A SIMPLE MODEL FOR THE EFFECT OF NORMALIZATION ON THE CONVERGENCE RATE OF ADAPTIVE FILTERS

Vítor H. Nascimento

Electronic Systems Eng. Dept., Escola Politécnica, Universidade de São Paulo
Av. Prof. Luciano Gualberto, trav. 3, nº 158, CEP 05508-900 — São Paulo, SP, Brazil
vitor@lps.usp.br

ABSTRACT

We propose a new simple model for the input regressor vectors in adaptive filters. This model allows more insight on the effect of normalization on the convergence rate and eigenvalue spread for the normalized least-mean-squares algorithm (NLMS). Using the new model, we show that NLMS will work best to reduce eigenvalue spread when the input regressor vector points to all directions with equal probability, but with direction-dependent power.

1. INTRODUCTION

The ϵ-NLMS (normalized least-mean-squares) algorithm is given by the following recursion [1]

$$\begin{aligned} e(n) &= d(n) - \boldsymbol{X}(n)^T \boldsymbol{W}(n), \\ \boldsymbol{W}(n+1) &= \boldsymbol{W}(n) + \frac{\mu}{\epsilon + \boldsymbol{X}(n)^T \boldsymbol{X}(n)} \boldsymbol{X}(n) e(n), \end{aligned} \tag{1}$$

where the scalar $d(n)$ is known as the *desired* signal, the vector $\boldsymbol{X}(n) \in \mathbb{R}^M$ is the *regressor* (both have zero mean), and the constant $\mu > 0$ is the *step-size*. The algorithm obtained with $\epsilon = 0$ is usually denominated normalized LMS (NLMS). When the denominator is set to 1, we obtain the standard LMS algorithm. An important matrix for the analysis of these algorithms is $R \triangleq \mathrm{E}\, \boldsymbol{X}(n)\boldsymbol{X}(n)^T$, the autocorrelation of $\boldsymbol{X}(n)$ ($\mathrm{E}(\cdot)$ is the expectation operator, and T denotes transposition).

Normalization is useful mainly because it guarantees stability if $0 < \mu < 2$. Considerable effort has been spent trying to find out whether and when normalization allows faster convergence for a given level of misadjustment [2, 3, 4, 5, 6]. As is well-known, a major problem with LMS is its slow convergence rate when R has a large eigenvalue spread (i.e., the ratio κ between its largest and smallest eigenvalues is much larger than unity). [6] showed that for $\epsilon = 0$ and Gaussian $d(n)$ and $\boldsymbol{X}(n)$, normalization reduces (or does not increase) κ. Given the difficulty of the problem, much of the work in the above references is devoted to the case where $R = \sigma_x^2 I$ (a multiple of the identity matrix, for which $\kappa = 1$). In this paper we propose a simple model, derived from that given in [4], that gives more intuition on when normalization will result in reduced eigenvalue spread and faster convergence.

This work was supported in part by the Brazilian Research Council (CNPq), and by the São Paulo State Research Council (FAPESP).

2. MEAN-SQUARE BEHAVIOR OF ϵ-NLMS

Recalling that for stationary $d(n)$ and $\boldsymbol{X}(n)$ there is a vector $\boldsymbol{W}_*$ (known as the *Wiener solution*) such that [1] $d(n) = \boldsymbol{X}(n)^T \boldsymbol{W}_* + e_0(n)$, with $\mathrm{E}(e_0(n)\boldsymbol{X}(n)) = \boldsymbol{0}$, we may define the weight error vector $\boldsymbol{V}(n) \triangleq \boldsymbol{W}_* - \boldsymbol{W}(n)$, and obtain the error equation

$$\begin{aligned} e(n) &= \boldsymbol{X}(n)^T \boldsymbol{V}(n) + e_0(n), \\ \boldsymbol{V}(n+1) &= \boldsymbol{V}(n) + \frac{\mu \boldsymbol{X}(n) e(n)}{\epsilon + \boldsymbol{X}(n)^T \boldsymbol{X}(n)}, \end{aligned} \tag{2}$$

We shall concentrate on the behavior of the mean-square error $\mathrm{E}(e(n)^2)$ and of the weight error autocorrelation matrix $K(n) = \mathrm{E}(\boldsymbol{V}(n)\boldsymbol{V}(n)^T)$. Assuming, as usual, that sequences $\{\boldsymbol{X}(n)\}$ and $\{e_0(n)\}$ are iid (independent, identically distributed) and independent of each other, one can show that [3] ($\sigma_0^2 = \mathrm{E}\, e_0(n)^2$, $\mathrm{Tr}(\cdot)$ is the trace of a matrix)

$$\mathrm{E}(e(n)^2) = \sigma_0^2 + \mathrm{Tr}(RK(n)), \tag{3}$$

$$\begin{aligned} K(n+1) &= K(n) - \mu R_1 K(n) - \mu K(n) R_1 + \\ &\quad + \mathrm{E}\left[\frac{\boldsymbol{X}(n)\boldsymbol{X}(n)^T K(n) \boldsymbol{X}(n)\boldsymbol{X}(n)^T}{(\epsilon + \boldsymbol{X}(n)^T \boldsymbol{X}(n))^2} \right] + R_2 \sigma_0^2, \end{aligned} \tag{4}$$

$$R_1 \triangleq \mathrm{E}\left[\frac{\boldsymbol{X}(n)\boldsymbol{X}(n)^T}{\epsilon + \|\boldsymbol{X}(n)\|^2} \right], \; R_2 \triangleq \mathrm{E}\left[\frac{\boldsymbol{X}(n)\boldsymbol{X}(n)^T}{(\epsilon + \|\boldsymbol{X}(n)\|^2))^2} \right].$$

One can see from (4) that the convergence of $K(n)$ is governed primarily by the eigenvalues of R_1, specially when μ is small. The equivalent result for LMS is similar, but with no divisions, so R_1 and R_2 would be replaced by R.

Elegant results are given in [2, 3], reducing the expectations defining R_1 and R_2 to one-dimensional integrals for

the case of Gaussian $d(n)$ and $\boldsymbol{X}(n)$. [6] extends the results, comparing the condition numbers of R (which affects the convergence of LMS) and R_1 (which affects the convergence speed of ϵ-NLMS). The expressions for R_1 and R_2 and for the remaining term in (4) are however quite complex, and closed-form expressions are provided only if $R = \sigma_x^2 I$ (i.e., the entries of $\boldsymbol{X}(n)$ are uncorrelated), or for contrived examples for which only one of the M eigenvalues λ_i of R differs from the others.

[5] considers the problem of finding the nonlinearity that gives the best compromise between convergence speed and steady-state error, concluding that one should choose (in our notation) $\mu = 1$ and ϵ large. However, the analysis assumes again that $R = \sigma_x^2 I$.

Another approach is taken by [7, 8], which propose a simple approximation to the expectations in (4), noting that for large enough filter length M it holds that, for example,

$$R_1 \approx \frac{\mathrm{E}\,\boldsymbol{X}(n)\boldsymbol{X}(n)^T}{\mathrm{E}(\epsilon + \boldsymbol{X}(n)^T\boldsymbol{X}(n))} = \frac{R}{\epsilon + \mathrm{Tr}(R)},$$

where $\mathrm{Tr}(R)$ is the trace of matrix R. This approximation gives good results for the steady-state and reasonable results for the transient behavior of the filter. The problem with the transient is caused by approximating R_1 by a multiple of R: the effect of eigenvalue-spread reduction observed in [6] is not taken into account. For large filter lengths the approximation becomes better; however, for some eigenvalue distributions [6] reports large differences for $M \approx 200$.

3. NEW INPUT MODEL

Recalling that the behavior of LMS with small step-size depends only on R, [4] proposes using a fictitious input sequence $\boldsymbol{X}(n)$ that has autocorrelation R, but with a simple structure. The proposed model is ($\|\boldsymbol{X}\|^2 = \boldsymbol{X}^T\boldsymbol{X}$)

$$\boldsymbol{X}(n) = s(n)r(n)\mathcal{X}(n), \tag{5}$$

where $\begin{cases} \Pr\{s(n) = \pm 1\} = 0.5, \ (\pm 1 \text{ have equal probab.}) \\ r(n) \text{ has the same distribution as } \|\boldsymbol{X}(n)\|, \\ \Pr\{\mathcal{X}(n) = \boldsymbol{q}_i\} = p_i = \frac{\lambda_i}{\mathrm{Tr}(R)}, \ \ i = 1 \ldots M. \end{cases}$

Vector $\boldsymbol{q}_i$ is the eigenvector of R with respect to eigenvalue λ_i. Since R is symmetric and positive-definite, it follows that $\boldsymbol{q}_i^T\boldsymbol{q}_j = \delta_{i,j}$, that is, $\{\boldsymbol{q}_i\}$ is an orthonormal set. Variables $s(n)$, $r(n)$ and $\mathcal{X}(n)$ are independent from each other, and are also assumed iid. Note that the autocorrelation matrix of such $\boldsymbol{X}(n)$ is R, since $\mathrm{E}\,r(n)^2 = \mathrm{Tr}(R)$, and

$$\mathrm{E}(\boldsymbol{X}(n)\boldsymbol{X}(n)^T) = \mathrm{E}(r(n)^2)\,\mathrm{E}(s(n)^2)\sum_{i=1}^{M} p_i\boldsymbol{q}_i\boldsymbol{q}_i^T = \sum_{i=1}^{M}\lambda_i\boldsymbol{q}_i\boldsymbol{q}_i^T = R.$$

The idea is to assume that the regressor vector is generated through (5). Since the model is very simple ($\boldsymbol{X}(n)$ can assume only finitely many directions), it simplifies the study of properties of adaptive filters. Using this model, [4] derives several properties of LMS and NLMS and compares both algorithms. The model however implicitly assumes the worst-case situation for NLMS, since it predicts that

$$R_1 \approx \mathrm{E}\left(\frac{r(n)^2}{(\epsilon + r(n)^2)\,\mathrm{Tr}(R)}\right) R$$

As (5) implies that R_1 is a multiple of R, the model ignores a possible reduction in eigenvalue spread given by the normalization.

This restriction can be overcome as follows. Let λ_i be the eigenvalues of R, ν_i the eigenvalues of R_1, and assume that $\boldsymbol{X}(n)$ is iid and generated from

$$\boldsymbol{X}(n) = s(n)\mathcal{X}(n), \tag{6}$$

where $\mathcal{X}(n) = \alpha_i\boldsymbol{q}_i$ with probability p_i, $s(n) = \pm 1$, with probability 0.5 each, $\alpha_i > 0$, $i = 1 \ldots M$, and $\boldsymbol{q}_i$ is as before. $s(n)$ and $\mathcal{X}(n)$ are independent of each other.

Using this model, we have

$$\mathrm{E}\left(\boldsymbol{X}(n)\boldsymbol{X}(n)^T\right) = \sum_{i=1}^{M} p_i\alpha_i^2\boldsymbol{q}_i\boldsymbol{q}_i^T = R,$$

if $p_i\alpha_i^2 = \lambda_i$ — that is, by choosing p_i and α_i properly, the model's autocorrelation matrix will be R. Similarly,

$$\mathrm{E}\left(\frac{\boldsymbol{X}(n)\boldsymbol{X}(n)^T}{\epsilon + \|\boldsymbol{X}(n)\|^2}\right) = \sum_{i=1}^{M} p_i\frac{\alpha_i^2}{\epsilon + \alpha_i^2}\boldsymbol{q}_i\boldsymbol{q}_i^T = R_1$$

if we choose α_i and p_i such that $p_i\alpha_i^2/(\epsilon + \alpha_i^2) = \nu_i$.

Substituting $p_i\alpha_i^2 = \lambda_i$ above, we solve for α_i, p_i:

$$\alpha_i^2 = \frac{\lambda_i}{\nu_i} - \epsilon, \qquad\qquad p_i = \frac{\lambda_i}{\alpha_i^2}.$$

This model needs knowledge of the ν_i, but as we shall see, it gives more accurate approximations for the learning curves of ϵ-NLMS. This model makes an implicit assumption, that the eigenvectors of R and of R_1 are the same — [2, 3] show that this is exact for Gaussian $\boldsymbol{X}(n)$ and $d(n)$.

4. WHEN DOES NORMALIZATION IMPLY FASTER CONVERGENCE?

Using (6), we can understand a little better what is the influence of normalization on the convergence rate of ϵ-NLMS. Let us consider two situations with the same R but with different R_1:

A. $\alpha_i^2 = \mathrm{E}\,\|\boldsymbol{X}(n)\|^2 = \mathrm{Tr}(R)$, $p_i = \lambda_i/\,\mathrm{Tr}(R)$. This is model (5) again, with constant $r(n) = \alpha_i$. As we saw

above, this means that R_1 is a multiple of R. We can also evaluate R_2 for this situation, resulting

$$R_2 = \frac{\mathrm{Tr}(R)}{(\epsilon + \mathrm{Tr}(R))^2} R,$$

and with this, for small step-size we have

$$K(n+1) \approx K(n) - \frac{\mu}{\epsilon + \mathrm{Tr}(R)} K(n) R - \\ - \frac{\mu}{\epsilon + \mathrm{Tr}(R)} R K(n) + \frac{\mu^2 \sigma_0^2}{(\epsilon + \mathrm{Tr}(R))^2} R,$$

and the convergence rate of ϵ-NLMS will be the same obtained with LMS with $\mu_{LMS} = \mu/(\epsilon + \mathrm{Tr}(R))$ (see [1]). For the limit $K(\infty) = \lim_{n\to\infty} K(n)$, we obtain

$$K(\infty) R + R K(\infty) \approx \frac{\mu \sigma_0^2}{\epsilon + \mathrm{Tr}(R)} R,$$

$$\text{and thus } K(\infty) \approx \frac{\mu}{2(\epsilon + \mathrm{Tr}(R))} \sigma_0^2 I, \qquad (7)$$

$$\text{and} \quad \lim_{n\to\infty} \mathrm{E}\, e(n)^2 \approx \sigma_0^2 \left(1 + \frac{\mu\, \mathrm{Tr}(R)}{2(\epsilon + \mathrm{Tr}(R))}\right). \qquad (8)$$

B. $p_i = 1/M$, $\alpha_i^2 = M\lambda_i$. Now we have

$$R_1 = \sum_{i=1}^{M} \frac{\lambda_i}{\epsilon + M\lambda_i} \boldsymbol{q}_i \boldsymbol{q}_i^T.$$

If $\epsilon \ll M\lambda_i$ for $i = 1 \ldots M$, $R_1 \approx Q/M$, where $Q = \sum_{i=1}^{M} \boldsymbol{q}_i \boldsymbol{q}_i^T$ is an orthogonal matrix: $QQ^T = I$, so its eigenvalue spread will be 1. If ϵ is not so small, we would still have a considerable reduction in the eigenvalue spread of R_1, since (assuming $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_M$)

$$\frac{\lambda_{\max}(R_1)}{\lambda_{\min}(R_1)} = \frac{\lambda_1(\epsilon + M\lambda_M)}{\lambda_M(\epsilon + M\lambda_1)} \leq \frac{\lambda_1}{\lambda_M}.$$

In addition,

$$R_2 = \frac{1}{M} \sum_{i=1}^{M} \frac{M\lambda_i}{(\epsilon + M\lambda_i)^2} \boldsymbol{q}_i \boldsymbol{q}_i^T.$$

For small step-sizes, we have

$$K(\infty) \sum_{i=1}^{M} \frac{\lambda_i}{\epsilon + M\lambda_i} \boldsymbol{q}_i \boldsymbol{q}_i^T + \sum_{i=1}^{M} \frac{\lambda_i}{\epsilon + M\lambda_i} \boldsymbol{q}_i \boldsymbol{q}_i^T K(\infty) \approx \\ \approx \mu\sigma_0^2 \sum_{i=1}^{M} \frac{\lambda_i}{(\epsilon + M\lambda_i)^2} \boldsymbol{q}_i \boldsymbol{q}_i^T. \qquad (9)$$

From (3) we obtain, using $\mathrm{Tr}(K(n)\boldsymbol{q}_i\boldsymbol{q}_i^T) = \boldsymbol{q}_i^T K(n) \boldsymbol{q}_i$,

$$\mathrm{E}\, e(n)^2 = \sigma_0^2 + \sum_{i=1}^{M} \big(\lambda_i \boldsymbol{q}_i^T K(n) \boldsymbol{q}_i\big),$$

Multiplying (9) to the left by $\boldsymbol{q}_k^T$ and to the right by $\boldsymbol{q}_k$, and recalling that $\boldsymbol{q}_k^T \boldsymbol{q}_i = \delta_{ki}$,

$$2\boldsymbol{q}_k^T K(\infty) \boldsymbol{q}_k = \frac{\mu\sigma_0^2}{\epsilon + M\lambda_k}.$$

Using the last two results we obtain

$$\lim_{n\to\infty} \mathrm{E}\, e(n)^2 \approx \sigma_0^2 \left(1 + \frac{\mu}{2} \sum_{i=1}^{M} \frac{\lambda_i}{\epsilon + M\lambda_i}\right). \qquad (10)$$

Comparing cases A (8) and B (10), we note that both steady-state mean-square errors are approximately equal for $\epsilon \ll M\lambda_M$ and for $\epsilon \gg \mathrm{Tr}(R)$, and that (10) is a little larger for intermediate values of ϵ. However, specially for small ϵ, case B will converge much faster than case A.

We conclude that normalization is particularly useful when the eigenvalue spread of R is caused by a mechanism similar to case B: all directions in $\mathbb{R}^M$ are "visited" by $\boldsymbol{X}(n)$ with similar probability, but the amplitude of $\boldsymbol{X}(n)$ is direction-dependent, resulting in a pattern of equal probability p_i and very dissimilar α_i. On the other hand, there will be little convergence-rate gain when the eigenvalue spread results from a mechanism similar to case A, that is, when some directions in $\mathbb{R}^M$ are "visited" less often by $\boldsymbol{X}(n)$, but always with the same average amplitude, giving a pattern of equal-size α_i and very dissimilar p_i.

5. EXAMPLES AND SIMULATIONS

Our new model may be used to understand what happens with a filter in some situations. For example, consider a two-tap filter with $\boldsymbol{X}(n) = \begin{bmatrix} x(n) & x(n-1) \end{bmatrix}^T$, $x(n) = a_1 x(n-1) + v(n)$, with $v(n)$ a white-noise Gaussian sequence with variance such that $\mathrm{E}\, x(n)^2 = 1$. If a_1 is close to 1, $x(n)$ and $x(n-1)$ will be highly correlated, and R will be nearly singular. Normalization will reduce eigenvalue spread to a certain degree: $\boldsymbol{X}(n)$ will tend to stay close to the direction $[\,1\ 1\,]^T$ (which is an eigenvector of R). For $\boldsymbol{X}(n)$ to point to the other eigenvector, $[\,1\ -1\,]^T$, $x(n-1)$ must necessarily be small (otherwise, given the large value of a_1, the probability of $x(n)$ and $x(n-1)$ having different signs would be very small). Therefore, there is a mixture between cases A and B: normalization will correct the problem of small magnitude, but not the small probability problem.

Fig. 1 shows this effect. For $a_1 = 0.99$, $\epsilon = 10^{-2}$, and $M = 2$, we plotted the points $\big(x(n),\ x(n-1)\big)$ (a) and $\big(x(n)/(\epsilon + x(n)^2 + x(n-1)^2),\ x(n-1)/(\epsilon + x(n)^2 + x(n-1)^2)\big)$ (b) for 10^4 points. The eigenvalue spread was reduced from 199 for R to ≈ 19 for R_1.

Our second example is an ϵ-NLMS with $M = 30$, $\mu = \epsilon = 0.1$, and

$$\boldsymbol{X}(n) = \begin{bmatrix} x(n) & \ldots & x(n-29) \end{bmatrix}]^T,$$

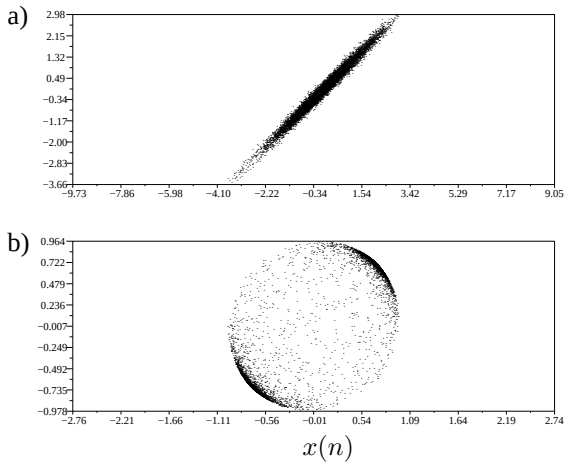

Fig. 1. Plots of points $\big(x(n),\ x(n-1)\big)$ (a) and $\big(x(n)/(x(n)^2+x(n-1)^2),\ x(n-1)/(x(n)^2+x(n-1)^2)\big)$ (b) for $M = 2$ with $\epsilon = 0.01$ and $a_1 = 0.99$ (see text).

with $x(n)$ taken from an AR filter with transfer function $H(z) = b_0/(1-0.3z^{-1}+0.8z^{-2})$, with b_0 such that $\mathrm{E}\,x(n)^2 = 1$. The filter was used to estimate an FIR filter with impulse response given by a von Hann window normalized so that $\boldsymbol{W}_*^T\boldsymbol{W}_* = 1$, with $\sigma_0^2 = 10^{-6}$. Fig. 2 shows a simulation (an average of 1,000 runs), the model proposed in [8], and an approximation obtained using (6) (with R_1 estimated from 3×10^6 points, and using the new model to approximate R_2.)

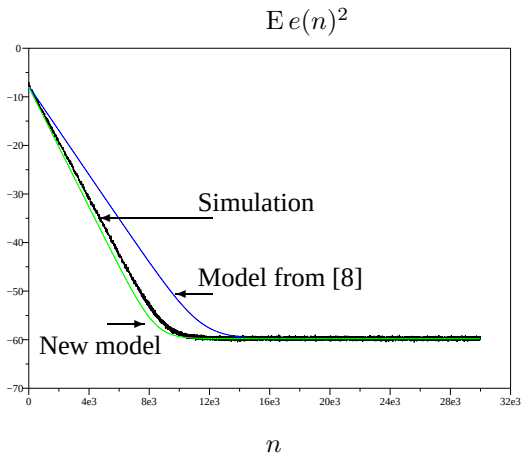

Fig. 2. Learning curves for ϵ-NLMS with $M = 30$, $\mu = 0.1$ and $\epsilon = 0.1$ (see text).

We performed simulations for M up to 200, and for μ up to 1. The new model gives closer predictions for all cases, although the difference between our model and that of [8] indeed gets smaller for larger M.

6. CONCLUSIONS

We proposed a new model for the input regressor vector in adaptive filters, modifying an idea first suggested in [4]. Our new model allows a better understanding of the influence of normalization on the convergence rate of a filter. The new model also gives accurate predictions for the learning curves, at the cost of needing the evaluation of the eigenvalues of the normalized autocorrelation matrix R_1.

7. REFERENCES

[1] Ali H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley-Interscience, 2003.

[2] Neil J. Bershad, "Analysis of the normalized LMS algorithm with Gaussian inputs," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 793–806, 1986.

[3] Neil J. Bershad, "Behavior of the ϵ-normalized LMS algorithm with Gaussian inputs," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, pp. 636–644, May 1987.

[4] D. T. M. Slock, "On the convergence behavior of the LMS and the normalized LMS algorithms," *IEEE Transactions on Signal Processing*, vol. 41, no. 9, pp. 2811–2825, Sept. 1993.

[5] S. C. Douglas and T. H.-Y. Meng, "Normalized data nonlinearities for LMS adaptation," *IEEE Transactions on Signal Processing*, vol. 42, no. 6, pp. 1352–1365, June 1994.

[6] P. E. An, M. Brown, and C. J. Harris, "On the convergence rate performance of the normalized least-mean-square adaptation," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1211–1214, Sept. 1997.

[7] Márcio H. Costa and José C. M. Bermudez, "An improved model for the normalized LMS algorithm with gaussian inputs and large number of coefficients," in *Proceedings of the ICASSP'02*, 2002, vol. 2, pp. 1385–1388.

[8] José C. M. Bermudez and Márcio H. Costa, "A statistical analysis of the ϵ-NLMS and NLMS algorithms for correlated Gaussian signals," in *Anais do XX Simpósio Brasileiro de Telecomunicações (CDROM)*, 2003, pp. 1–5, may be obtained at http://www.lps.usp.br/∼vitor/bermudez.pdf.