.2 A probabilistic Bayesian learning framework

Consider now a framework where agents are interested in estimating the value of an unobservable variable \( \theta_t, t \geq 1 \). Nature draws the value \( \theta_t \) at some time \( t = 0 \) from an improper uniform distribution over \( \mathbb{R} \). Consistently, agents have a flat (uninformative) prior on its value at time \( t = 1 \). Nature can also re-draw, with some fixed and known probability \( 0 \leq \pi \leq 1 \), a new value for the variable, again from an improper distribution over \( \mathbb{R} \), at the beginning of each period \( t > 1 \). At every period \( t \geq 1 \) agents receive a signal \( x_t \) on the value of \( \theta_t \), in the form

\[
x_t = \theta_t + v_t,
\]

where \( v_t \) is an i.i.d. random variable, normally distributed with zero mean and constant variance \( \sigma_v^2 \).

I first define

\[
\bar{x}^f_j = \frac{1}{t - j + 1} \sum_{z=j}^{t} x_z, \quad \text{for } j \leq t
\]

the best estimate of \( \theta_t \) at time \( t \) if Nature had last re-drawn at (the beginning of) time \( j \leq t \). This is simply the mean of the sample of relevant observations for the signal, since the last change in \( \theta_t \) took place.

Given that agents don’t know if and when a change in the estimated variable took place, their posterior mean for it at time \( t \) is then given by

\[
\bar{x}_t = \sum_{j=1}^{t} a'_j \bar{x}^f_j
\]

where

\[
a'_j = (1 - \pi)^{t-1}
\]
\[ a^j_t = (1 - \pi)^{-j} \pi, t \geq j > 1, \]  

(8)

with

\[ \sum_{j=1}^{t} a^j_t = 1. \]

The coefficients \( a^j_t \) capture the probability that each (truncated) series \( \bar{x}^j_t \) is the appropriate one for computing the conditional expected value of \( \theta_t \), that is, the probability that Nature re-drew at the beginning of time \( j \) and never after. Clearly, (7)–(8) are the same as (3)–(4) for \( \pi = g \).

Substituting (7)–(8) into (6) and collecting terms for each \( x_j \), it is then possible to rewrite the posterior mean \( \bar{x}_t \) as a weighted sum of current and past values of \( x_j \) as

\[ \bar{x}_t = \sum_{j=1}^{t} h^j_t x_j, \]

(9)

where

\[ h^1_t = \frac{(1 - \pi)^{t-1}}{t}, \]

(10)

\[ h^j_t = h^{j-1}_t + \frac{(1 - \pi)^{t-j} \pi}{t - j + 1}, t \geq j > 1. \]

(11)

It can be shown that \( \sum_{j=1}^{t} h^j_t = 1 \). Clearly if \( \pi = 1 \) (\( \theta_t \) changes for sure every period), \( h^j_t = 0 \) for \( j < t \) and \( h^j_t = 1 \) for \( j = t \): only the last observation matters. If instead \( \pi = 0 \) (\( \theta_t \) constant) then all observations receive the same weight \( 1/t \). This last case gives rise to a decreasing gain algorithm, implementing recursive least squares (equivalent to stochastic gradient in this case)

\[ \bar{x}_t = \bar{x}_{t-1} + \frac{1}{t} (x_t - \bar{x}_{t-1}) \]

with \( \bar{x}_0 = 0 \) (that is, \( \bar{x}_1 = x_1 \)), or, in non-recursive form,

\[ \bar{x}_t = \frac{1}{t} \sum_{j=1}^{t} x_j, \]

(12)

which is simply the sample mean.

To better understand the weighting structure defined by (10)–(11), I propose Figure (1). Observation \( x_1 \) is relevant for inference about the current value of \( \theta_t \) only if Nature never re-drew over the whole sample period from 1 to \( t \), which happened with probability \( (1 - \pi)^{t-1} \): in such case each observation in that sample should be weighted equally, with weight \( 1/t \). Observation \( x_2 \) is relevant if Nature never re-drew (again, with weight \( 1/t \)), which happened with probability \( (1 - \pi)^{t-1} \), or
if it re-drew at the beginning of period 2 and never after (and in this case, with weight \( \frac{1}{t-1} \)), which happened with probability \((1 - \pi)^{t-2}\pi\). And so on.

### 2.2.1 Flat prior and independent draws

In the proposed framework, Nature can re-draw the variable of interest at every period from an improper distribution over \(\mathbb{R}\), with some fixed probability \(0 \leq \pi \leq 1\). This means that Nature can draw any value for \(\theta_t\), from an unbounded set. This assumption justifies the flat prior on the new draw, and allows for the weighting structure proposed above. Having a model-free evolution of the variable of interest is crucial for the proposed framework, and contrasts it with alternative settings where the estimated variable changes according to a known dynamic model. For example, if agents knew that \(\theta_t\) was to follow an AR(1) process with known autoregressive coefficient (say, a random walk), with Nature drawing each period the i.i.d. innovation, then a Kalman filter would be the optimal algorithm to employ for tracking and estimating the value of \(\theta_t\). It is the model-free assumption that allows our new interpretation of the weights in terms only of \(\pi\) (and time) and not of other elements of the assumed dynamic model for \(\theta_t\).

From an applied perspective, this modelling choice can be justified in particular when there is no a priori knowledge of what the changes in \(\theta_t\) could look like, and when there is no reason to believe that the current value of the estimated variable should give any insight into the possible new value. These are of course extreme assumptions, but provide a sharp characterization of the weighting structure on past observations and allow for its interpretation purely in terms of probabilities.

### 3 A COMPARISON

In light of the proposed framework, it is instructive to analyse the relationship between the adaptive learning gain \(g\) and the probability \(\pi\) in the proposed Bayesian learning model. The gain parameter in an adaptive learning algorithm determines the weight put on past observations: with a decreasing gain \(1/t\), all observations receive equal weight; with a constant gain \(g\), instead, the weight decays exponentially with past observations. A similar interpretation can be given to \(\pi\), which represents the probability of a change in the variable \(\theta_t\), happening at each time \(t\): this determines the probability that each observation from time \(j, 0 < j < t\) is relevant for time \(t\) inference, which, together with the number of observations, determines individual weights.

The weighting structure represented by (10)–(11) cannot be generated exactly by a CG algorithm for finite \(t\). Nevertheless, even if an approximation only, it provides a means to interpret the weighting implied by such algorithm. Clearly, if one sets \(\pi = g\) then \(b^{j}_{i} = a^{j}_{i}\): if the constant
gain is to be interpreted as the probability \( \pi \), the weights put on individual observations through the CG algorithm are the weights put on past truncated series of observations in the probabilistic Bayesian setting. In such setting, weights on individual observations are instead given by (10)–(11), which, in a non-recursive way, can be rewritten as

\[
\begin{align*}
h^t_i &= \frac{(1 - \pi)^{t-1}}{t} \\
h^t_j &= \frac{(1 - \pi)^{t-1}}{t} + \sum_{m=2}^{j} \frac{(1 - \pi)^{t-m} \pi^m}{t-m+1}.
\end{align*}
\]

While the weighting structure on individual observations in the Bayesian framework is more convoluted than that in the CG algorithm, both \( b^t_j \) and the leading term in \( h^t_j \) (represented by \( \frac{(1 - \pi)^{t-1}}{t} \)) decays exponentially, leading to similar weight profiles on older observations. In fact, for \( \pi = g \), the leading term in \( h^t_j \) is equal to \( \frac{b^t_j}{t-j+1} \).

Figure 2 shows \( b^t_j \) and \( h^t_j \), computed for \( \pi = g = 0.025 \) with \( t = 100 \). Figure 3 then shows the same series, but for \( t = 1,000 \) (the second and third frames zoom in, respectively, on the first and last 100 points of the series). It can be seen that as \( t \) increases, \( b^t_j \) and \( h^t_j \) get closer to each other for small values of \( j \), while for high values of \( j \) (that is, for observations closer to the time of estimation) the difference between the two terms remains largely the same. Weights \( b^t_j, j \geq 1, \) are independent of \( t \) (they depend instead on \( t - j \); that is, \( b^t_j = b^{t+k}_{j+k} \)). The same is not exactly true for \( h^t_j \), though quantitatively it is indeed the case that \( h^t_j \simeq h^{t+k}_{j+k} \). This is due to the fact that the leading \( j-1 \) terms out of the total \( j + k \) terms in \( h^{t+k}_{j+k} \) are the same as the leading \( j-1 \) terms out of the total \( j \) terms in \( h^t_j \) (the \( j^\text{th} \) term differs by \( \pi \)), with the additional \( k \) terms in \( h^{t+k}_{j+k} \) negligible in size. Thus the final end of the \( b \) and \( h \) curves tend to remain at the same distance as \( t \) increases. The two curves, instead, get closer and closer to each other on their initial part as \( t \) increases.

Figure 2  Values of \( b^t_j \) and \( h^t_j \) for \( t = 100 \) [Colour figure can be viewed at wileyonlinelibrary.com]
because both tend to zero (weights on observations farther back in time converge to zero under both weighting structures).

It can be seen that, despite being derived in different frameworks, the shape of the two weighting structures is remarkably similar, leading to similar weighting on past information in the two cases. Figure 4 shows the difference $\Delta^t = b^t - h^t$ (where $b^t$ and $h^t$ are the vectors $\{b^t_j\}_{j=1}^t$ and $\{h^t_j\}_{j=1}^t$) for $t = 1,000$ and $\pi = g = 0.025$.

4 | CONSTANT GAINS IN THE EMPIRICAL LITERATURE

Using the framework developed above, one can interpret the constant gain coefficients that have been found to fit the data well in empirical macroeconomic studies in terms of the implied probability of changes in the estimated variables. Typical values proposed (estimated or calibrated) in the empirical literature for the constant gain range from close to zero to over 0.2, with most
FIGURE 5 Values of $\hat{\pi}(g = 0.025)$
[Colour figure can be viewed at wileyonlinelibrary.com]

Irrespective of the specific value selected for the gain, one can use the framework proposed here in order to compute the implied probability of changes in the estimated variables that corresponds to such coefficient by finding the $\pi$ that, for a given $g$, minimizes the sum of (squared) deviations between the weighting structure implied by the gain and the weighting structure of the Bayesian framework. That is, one can compute

$$\hat{\pi}(g) = \arg\min_{\pi} \Delta(g)'\Delta(g)$$

where the notation for $\hat{\pi}(g)$ and $\Delta(g)$ makes explicit the dependence on both $t$ and $g$. Fixing $g$, one can find the implied probability for a certain gain coefficient as a function of the number of observations. Figure 5 shows such measure for $g = 0.025$. It can be seen that for large enough values of $t$, $\hat{\pi}(g = 0.025)$ stabilizes and becomes constant. One can thus compute the value of $\hat{\pi}(g)$ for large $t$, obtaining a function that gives the implied (asymptotic) probability $\hat{\pi}$ for any value of $g$. In particular, I restrict the range of $g$ between 0.01 and 0.1, which contains most values used in the empirical literature: Figure 6 shows the results.

---

1 A similar value of 0.29 has been found by Dave and Tsang (2014), also in an asset pricing setting.
2 This value is remarkably close to the one found by Milani (2007), who estimates the gain to be 0.0183.
3 I set $t = 1,000$ in the computations.
It can be seen that, for gains between 0.01 and 0.1, the implied probability of changes in the estimated variable(s) each period ranges from 0.31 per cent \((g = 0.01)\) to 3.59 per cent \((g = 0.1)\).

5 | CONCLUSIONS

This paper has proposed a probabilistic Bayesian framework that allows for a novel interpretation of the weighting structure on past observations implied by a CG learning algorithm. By assuming a model-free representation of the evolution of the variable being estimated, one does not need to specify the variance-covariance structure of state and measurement noise in order to derive the optimal Bayesian weight on past information and the gain coefficient can, under this interpretation, be linked only to the probability of changes in the estimated variables.

Using this framework, it is possible to map the gain coefficient used in empirical studies into implied probabilities. In particular, most works in the macroeconomic empirical literature use a gain coefficient between 0.01 and 0.1, which map into per-period probabilities of changes in the estimated variable between 0.31 per cent and 3.59 per cent.

Such new understanding of the role played by the gain parameter in a learning algorithm could be useful for researchers in choosing an appropriate value in empirical studies.

ORCID

Michele Berardi \(\odot\) https://orcid.org/0000-0001-8145-5207

REFERENCES

Benhabib, J., & Dave, C. (2014). Learning, large deviations and rare events. *Review of Economic Dynamics, 17*, 367–382.

Berardi, M., & Galimberti, J. K. (2013). A note on exact correspondences between adaptive learning algorithms and the kalman filter. *Economics Letters, 118*, 139–142.

Berardi, M., & Galimberti, J. K. (2017). Empirical calibration of adaptive learning. *Journal of Economic Behavior and Organization, 144*, 219–237.

Electronic copy available at: https://ssrn.com/abstract=3721354
Dave, C., & Feigenbaum, J. (2018). Precautionary learning and inflationary biases. *Macroeconomic Dynamics*, forthcoming.

Dave, C., Tsang, K. P. (2014). Recursive preferences, learning and large deviations. *Economics Letters*, 124, 329–334.

Evans, G. W., & Honkapohja, S. (2001). *Learning and Expectations in Macroeconomics*. Princeton, NJ: Princeton University Press.

Evans, G. W., Honkapohja, S., & Williams, N. (2010). Generalized stochastic gradient learning. *International Economic Review*, 51, 237–262.

Eusepi, S., & Preston, B. (2011). Expectations, learning, and business cycle fluctuations. *American Economic Review*, 101, 2844–2872.

Kostyshyna, O. (2012). Application of an adaptive step-size algorithm in models of hyperinflation. *Macroeconomic Dynamics*, 16, 355–375.

Malmendier, U., & Nagel, S. (2016). Learning from inflation experiences. *Quarterly Journal of Economics*, 131, 53–87.

Marcet, A., & Nicolini, J. P. (2003). Recurrent hyperinflations and learning. *American Economic Review*, 93, 1476–1498.

Markiewicz, A., & Pick, A. (2014). Adaptive learning and survey data. *Journal of Economic Behavior & Organization*, 107, 685–707.

Milani, F. (2007). Expectations, learning and macroeconomic persistence. *Journal of Monetary Economics*, 54, 2065–2082.

Milani, F. (2011). Expectation shocks and learning as drivers of the business cycle. *The Economic Journal*, 121, 379–401.

Milani, F. (2014). Learning and time-varying macroeconomic volatility. *Journal of Economic Dynamics and Control*, 47, 94–114.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55, 299–306.

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

Sargent, T. J., & Williams, N. (2005). Impacts of priors on convergence and escapes from Nash inflation. *Review of Economic Dynamics*, 8, 360–391.

Sargent, T. J., Williams, N., & Zha, T. (2006). Shocks and government beliefs: the rise and fall of American inflation. *American Economic Review*, 96, 1193–1224.

**How to cite this article:** Berardi M. A probabilistic interpretation of the constant gain learning algorithm. *Bull Econ Res*. 2020;72:393–403. [https://doi.org/10.1111/boer.12256](https://doi.org/10.1111/boer.12256)