

Likelihood Ratio and Score Processes

Contents

- 6.1. Introduction
- 6.2. Multiplicative martingales
- 6.3. Multiple models
- 6.4. Some large sample properties
- 6.5. Log likelihoods
- 6.6. Score processes
- 6.7. Using a multiplicative martingale for model selection

[↓ Download PDF here](#)

Authors: Lars Peter Hansen (University of Chicago) and Thomas J. Sargent (NYU)

6.1. Introduction

In this chapter we study the behavior of likelihood ratio processes and score processes. We treat these as processes so that we can study their dynamic behavior when additional observations are included. We start by showing how what we call multiplicative martingales imply alternative probability models. We then investigate some limiting behavior that reveals which among multiple models generates the data without imposing an *ex ante* distribution (a prior) over the alternative models. We then provide examples of such martingales constructed from likelihood ratio processes for alternative models. We end by studying the behavior of score processes that are constructed by differentiating log-likelihood with respect to an underlying parameter vector.

6.2. Multiplicative martingales

Let $\{M_t : t \geq 0\}$ denote a process whose logarithm evolves as:

$$\log M_{t+1} - \log M_t = \kappa(X_t, W_{t+1}).$$

Thus the logarithmic counterpart is recognizable as an additive functional as studied in [Chapter 4](#). We call such an $\{M_t : t \geq 0\}$ a multiplicative functional. We explore properties of such processes in some generality in [Chapter 8](#). In this chapter, the processes of interest are multiplicative martingales:

Definition 6.1

The process M is a **multiplicative martingale** if

$$E(\exp[\kappa(X_t, W_{t+1})] \mid X_t) = 1,$$

and thus $E(M_{t+1} \mid M_t, X_t) = M_t$.

Multiplicative martingales provide a convenient way to construct alternative probabilities.

Proposition 6.1

Suppose that $\{M_t : t \geq 0\}$ is a multiplicative martingale. The ratio M_{t+1}/M_t implies a transition probability conditioned on X_t via the formula:

$$\tilde{E}[f(X_t, W_{t+1}) \mid X_t] = E\left[\left(\frac{M_{t+1}}{M_t}\right) f(X_t, W_{t+1}) \mid X_t\right].$$

The resulting process for X remains Markovian, and the implied τ -period transition probability can be represented as:

$$\tilde{E}[f(X_{t+\tau-1}, W_{t+\tau}) \mid X_t] = E\left[\left(\frac{M_{t+\tau}}{M_t}\right) f(X_{t+\tau-1}, W_{t+\tau}) \mid X_t\right].$$

Proof. The linear operator:

$$\tilde{E}(\cdot \mid X_t = x)$$

maps nonnegative functions into nonnegative functions and the unit function into itself. Thus this operator is a conditional expectation. Moreover, it maps functions of (w, x) into functions of x alone. This implies that $\{X_t : t \geq 0\}$ is a first-order Markov process under the implied change of

probability. The representation of the implied τ -period conditional expectation operator follows from the Law of Iterated Expectations.

Remark 6.1

Under the change of measure induced by M_{t+1}/M_t , the shock W_{t+1} typically does not have conditional mean zero.

So far we have not restricted M_0 , the initial condition M_0 . We have only characterized its stochastic evolution. Suppose that the process $\{X_t\}$ is stationary with probability denoted Q , and consider $M_0 = \tilde{q}(X_0)$ where \tilde{q} satisfies:

$$\begin{aligned} E[M_0 f(X_0)] &= E[M_1 f(X_1)] \\ \int \tilde{q}(x) dQ(x) &= 1 \end{aligned}$$

for all bounded (measurable) functions f of the Markov state x . Then $\tilde{Q}(dx) \stackrel{\text{def}}{=} \tilde{q}(x)Q(dx)$ is a stationary distribution under the implied $\tilde{\cdot}$ probability distribution.

In what follows we will impose such an initial condition on M_0 in order that we can apply a Law of Large Numbers as characterized in [Chapter 1](#). In addition we will impose the ergodic restriction that the only solutions to the equation:

$$\tilde{E} \left[\left(\frac{M_{t+1}}{M_t} \right) f(X_{t+1}) \mid X_t = x \right] = f(x)$$

are constant functions with \tilde{Q} measure one as investigated in [Chapter 2](#).

6.3. Multiple models

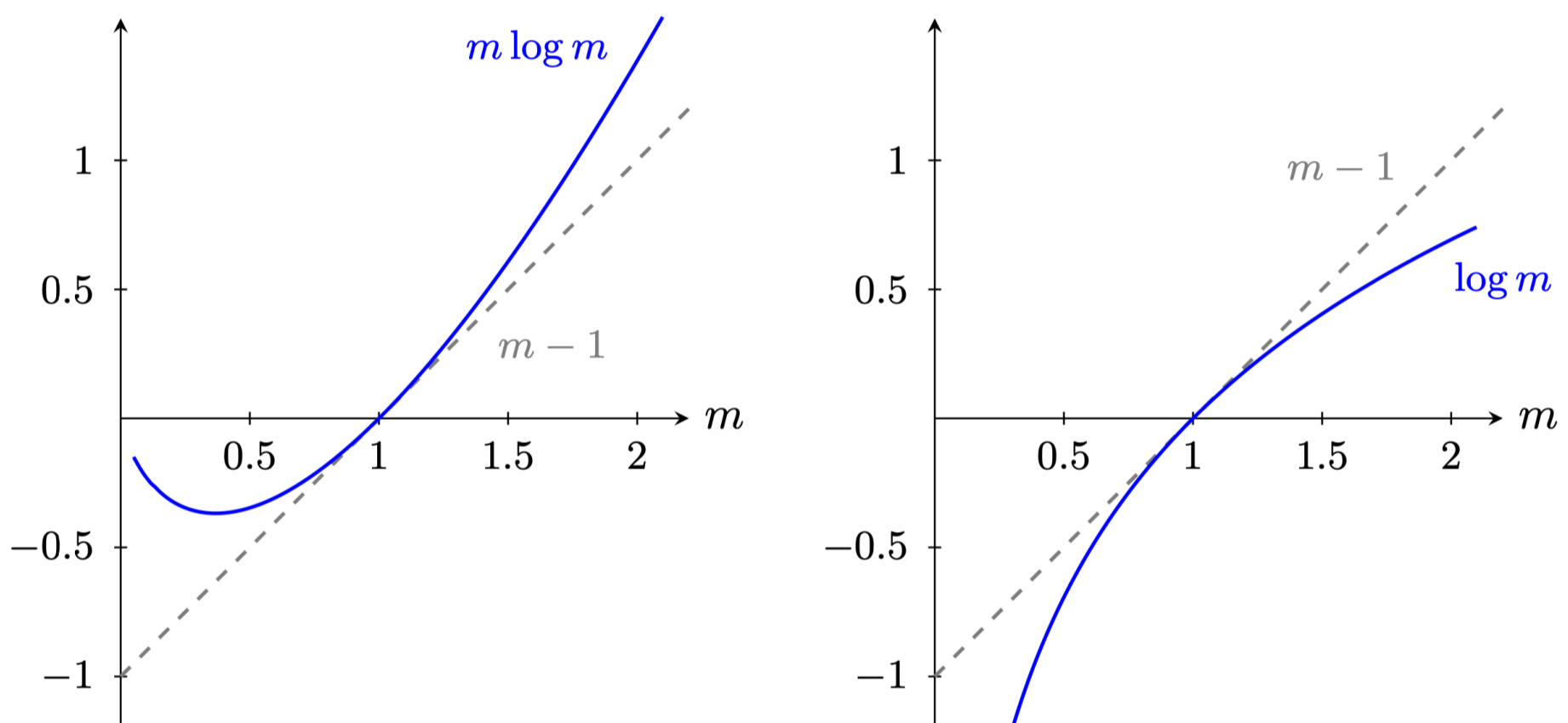
So far, we have considered two models, an initial one and a second one implied by a multiplicative martingale. Now suppose we have $\ell + 1$ such models, given an initial one and ℓ multiplicative martingales. We denote each such martingale as $\{M_t^i : t \geq 0\}$ where each one induces a process that is stationary and ergodic. We may view the initial probability specification as one in which $M_t^0 = 1$ for all $t \geq 0$. In terms of the [Chapter 1](#), think of this as partitioning sample space into $\ell + 1$ partitions. Each partition is matched to one of the martingales, which in turn gives the probabilities conditioned on being in that partition.

In this construction, we choose to normalize model zero to be the probability distribution used for representing all of the probabilities in conjunction with multiplicative martingales. Suppose instead we had

used model one for this purpose. The $\{M_t^i : t \geq 0\}$ processes cease to be martingales because we have changed the underlying probability. Instead, the ratio processes $\{M_t^i/M_t^1 : t \geq 0\}$ are multiplicative martingales with this change in the underlying probabilities. This outcome of altering the baseline probability will be central in the discussion that follows.

6.4. Some large sample properties

We next consider two gradient inequalities that we will use prominently. The function $m \log m$ is convex and the function $\log m$ is concave. A convex function lies above its gradient approximation and a concave function below its gradient. Observe that the gradient is one for both functions when $m = 1$ and the gradient approximation is $m - 1$. [Fig. 6.1](#) illustrates these two inequalities.



(a) The inequality $m \log m \geq m - 1$.

(b) The inequality $m - 1 \geq \log m$.

Fig. 6.1 Gradient inequalities

We use these inequalities in conjunction with conditional expectations to obtain the targets of interest. This amounts to application of what is called Jensen's Inequality and implies

$$\begin{aligned} E(\log M_{t+1} - \log M_t) &\leq 0 \\ E[(\log M_{t+1} - \log M_t)M_{t+1}] &\geq 0. \end{aligned} \tag{6.1}$$

These weak inequalities become strict when $\log M_{t+1} - \log M_t$ is not equal to zero with probability one. Observe that $\{\log M_t : t \geq 0\}$ is an additive functional of the type studied in [Chapter 4](#) with a trend coefficient that is negative. A consequence of the Law of Large Numbers and the first inequality in [\(6.1\)](#) is that:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log M_t = E(\log M_{t+1} - \log M_t) \leq 0.$$

Suppose now we consider expectations under the probabilities implied by the multiplicative martingale $\{M_t \geq 0\}$. Then the second inequality in (6.1) implies that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log M_t = \tilde{E}(\log M_{t+1} - \log M_t) \geq 0.$$

These two limits give a large sample justification for use of the criterion

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log M_t = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=1}^t (\log M_\tau - \log M_{\tau-1}) + \frac{1}{t} \log M_0$$

to determine which of the two models generates the data since the inequalities are strict when the multiplicative martingale is not degenerate.

Proposition 6.2

The process $\{\log M_t : t \geq 0\}$ is a **supermartingale**. That is,

$$E(\log M_{t+1} \mid X_t, M_t) \leq \log M_t, \quad t \geq 0.$$

The process $\{M_t \log M_t : t \geq 0\}$ is a **submartingale**. That is,

$$E(M_{t+1} \log M_{t+1} \mid X_t, M_t) \geq M_t \log M_t, \quad t \geq 0.$$

These insights extend directly to the case of multiple models captured by alternative multiplicative martingales. It leads to the use of

$$\max_{i=0,1,\dots,\ell} \frac{1}{t} \log M_t^i$$

as a way to select models with a large sample justification. As we will show, this approach has a direct application to the method of maximum likelihood.

We conclude this section by pointing to an unusual sample path property of multiplicative martingales.

Proposition 6.3

For a nondegenerate multiplicative martingale, $\{M_t : t \geq 0\}$ converges almost surely to zero even though $E\left(\frac{M_t}{M_0} \mid X_0\right) = 1$ for each t .

Proof. We observed that by the Law of Large Numbers:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log M_t$$

converges to a negative limit almost surely. This in turn implies that

$$\lim_{t \rightarrow \infty} M_t = 0$$

almost surely.

[[Martin, 2012](#)] uses this type of result to argue that cumulative returns on long-dated assets that are positive martingales necessarily have distributions with fat tails. In [Chapter 8](#) we investigate the large deviation behavior of positive martingales.

6.5. Log likelihoods

We now construct the function κ as the increment to a log-likelihood ratio. We let θ denote a parameter vector, and suppose that a vector of date $t + 1$ observations Z_{t+1} given by

$$Z_{t+1} = \phi(X_t, W_{t+1}, \theta).$$

Form:

$$\begin{aligned} \kappa(x, w^* \mid \theta) &= \log \psi(z^* \mid x, \theta) - \log \psi(z^* \mid x, \theta_0) \\ &= \log \psi[\phi(x, w^*) \mid x, \theta] - \log \psi[\phi(x, w^*) \mid x, \theta_0] \end{aligned}$$

where $\psi(z^* \mid x, \theta)$ is the density of Z_{t+1} given X_t and θ . We further assume that given (x, θ) , Z_{t+1} is informationally equivalent to W_{t+1} . This can be justified by an assumption that we can invert ϕ as a function of W_{t+1} . We assume that given θ , X_t is revealed by past Z_t and possibly a finite number of lags. With this construction, we think of model θ_0 as the baseline used for computing expectations.

By factoring a joint density we obtain a likelihood process conditioned on (X_0, θ) as:

$$L_t(\theta) = \sum_{j=1}^t \log \psi(Z_j | X_{j-1}, \theta).$$

We apply our previous arguments to justify finding the

$$\max_{i=0,1,\dots,\ell} L_t(\theta_i)$$

for a discrete set of models.^[1] In what follows, we suppose there is a continuum of models.

6.6. Score processes

In this section, we study first-order necessary conditions associated with the maximum likelihood estimator. For simplicity suppose that the parameter space Θ is an open interval of \mathbb{R} containing the parameter value θ_0 . Inequality (6.1) implies that the parameter value θ_0 necessarily maximizes the population objective

$$\int_{\mathcal{Z}} \log \psi(z^* | x, \theta) \psi(z^* | x, \theta_0) \tau(dz^*). \quad (6.2)$$

Importantly, this objective conditions on x and θ_0 . Suppose that we can differentiate under the integral sign in (6.2) to get the first-order condition

$$\int_{\mathcal{Z}} \left[\frac{d}{d\theta} \log \psi(z^* | x, \theta_0) \right] \psi(z^* | x, \theta_0) \tau(dz^*) = 0. \quad (6.3)$$

Definition 6.2

The **score process** $\{S_t : t = 0, 1, \dots\}$ is defined as

$$S_t = \frac{d}{d\theta} L_t(\theta) \Big|_{\theta_0} = \sum_{j=1}^t \frac{d}{d\theta} \log \psi(Z_j | X_{j-1}, \theta_0).$$

Theorem 6.1

$E(S_{t+1} - S_t | X_t) = 0$, so the score process is an additive martingale with increment $\frac{d}{d\theta} \log \psi(Z_j | X_{j-1}, \theta_0)$ under the θ_0 probability model. From the analysis in [Chapter 3](#) and [\[Billingsley, 1961\]](#), we obtain a central limit approximation:

$$\frac{1}{\sqrt{t}} S_t \rightarrow \mathcal{N}(0, V)$$

where $V = E \left(\left[\frac{d}{d\theta} \log \psi(Z_{t+1} | X_t, \theta_0) \right]^2 \right)$.

Proof. This follows directly from equation (6.3) and Proposition 3.2.

[Theorem 6.1](#) justifies using the martingale central limit theorem [Proposition 3.2](#) to characterize the large sample behavior of the score process. The resulting central limit approximation yields a large sample characterization of the maximum likelihood estimator of θ in a Markov setting. With additional regularity conditions, the following result is typical:

$$\sqrt{t}(\theta_t - \theta_0) \rightarrow \mathcal{N} \left(0, \frac{1}{V} \right),$$

where θ_t maximizes the log likelihood function $L_t(\theta)$. This kind of result motivates interpreting the covariance of the martingale increment of the score process

$$V = E \left(\left[\frac{d}{d\theta} \log \psi(Z_{t+1} | X_t, \theta_0) \right]^2 \right)$$

as a measure of the information in the data about the parameter θ_0 . The scalar V is a measure of what is called **Fisher information** after the statistician R.A. Fisher. This analysis has a direct extension to the case where θ is a vector.

6.6.1. Nuisance parameters

Consider extending the notion of Fisher information to the following situation. There is a vector θ of unknown parameters. We want a measure of information about one component of the parameter vector, although to get that information we have to estimate other components too. Call the first component of θ the ‘parameters of interest’ θ_0 and call the other component the ‘nuisance parameters’ ϑ_0 . Suppose that the likelihood is parameterized on an open set Θ in a finite dimensional Euclidean space and that the true parameter vector is $(\theta_0, \vartheta_0) \in \Theta$. Write the multivariate score process as

$$\left\{ \begin{bmatrix} S_{t+1} \\ \tilde{S}_{t+1} \end{bmatrix} : t = 0, 1, \dots \right\}$$

where $\{S_{t+1} : t = 0, 1, \dots\}$ is the partial derivative of the log-likelihood with respect to θ and $\{\tilde{S}_{t+1} : t = 0, 1, \dots\}$ is the partial derivative with respect to ϑ . Estimating ϑ_0 simultaneously with θ_0 is more difficult than estimating θ_0 when knowing ϑ_0 . Fisher's measure of information corrects for the additional challenge of estimating ϑ_0 simultaneously in order to make inferences about θ_0 . In particular, Fisher's measure of information about θ is the inverse of the (1, 1) component of the covariance matrix partitioned conformably with θ, ϑ :

$$V = \begin{bmatrix} E(S_{t+1} - S_t)^2 & E(S_{t+1} - S_t)(\tilde{S}_{t+1} - \tilde{S}_t)' \\ E(\tilde{S}_{t+1} - \tilde{S}_t)(S_{t+1} - S_t) & E(\tilde{S}_{t+1} - \tilde{S}_t)(\tilde{S}_{t+1} - \tilde{S}_t)' \end{bmatrix}.$$

To represent this inverse in a revealing way, compute the population regression

$$S_{t+1} - S_t = \beta'(\tilde{S}_{t+1} - \tilde{S}_t) + U_{t+1},$$

where β is the population regression coefficient and U_{t+1} is the population regression residual that by construction is orthogonal to the regressor $(\tilde{S}_{t+1} - \tilde{S}_t)$. The population regression induces the representation

$$\begin{bmatrix} S_{t+1} - S_t \\ \tilde{S}_{t+1} - \tilde{S}_t \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \beta' \\ \mathbb{O} & \mathbb{I} \end{bmatrix} \begin{bmatrix} U_{t+1} \\ \tilde{S}_{t+1} - \tilde{S}_t \end{bmatrix},$$

where \mathbb{O} is a vector of zeros. Since U_{t+1} is orthogonal to $\tilde{S}_{t+1} - \tilde{S}_t$,

$$\mathbb{V} = \begin{bmatrix} \mathbf{1} & \beta' \\ \mathbb{O} & \mathbb{I} \end{bmatrix} \begin{bmatrix} EU_{t+1}^2 & \mathbb{O}' \\ \mathbb{O} & E(\tilde{S}_{t+1} - \tilde{S}_t)(\tilde{S}_{t+1} - \tilde{S}_t)' \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbb{O}' \\ \beta & \mathbb{I} \end{bmatrix}.$$

The inverse matrix is:

$$\mathbb{V}^{-1} = \begin{bmatrix} \mathbf{1} & \mathbb{O}' \\ -\beta & \mathbb{I} \end{bmatrix} \begin{bmatrix} \frac{1}{EU_{t+1}^2} & \mathbb{O}' \\ \mathbb{O} & [E(\tilde{S}_{t+1} - \tilde{S}_t)(\tilde{S}_{t+1} - \tilde{S}_t)']^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1} & -\beta' \\ \mathbb{O} & \mathbb{I} \end{bmatrix}.$$

This equation reveals that the (1, 1) component of the partition of matrix \mathbb{V}^{-1} is

$$V_{1,1}^{-1} = \frac{1}{E(U_{t+1}^2)}.$$

We take the Fisher information to be the reciprocal $E(U_{t+1}^2)$. By least squares theory, this measure is

$$E(U_{t+1})^2 \leq E(S_{t+1} - S_t)^2. \quad (6.4)$$

This inequality asserts that the likelihood function contains more information about θ when ϑ is known to be ϑ_0 than when θ and ϑ are both unknown.

Inequality (6.4) offers reasons to be cautious when ignoring the joint estimation challenge and pretending you know ϑ . This latter practice leads you to overstate the information in the sample about the parameter θ of interest. Note that since the multivariate score $\begin{bmatrix} S_{t+1} \\ \tilde{S}_{t+1} \end{bmatrix}$ is a vector of additive martingales, the score regression residual process $\{U_{t+1} : t = 0, 1, \dots\}$ is itself an additive martingale.

6.7. Using a multiplicative martingale for model selection

We now return to the single multiplicative martingale as a way to model an alternative probability.

Suppose we use the martingale at time t , M_t , to construct a criterion for choosing between models. One choice is to check whether M_t is bigger or less than one for a sample of size t . When M_t exceeds one, we choose the probability model induced by the multiplicative martingale. For motivation, recall from [Proposition 6.3](#) that the martingale converges almost surely to zero. Thus the probability of exceeding a fixed threshold must decline to zero under the baseline specification. To provide a more refined result, we study probabilities of making mistakes when using this criterion. We follow [\[Chernoff, 1952\]](#) and others by applying what is called large deviation theory study the probability of making a mistake if the baseline model used to represent expectations governs the data generation. We show how to characterize this probability for large sample sizes.

For notational convenience, we initialize $M_0 = 1$ and condition on the date zero state X_0 . Should the $\{M_t : t \geq 0\}$ of interest be initialized differently, in what follows we would use M_t/M_0 and include M_0 in the date zero conditioning information. Instead of carrying along the extra notation, we just change how we normalize the martingale. We construct a bound on the mistake probability by using an inequality that is implied by two facts:

- The probability that M_t exceeds one is the expectation of the indicator function:

$$\mathbf{1}_{[1, \infty)}(m) \stackrel{\text{def}}{=} \begin{cases} 0 & m < 1 \\ 1 & m \geq 1 \end{cases}$$

- For a scalar $\alpha > 0$,

$$\mathbf{1}_{[1, \infty)}(m) \leq m^\alpha.$$

The left panel of [Fig. 6.2](#) illustrates this inequality. Values of $0 < \alpha < 1$ are of particular interest in which case m^α is a concave increasing function. Thus we use the inequality

$$E [\mathbf{1}_{[1,\infty)}(M_t) | X_0] \leq E [(M_t)^\alpha | X_0].$$

The expectation on the right side of this relation is more tractable to analyze than the one on the left side when we investigate behavior as $t \rightarrow \infty$. Since the inequality applies for arbitrary α , to produce approximations that are as sharp as possible, we will minimize over the choice of α .

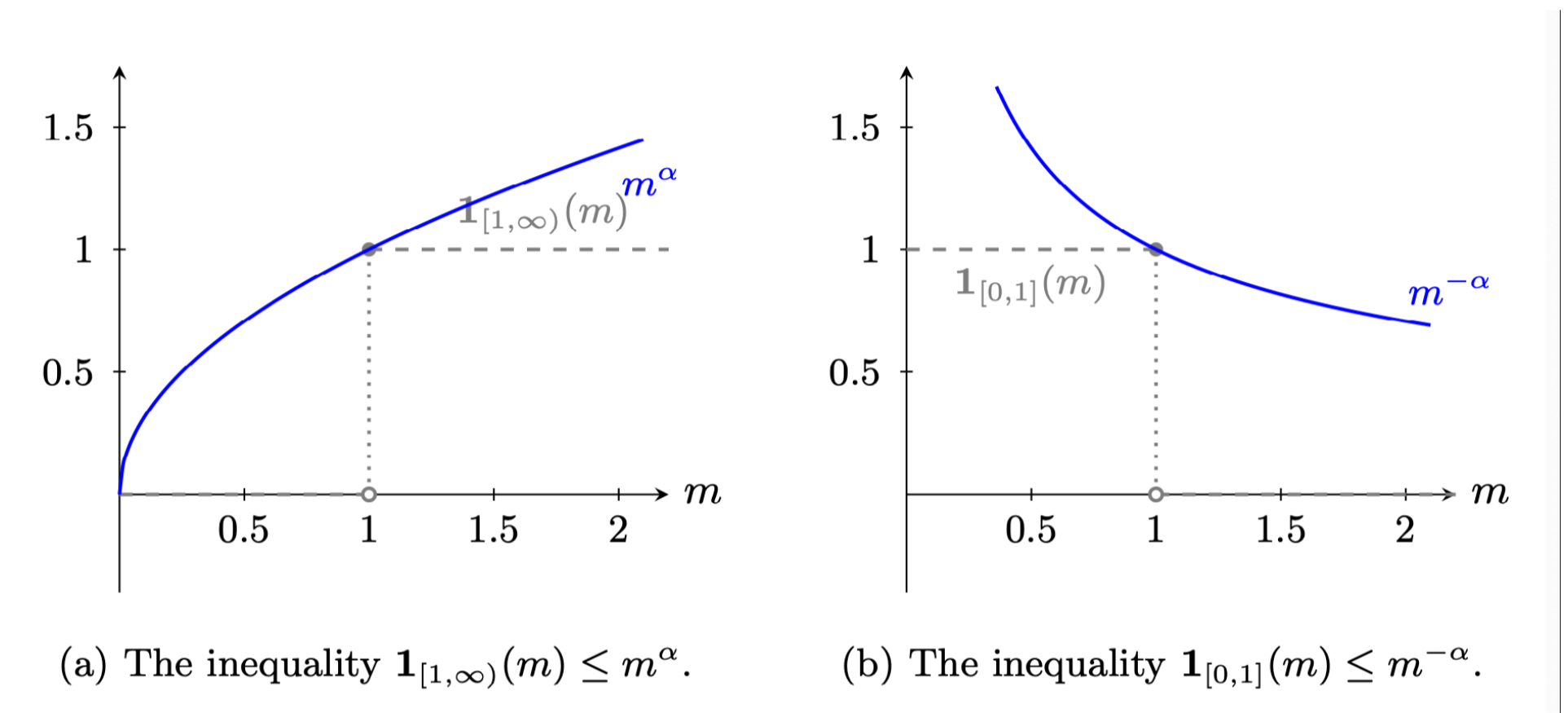


Fig. 6.2 Two indicator function inequalities with $\alpha = \frac{1}{2}$.

Note that m^α for $0 < \alpha < 1$ is a concave function. From the associated gradient inequality, the process $\{(M_t)^\alpha : t \geq 0\}$ is a supermartingale. This uses the same logic as we used when showing that $\{\log M_t : t \geq 0\}$ is a supermartingale. The limit of interest is

$$-\lim_{t \rightarrow \infty} \frac{1}{t} \log E [(M_t)^\alpha | X_0] \stackrel{\text{def}}{=} \eta(\alpha).$$

The resulting $\eta(\alpha)$ is interpretable as an **asymptotic decay rate** for the process $\{(M_t)^\alpha : t \geq 0\}$. To obtain a sharp bound on the limiting behavior of the mistake probability we solve

$$\eta^* \stackrel{\text{def}}{=} \max_{0 < \alpha < 1} \eta(\alpha).$$

Then η^* tells us the asymptotic decay rate in the probability of choosing the multiplicative martingale-based model when the original baseline model actually generates the data.

So far, we have only analyzed one type of mistake. The other possibility is that the alternative model with probabilities induced by the multiplicative martingale generates the data, and $M_t < 1$. In this case, the indicator function and the dominating function are given by:

$$\mathbf{1}_{[0,1]}(m) \leq m^{-\alpha},$$

as is illustrated in the right panel of [Fig. 6.2](#). Since we are interested in computations under the alternative probability measure, we are led to study

$$-\lim_{t \rightarrow \infty} \frac{1}{t} \log E \left[(M_t)^{1-\alpha} \mid X_0 \right] \stackrel{\text{def}}{=} \tilde{\eta}(\alpha)$$

Again, we maximize over α and note that

$$\max_{0 < \alpha < 1} \tilde{\eta}(\alpha) = \max_{0 < \alpha < 1} \eta(\alpha) \stackrel{\text{def}}{=} \eta^*$$

Remarkably, the asymptotic decay rates of the two mistakes are *equated* using one as a threshold. The resulting decay rate is called **Chernoff entropy**. This finding does not imply that the two mistake probabilities are equated. But the rate equalization extends to constant thresholds other than one, including a Bayesian procedure with prior probabilities assigned to each model. Alternative constant thresholds will change the probabilities but not the asymptotic decay rates. Commonly employed classical methods that hold fixed one of the error probabilities independent of sample size do not have fixed thresholds and thus are not covered by this analysis. Nor are they well justified. In contrast to this common approach, fixed threshold rules have both mistake probabilities decay as more data becomes available.

In [Chapter 8](#), we develop methods that, among other things, show how to compute the asymptotic decay rates.

Remark 6.2

Consider a Bayesian decision maker selecting a model. Let π_0 denote the prior probability of that the baseline model generates, and let π_m denote prior probability that the alternative model implied by the multiplicative martingale generates the data. Let v_0 denote the utility reward if the baseline model is chosen correctly, and let v_m denote the utility reward if the multiplicative martingale model is chosen correctly. At time t , the decision maker will select the baseline model by checking if the utility-weighted posterior probability of the baseline model exceeds the multiplicative model counterpart:

$$v_m \frac{\pi_m M_t}{\pi_0 + \pi_m M_t} \leq v_0 \frac{\pi_0}{\pi_0 + \pi_m M_t}$$

With a straightforward computation, this inequality simplifies to:

$$M_t \leq \frac{v_0 \pi_0}{v_m \pi_m},$$

which is a constant threshold rule of the type we have been considering.

[↓ Homework For Econ 312](#)

- [1] This inequality analysis extends pointwise in θ . A pointwise use of the Law of Large Numbers is typically not sufficient to justify statistical consistency.