บทที่ 2
ขอบเขตของอัตราการเรียนรู้ในการลดระดับตามแนวเกรเดียนต์

การปรับอัตราการเรียนรู้ในข่ายงานระบบประสาท สามารถทำให้ประสิทธิภาพการเรียนรู้ดีขึ้นได้ดังที่กล่าวไว้ในความเบื้องต้นของบทที่ 1 แต่ต้องคำนึงถึงขอบเขตที่สามารถปรับได้โดยระบบไม่เสียเสถียรภาพในการเรียนรู้ ในบทนี้จึงได้กล่าวถึงลักษณะการลู่เข้าสู่ค่าเหมาะสมที่สุดของพารามิเตอร์ในระบบ การปรับค่าพารามิเตอร์ที่ใช้จะเป็นไปตามสมการ (1.5) ซึ่งอาศัยเกรเดียนต์เพื่อกำหนดค่าถัดไป

การปรับพารามิเตอร์โดยอาศัยเกรเดียนต์นี้ เรียกว่า การลดระดับตามแนวเกรเดียนต์ วิธีนี้จะกำหนดเวกเตอร์พารามิเตอร์ขึ้นใหม่เป็นลำดับเพื่อให้เข้าใกล้ค่าเหมาะสมที่สุดมากขึ้น โดยกำหนดให้ไปในทิศทางตรงข้ามเกรเดียนต์ และก้าวไปด้วยระยะเท่ากับ $\beta/2$ ตามสมการ (1.5) ซึ่ง β ก็คือค่าอัตราการเรียนรู้ เป็นตัวกำหนดความยาวก้าวและเป็นตัวควบคุมลักษณะการลู่เข้า

นอกจากนี้ ยังได้กล่าวถึงบทบาทของอัตราการเรียนรู้ที่มีต่อการลดระดับตามแนวเกรเดียนต์ กรณีที่สัญญาณขาเข้ามีลักษณะเป็นกระบวนการสโตคาสติกส์ (stochastic process) ทั้งที่ทราบและไม่ทราบค่าทางสถิติของสัญญาณ ในกรณีที่ไม่ทราบค่าทางสถิติของสัญญาณจึงมีการปรับวิธีการหาค่าเกรเดียนต์เป็นวิธีเกรเดียนต์แบบสโตคาสติกส์ (Stochastic Gradient) หรือเรียกอีกอย่างว่า วิธีกำลังสองน้อยที่สุด (Least Mean Square) ซึ่งเป็นผลให้การปรับพารามิเตอร์ไม่สามารถก้าวถึงค่าเหมาะสมที่สุดได้ ความเข้าใจถึงบทบาทของอัตราการเรียนรู้นี้เป็นแนวทางสำหรับการปรับค่าให้เปลี่ยนแปลงตามเวลาอย่างเหมาะสม

เนื้อหาที่นำมากล่าวถึงในบทนี้ ได้นำมาจากหนังสือ Adaptive Filters ของ Honig M. L. และ Messerschmitt D. G. [14] โดยเลือกเนื้อหาในส่วนที่เกี่ยวข้องและนำมาเรียบเรียงและแก้ไขตามความเหมาะสม

2.1 จุดเหมาะสมที่สุดของการลดระดับตามแนวเกรเดียนต์

ในหัวข้อนี้ แสดงถึงค่าเหมาะสมที่สุดของพารามิเตอร์ที่ทำการปรับโดยวิธีการลดระดับตามแนวเกรเดียนต์ (Gradient Descent) เมื่อใช้ฟังก์ชันเป้าหมายเป็นความผิดพลาดกำลังสองเฉลี่ย (mean-square error) หรือ MSE ในกรณีที่ทราบค่าทางสถิติของสัญญาณขาเข้า โดยใช้วิธี

การทางคณิตศาสตร์ เพื่อแสดงค่าเหมาะสมที่สุดที่เป็นเป้าหมายของการลดระดับตามแนวเกรเดียนต์

เมื่อกำหนดให้ $\underline{w}$ เป็นเวกเตอร์ที่มีส่วนประกอบ (component) เป็นพารามิเตอร์ของระบบ มีขนาดมิติเท่ากับ N $e(t)$ เป็นความผิดพลาดระหว่างสัญญาณอ้างอิง $d(t)$ และสัญญาณขาออก $o(t)$ พิจารณาข่ายงานระบบประสาทแบบชั้นเดียวที่มีลักษณะเป็นเชิงเส้น จะได้ว่า $o(t) = \underline{w}^T \underline{x}(t)$ เมื่อ $\underline{x}(t)$ เป็นเวกเตอร์สัญญาณขาเข้าของระบบที่มีขนาดมิติเท่ากับ N ค่าของ $e(t)$ จะขึ้นอยู่กับ $\underline{w}$, $\underline{x}(t)$ และ $d(t)$ ดังนี้

$$e(t) = d(t) - \underline{w}^T \underline{x}(t) \tag{2.1.1}$$

เราสามารถสร้างฟังก์ชันเป้าหมายในการหาจุดเหมาะสมที่สุดเป็นฟังก์ชันกำลังสอง (quadratic) ของ $e(t)$ เพื่อให้มีค่ามากกว่าหรือเท่ากับศูนย์ และมีจุดเหมาะสมที่สุดเพียงจุดเดียว ซึ่งแสดงเป็นสมการได้ดังนี้

$$e^2(t) = d^2(t) - 2\underline{w}^T d(t)\underline{x}(t) + \underline{w}^T \underline{x}(t)\underline{x}^T(t)\underline{w} \tag{2.1.2}$$

ในกรณีที่สัญญาณขาเข้าของระบบมีลักษณะเป็นกระบวนการสโตคาสติกส์ เราไม่สามารถหาค่าที่แน่นอนของระบบได้ ดังนั้น เราจึงใช้ค่าทางสถิติเพื่อหาจุดเหมาะสมที่สุดแทน ซึ่งทำได้โดยใช้ค่าคาดหวัง (expectation value) ของฟังก์ชันเป้าหมายมาทำการหาจุดเหมาะสมที่สุด เราสามารถหาค่าคาดหวังของสมการ (2.1.2) ได้เป็น

$$\begin{aligned} E\left[e^2(t)\right] &= E\left[d^2(t)\right] - 2\underline{w}^T E\left[d(t)\underline{x}(t)\right] + \underline{w}^T E\left[\underline{x}(t)\underline{x}^T(t)\right]\underline{w} \\ &= E\left[d^2(t)\right] - 2\underline{w}^T \underline{p} + \underline{w}^T \Phi \underline{w} \end{aligned} \tag{2.1.3}$$

โดยที่สัญญลักษณ์ E แทนการหาค่าคาดหวัง สมการ (2.1.3) นี้จะเป็นฟังก์ชันกำลังสองของเวกเตอร์ $\underline{w}$ โดยเวกเตอร์ $\underline{p}$ และเมตริกซ์ Φ ถูกกำหนดขึ้นเพื่อใช้แทนค่าในสมการดังนี้

$$\underline{p} = E\left[d(t)\underline{x}(t)\right] \tag{2.1.4}$$

$$\Phi = E\left[\underline{x}(t)\underline{x}^T(t)\right] = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,N} \\ \phi_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \phi_{N,1} & \cdots & \cdots & \phi_{N,N} \end{bmatrix} \tag{2.1.5}$$

$$\phi_{ij} = E\left[x_i(t)x_j(t)\right]; \qquad 1 \le i, j \le N \tag{2.1.6}$$

เมตริกซ์ Φ เป็นเมตริกซ์อัตสหสัมพันธ์ (autocorrelation) ของสัญญาณขาเข้า $\underline{x}(t)$ จึงมีคุณสมบัติเป็นเมตริกซ์สมมาตรและเป็นบวกกึ่งแน่นอน (positive semidefinite) ส่วนใหญ่เราจะกำหนดให้เมตริกซ์ Φ เป็นบวกแน่นอน (positive definite) เพื่อให้เป็นเมตริกซ์ไม่เอกฐานทำให้สามารถหาเมตริกซ์ผกผัน (inverse matrix) ได้ และมีค่าเฉพาะ (eigenvalue) เป็นจำนวนจริงบวก

สมการ (2.1.3) สามารถทำเป็นกำลังสองสมบูรณ์ ได้เป็น

$$E\left[e^2(t)\right] = E\left[d^2(t)\right] - \underline{p}^T\Phi^{-1}\underline{p} + \left(\Phi^{-1}\underline{p} - \underline{w}\right)^T \Phi\left(\Phi^{-1}\underline{p} - \underline{w}\right) \tag{2.1.7}$$

เมื่อ Φ เป็นเมตริกซ์บวกแน่นอน เทอมที่ 3 ของสมการ (2.1.7) ซึ่งเป็นเทอมกำลังสองจะมีค่ามากกว่าหรือเท่ากับศูนย์ เราสามารถหาเวกเตอร์พารามิเตอร์ที่จุดเหมาะสมที่สุด โดยกำหนดให้เทอมที่ 3 นี้เป็นศูนย์ เพราะเป็นเทอมเดียวที่มีเวกเตอร์ $\underline{w}$ เกี่ยวข้อง ดังนี้

$$\left(\Phi^{-1}\underline{p} - \underline{w}\right)^T \Phi\left(\Phi^{-1}\underline{p} - \underline{w}\right) = 0$$

และหาเวกเตอร์พารามิเตอร์ที่จุดเหมาะสมที่สุด ได้เป็น

$$\underline{w}_{opt} = \Phi^{-1}\underline{p} \tag{2.1.8}$$

เมื่อเทอมที่ 3 ในสมการ (2.1.7) เป็น 0 ดังนั้น ค่าฟังก์ชันเป้าหมายที่จุดเหมาะสมที่สุด จะเป็น

$$E\left[e^2(t)\right] = E\left[d^2(t)\right] - \underline{p}^T\Phi^{-1}\underline{p} \tag{2.1.9}$$

$$= E\left[d^2(t)\right] - \underline{p}^T\underline{w}_{opt} \tag{2.1.10}$$

จากสมการ (2.1.8) ถึง (2.1.10) แสดงว่า เวกเตอร์พารามิเตอร์และฟังก์ชันเป้าหมายที่จุดเหมาะสมที่สุด สามารถหาได้เมื่อทราบเมตริกซ์อัตสหสัมพันธ์ Φ และเวกเตอร์สหสัมพันธ์ข้าม $\underline{p}$

2.2 หลักการความตั้งฉาก (Orthogonality Principle)

ในหัวข้อที่ 2.1 เราหาจุดเหมาะสมที่สุดได้โดยใช้ระบบสมการเชิงเส้นและเมตริกซ์ แต่เรายังสามารถใช้การหาอนุพันธ์ของฟังก์ชันเป้าหมายเทียบกับพารามิเตอร์เพื่อหาจุดเหมาะสมที่สุดได้เช่นกัน ผลที่ได้จะตรงกัน แต่สมการที่ได้จากการหาอนุพันธ์นี้สามารถแปลความหมายเพื่อแสดงความสัมพันธ์ระหว่างสัญญาณขาเข้า $\underline{x}(t)$ และความผิดพลาด $e(t)$ ที่จุดเหมาะสมที่สุดได้

เราสามารถหาอนุพันธ์ของฟังก์ชันเป้าหมายเทียบกับพารามิเตอร์ และกำหนดให้เท่ากับศูนย์เพื่อหาจุดเหมาะสมที่สุดได้ดังนี้

$$0 = \frac{\partial}{\partial w_m} E\left[e^2(t)\right], \qquad 1 \le m \le N \tag{2.2.1}$$

$$= 2E\left[e(T)\frac{\partial}{\partial w_m}e(t)\right], \qquad 1 \le m \le N \tag{2.2.2}$$

$$= -2E\left[e(t)x_m(t)\right], \qquad 1 \le m \le N \tag{2.2.3}$$

สมการ (2.2.3) แสดงให้เห็นว่า ความผิดพลาดไม่มีสหสัมพันธ์ (uncorrelate) กับสัญญาณขาเข้าของระบบ เพราะค่าคาดหวังของผลคูณที่เกิดจากสัญญาณทั้งสองมีค่าเป็นศูนย์ เช่นเดียวกับลักษณะของเวกเตอร์ตั้งฉาก (orthogonal) ที่ผลคูณภายใน (inner product) มีค่าเป็นศูนย์ จึงตั้งชื่อหลักการนี้ว่า “หลักการความตั้งฉาก”

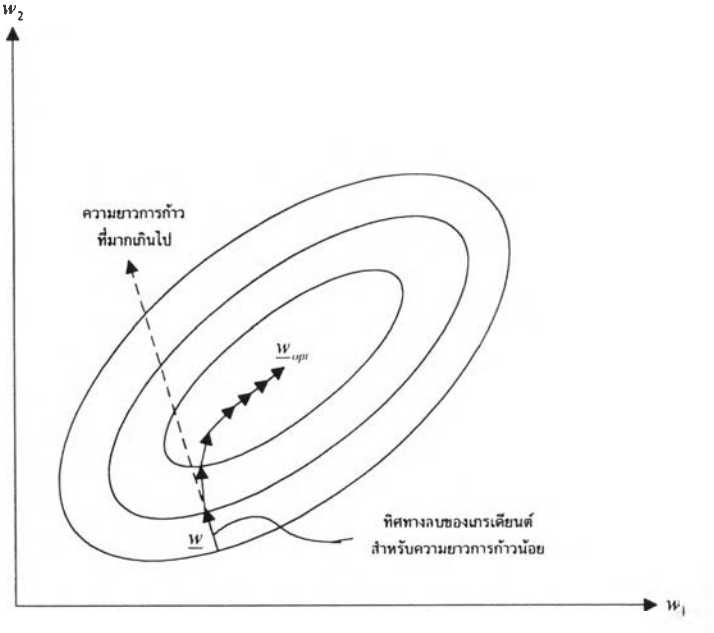
2.3 การหาจุดเหมาะสมที่สุดด้วยวิธีการลดระดับตามแนวเกรเดียนต์

เราสามารถหาจุดเหมาะสมที่สุดโดยใช้วิธีการทางคณิตศาสตร์ อย่างระบบสมการเชิงเส้นหรือการหาเกรเดียนต์ ในกรณีที่เราไม่สามารถหาเกรเดียนต์ของฟังก์ชันเป้าหมายหรือการคำนวณหาเมตริกซ์ผกผัน Φ^{-1} มีความยุ่งยากซับซ้อน โดยเฉพาะในกรณีที่สัญญาณขาเข้ามีจำนวนมาก

เราสามารถใช้ขั้นตอนก ารปรับพารามิเตอร์เพื่อให้เข้าใกล้จุดเหมาะสมมากขึ้นเรื่อย ๆ จนไปถึงจุดเหมาะสมที่สุดได้ ซึ่งวิธีที่ใช้กันโดยทั่วไปสำหรับการหาจุดเหมาะสมที่สุดแบบนี้ คือ การลดระดับตามแนวเกรเดียนต์

การลดระดับตามแนวเกรเดียนต์ ใช้การกำหนดเวกเตอร์พารามิเตอร์ขึ้นใหม่เป็นลำดับ การกำหนดเวกเตอร์แต่ละครั้งจะเข้าใกล้จุดเหมาะสมที่สุดมากขึ้นโดยอาศัยเกรเดียนต์เพื่อกำหนดเวกเตอร์ใหม่ในทิศทางตรงข้าม เพราะว่า ฟังก์ชันเป้าหมายมีอัตราการเพิ่มสูงสุดเมื่อปรับพารามิเตอร์ไปในทิศทางของเกรเดียนต์ ดังนั้น เมื่อกำหนดเวกเตอร์ใหม่โดยให้ก้าวไปในทิศทางตรงข้ามเพียงเล็กน้อย ทำให้ฟังก์ชันเป้าหมายมีค่าลดลงและเวกเตอร์ใหม่เข้าใกล้จุดเหมาะสมที่สุดมากขึ้น

ในกรณีที่มีพารามิเตอร์ของระบบ 2 ตัวสามารถแสดงการก้าวในวิธีการลดระดับตามแนวเกรเดียนต์ได้ดังในรูปที่ 2.1 เห็นได้ว่าคอนทัวร์ (contour) ของฟังก์ชันเป้าหมายซึ่งเป็นฟังก์ชันกำลังสองเป็นรูปวงรี ส่วนลูกศรในรูปที่ 2.1 จะแสดงทิศทางที่กำหนดเวกเตอร์ใหม่ขึ้น ซึ่งต้องกำหนดความยาวก้าว (step size) ให้มีค่าน้อยพอที่จะทำให้ลู่เข้าสู่จุดเหมาะสมที่สุดได้ ถ้ากำหนดให้มีค่ามากเกินไปจะทำให้ก้าวข้ามจุดเหมาะสมที่สุด



รูปที่ 2.1 การหาจุดเหมาะสมที่สุดโดยวิธีการลดระดับตามแนวเกรเดียนต์

การลดระดับตามแนวเกรเดียนต์ สามารถแสดงเป็นสมการสำหรับปรับพารามิเตอร์ ได้ดังนี้

$$\underline{w}_{j+1} = \underline{w}_j - \frac{\beta}{2}\nabla_{\underline{w}_j}\left\{E\left[e^2(t)\right]\right\} \tag{2.3.1}$$

เมื่อ $\underline{w}_{j+1}$ เป็นเวกเตอร์ที่กำหนดขึ้นใหม่จากเวกเตอร์ $\underline{w}_j$ ที่ถูกกำหนดขึ้นก่อนหน้า โดยที่ j เป็นดัชนีของครั้งที่ทำการปรับ ส่วน β เป็นค่าคงตัวที่มีค่าเป็นบวกน้อย ๆ ค่า β นี้เป็นความยาวก้าวซึ่งจะต้องกำหนดให้น้อยพอที่จะทำให้ลู่เข้าจุดเหมาะสมที่สุดได้

เราสามารถหาเกรเดียนต์ของฟังก์ชันเป้าหมายเทียบกับพารามิเตอร์ในสมการ (2.1.7) ได้เป็น $2\Phi\left(\Phi^{-1}\underline{p} - \underline{w}\right)$ และนำมาแทนในสมการ (2.3.1) ซึ่งจะได้สมการสำหรับปรับพารามิเตอร์เป็น

$$\underline{w}_{j+1} = \underline{w}_j + \beta\left(\underline{p} - \Phi\underline{w}_j\right) \tag{2.3.2}$$

$$= \left(I - \beta\Phi\right)\underline{w}_j + \beta\underline{p} \tag{2.3.3}$$

เมื่อ I เป็นเมตริกซ์เอกลักษณ์ (Identity matrix)

ขั้นตอนของการลดระดับตามแนวเกรเดียนต์เริ่มจากกำหนดเวกเตอร์เริ่มต้น $\underline{w}_0$ และปรับค่าเวกเตอร์พารามิเตอร์เพื่อก้าวเข้าใกล้เวกเตอร์ที่จุดเหมาะสมที่สุด $\underline{w}_{opt}$ ดังแสดงไว้ในหัวข้อที่ 2.1 ทำให้เกิดเวกเตอร์พารามิเตอร์เป็นลำดับ เราจะพิจารณาการลู่เข้าสู่จุดเหมาะสมที่สุดของลำดับเวกเตอร์ที่เกิดขึ้นนี้ โดยกำหนด

$$\underline{q}_j = \underline{w}_j - \Phi^{-1}\underline{p} \tag{2.3.4}$$

ซึ่งเป็นเวกเตอร์ผลต่างระหว่างเวกเตอร์พารามิเตอร์ที่กำหนดขึ้นในปัจจุบันกับเวกเตอร์ที่จุดเหมาะสมที่สุด เมื่อนำค่า $\underline{w}_{opt}$ ลบจากทั้ง 2 ข้างของสมการ (2.3.3) จะได้สมการใหม่ในรูปของเวกเตอร์ $\underline{q}_j$ ดังนี้

$$\underline{q}_{j+1} = \left(I - \beta\Phi\right)\underline{q}_{j} \tag{2.3.5}$$

เราสามารถหาค่า $\underline{q}_{j+1}$ ด้วยสมการ (2.3.5) ไปเรื่อย ๆ จนอยู่ในรูปของ $\underline{q}_0$ ซึ่งเป็นค่าเริ่มต้นของเวกเตอร์ $\underline{q}$ ซึ่งจะได้เป็น

$$\underline{q}_{j+1} = \left(I - \beta\Phi\right)^{j}\underline{q}_{0} \tag{2.3.6}$$

ถ้าเวกเตอร์พารามิเตอร์ก้าวเข้าใกล้สู่จุดเหมาะสมที่สุด เวกเตอร์ $\underline{q}_j$ จะต้องเข้าใกล้เวกเตอร์ศูนย์ $\underline{0}$ ดังนั้น สิ่งที่ต้องพิจารณาคือ ในสมการที่ (2.3.6) เวกเตอร์ $\underline{q}_j$ สามารถลู่เข้าใกล้เวกเตอร์ศูนย์หรือไม่?

จากสมการ (2.3.5) และ (2.3.6) เวกเตอร์ $\underline{q}_j$ จะเข้าใกล้เวกเตอร์ศูนย์หรือไม่ ขึ้นกับค่าเฉพาะ (eigenvalue) ของเมตริกซ์อัตสหสัมพันธ์ Φ ซึ่งจะกำหนดเป็น $\lambda_1, \lambda_2, \ldots, \lambda_N$ เมื่อ N เป็นมิติ (dimension) ของเมตริกซ์ Φ ถ้าเมตริกซ์ Φ เป็นแบบบวกแน่นอนจะได้ค่าเฉพาะเป็นจำนวนจริงบวก ถ้าเป็นเมตริกซ์บวกกึ่งแน่นอนจะได้ค่าเฉพาะเป็นจำนวนจริงที่มากกว่าหรือเท่ากับศูนย์เท่านั้น ค่าเฉพาะเหล่านี้ถ้ามีค่าไม่ซ้ำกันจะได้เวกเตอร์เฉพาะ (eigenvector) จากค่าเฉพาะเหล่านี้ N เวกเตอร์ กำหนดให้เป็น $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_N$ และเป็นเวกเตอร์ปกติ (normalized vector) มีขนาดเป็น 1 และตั้งฉาก (orthogonal) ซึ่งกันและกัน ทำให้ผลคูณภายในของเวกเตอร์เป็น

$$\underline{v}_i^{T}\underline{v}_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \tag{2.3.7}$$

เมตริกซ์ Φ สามารถเขียนในรูปผลรวมของเทอมที่มีค่าเฉพาะและเวกเตอร์เฉพาะเหล่านี้ ได้เป็น

$$\Phi = \sum_{i=1}^{N} \lambda_i \underline{v}_i \underline{v}_i^{T} \tag{2.3.8}$$

จากสมการ (2.3.8) เราสามารถเขียนค่าในวงเล็บของสมการ (2.3.6) ได้เป็น

$$\left(I - \beta\Phi\right)^{j} = \sum_{i=1}^{N} \left(1 - \beta\lambda_i\right)^{j} \underline{v}_i \underline{v}_i^{T} \tag{2.3.9}$$

พิจารณาสมการ (2.3.9) และ (2.3.6) จะเห็นว่า เวกเตอร์ $\underline{q}_j$ เกิดจากผลรวมของเทอมที่มีค่าเฉพาะทั้ง N ตัว และแต่ละเทอมจะแปรผันตาม $(1-\beta\lambda_i)^j$ ดังนั้น เทอมที่ลดลงได้เร็วที่สุด คือเทอมของค่าเฉพาะสูงสุด และเทอมที่ลดลงได้ช้าที่สุดจะเป็นเทอมของค่าเฉพาะต่ำสุด ความเร็วในการลู่เข้าสู่จุดเหมาะสมที่สุดจะถูกกำหนดโดยค่า β ถ้าค่า β มากจนทำให้ค่าในวงเล็บมีขนาดมากกว่า 1 เวกเตอร์ $\underline{q}_j$ จะมีค่ามากขึ้นในแต่ละครั้งที่กำหนดเวกเตอร์ใหม่ เช่นเดียวกับที่แสดงไว้ในรูปที่ 2.1 ที่ความยาวก้าวมากเกินไปทำให้ก้าวข้ามจุดเหมาะสมที่สุด และค่าฟังก์ชันเป้าหมายจะเพิ่มขึ้นได้

เราสามารถหาช่วงของค่า β ที่ทำให้ค่าในวงเล็บ $(1-\beta\lambda_i)^j$ น้อยกว่า 1 ได้โดยใช้ค่าเฉพาะมากที่สุดและน้อยที่สุด กำหนดเป็น λ_{max} และ λ_{min} ตามลำดับ เราสามารถตั้งอสมการโดยกำหนดให้ค่าในวงเล็บมีค่าระหว่าง -1 ถึง 1 ดังนี้

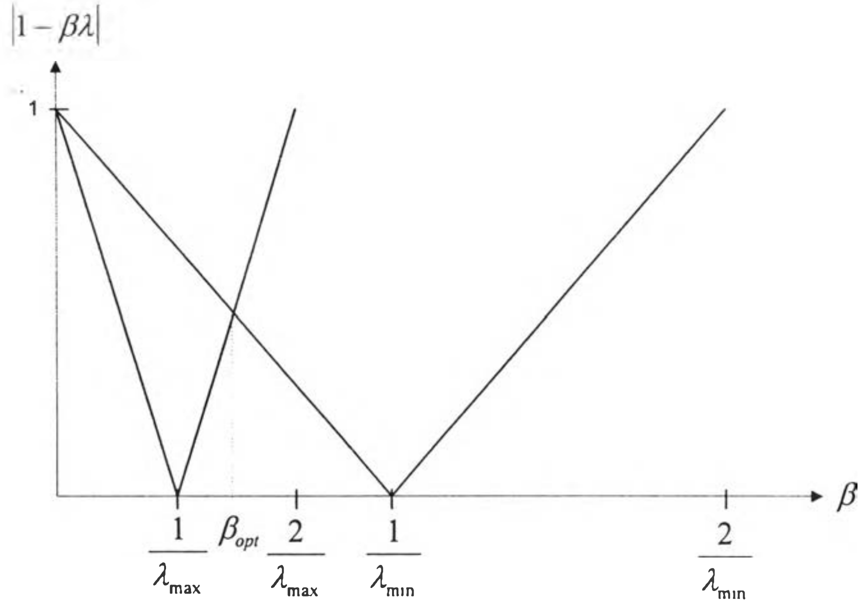
$$-1<(1-\beta\lambda)<1 \tag{2.3.10}$$

หรือ
$$0<\beta<\frac{2}{\lambda_{max}} \tag{2.3.11}$$

λ_{max} ทำให้ได้ขอบเขตแคบที่สุด ดังนั้น ค่า β ที่ทำให้ผลรวมทุกเทอมในสมการ (2.3.9) ลู่เข้าสู่ศูนย์ต้องน้อยกว่า $\frac{2}{\lambda_{max}}$ ดังนั้น (2.3.11) จึงเป็นขอบเขตที่ทำให้ลู่เข้า เมื่อกำหนดให้ β เป็นค่าคงตัว ก็จะมีค่าหนึ่งที่ทำให้ลู่เข้าเร็วที่สุดในช่วง 0 ถึง $\frac{2}{\lambda_{max}}$ การหา β ค่านี้ ให้พิจารณารูปที่ 2.2 ซึ่งเป็นกราฟแสดงขนาดของค่าในวงเล็บ $|1-\beta\lambda_i|$ เทียบกับ β

จากรูปที่ 2.2 การเลือกค่า $\beta=\beta_{opt}$ จะทำให้ลู่เข้าเร็วที่สุด เพราะว่าความยาวก้าวนี้ทำให้ขนาดของค่าในวงเล็บสำหรับเทอมของ λ_{min} และ λ_{max} ในสมการ (2.3.9) เท่ากัน ความเร็วในการลู่เข้าของทั้งสองเทอมก็จะเท่ากัน ค่า β_{opt} นี้หาได้จาก

$$\beta_{opt}=\frac{2}{\lambda_{min}+\lambda_{max}} \tag{2.3.12}$$



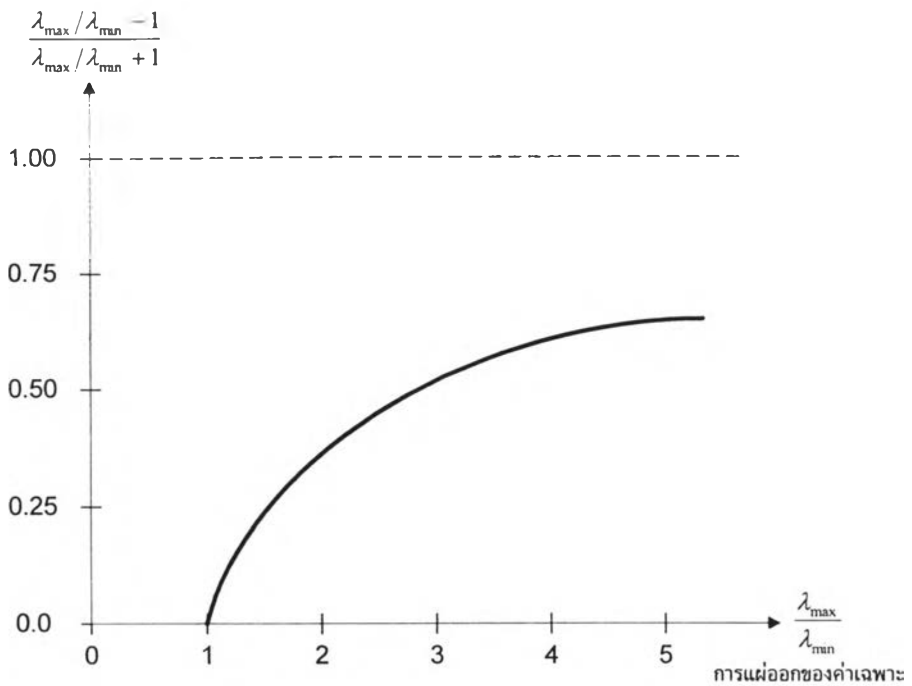
รูปที่ 2.2 กราฟแสดงค่า $|1-\beta\lambda|$ เทียบกับ β

เมื่อแทน β_{opt} นี้ในวงเล็บ จะได้เป็น

$$\left(\left(\tfrac{\lambda_{max}}{\lambda_{min}}-1\right)\Big/\left(\tfrac{\lambda_{max}}{\lambda_{min}}+1\right)\right)^{j} \tag{2.3.13}$$

โดยแทนค่า λ ด้วย λ_{min} เนื่องจากเป็นเทอมที่ลู่เข้าช้าที่สุด จึงเป็นเทอมที่กำหนดความเร็วในการลู่เข้า ในรูปที่ 2.3 แสดงกราฟของค่าในวงเล็บ $\left(\tfrac{\lambda_{max}}{\lambda_{min}}-1\right)\Big/\left(\tfrac{\lambda_{max}}{\lambda_{min}}+1\right)$ เทียบกับอัตราส่วน $\dfrac{\lambda_{max}}{\lambda_{min}}$

อัตราส่วน $\dfrac{\lambda_{max}}{\lambda_{min}}$ ถูกเรียกว่า การแผ่ออกของค่าเฉพาะ (eigenvalue spread) ซึ่งมีค่าได้ตั้งแต่ 1 ถึงไม่จำกัดและมีความสำคัญต่อความเร็วในการลู่เข้าอย่างมาก จากรูปที่ 2.3 เห็นได้ว่าเมื่อค่าอัตราส่วนนี้มากขึ้น ค่าในวงเล็บจะเข้าใกล้ 1 มากขึ้น ทำให้การลู่เข้าสู่ศูนย์ของสมการ (2.3.9) ช้าลง เมื่อการแผ่ออกของค่าเฉพาะมีค่าเป็น 1 จะทำให้ลู่เข้าเร็วที่สุด



รูปที่ 2.3 กราฟแสดงค่า $\left(\left(\frac{\lambda_{max}}{\lambda_{min}}-1\right)/\left(\frac{\lambda_{max}}{\lambda_{min}}+1\right)\right)$ เทียบกับอัตราส่วน $\frac{\lambda_{max}}{\lambda_{min}}$

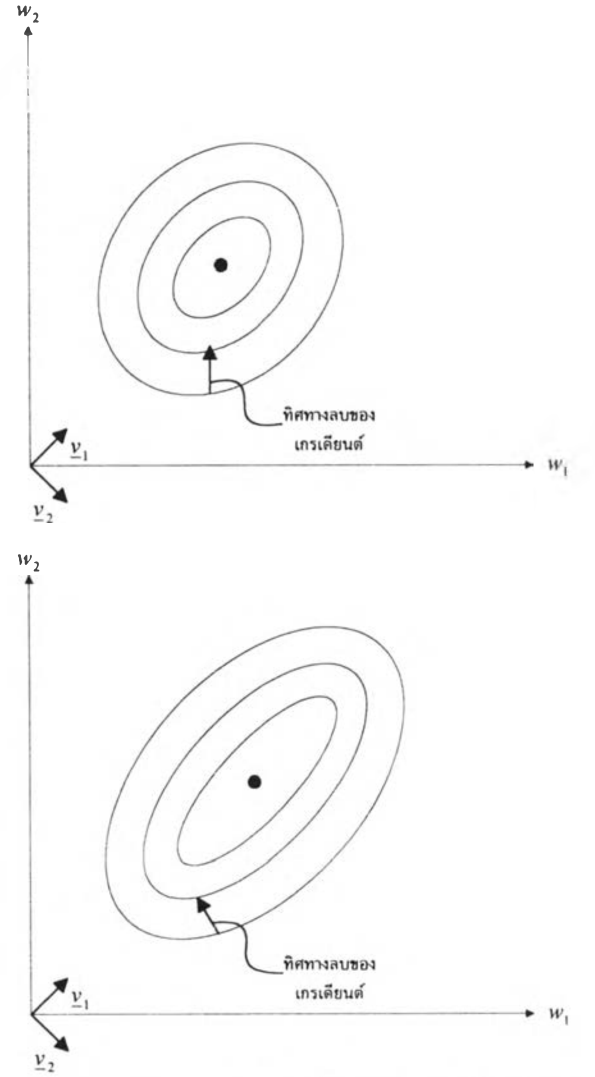
เมื่อการแผ่ออกของค่าเฉพาะมากขึ้นแล้วความเร็วในการลู่เข้าจะน้อยลง สามารถพิจารณาได้จากสมการแสดงค่าความแตกต่างระหว่างความผิดพลาดกำลังสองเฉลี่ยที่เกิดขึ้นในปัจจุบันกับค่าต่ำสุด ซึ่งแสดงเป็นสมการได้ดังนี้

$$E\left[e^2(t)\right]-E\left[e^2(t)\right]_{\min}=\left(\underline{w}-\underline{w}_{opt}\right)^T\Phi\left(\underline{w}-\underline{w}_{opt}\right) \qquad (2.3.14)$$

$$=\sum_{i=1}^{N}\lambda_i\left[\left(\underline{w}-\underline{w}_{opt}\right)^T\underline{v}_i\right]^2 \qquad (2.3.15)$$

จากสมการ (2.3.15) ผลต่างของความผิดพลาดกำลังสองเฉลี่ยเท่ากับผลรวมของขนาดเวกเตอร์ผลต่าง $\underline{w}-\underline{w}_{opt}$ ในทิศทางของเวกเตอร์เฉพาะ $\underline{v}_i;\ 1\le i\le N$ กำลังสอง คูณด้วยค่าเฉพาะ $\lambda_i;\ 1\le i\le N$ ดังนั้น ค่าเฉพาะที่ทำให้เกิดค่าผลต่างมากที่สุดคือ λ_{max} และที่ทำให้ผลต่างน้อยที่สุดคือ λ_{min} หมายความว่า เกิดการเปลี่ยนแปลงมากที่สุดในทิศของ λ_{max} และน้อยที่สุดในทิศของ λ_{min} ดังนั้น ทิศทางของเวกเตอร์เฉพาะที่เกิดจาก λ_{max} จะลู่เข้าสู่จุดเหมาะสมที่สุดได้เร็วที่สุด และทิศทางของเวกเตอร์เฉพาะที่เกิดจาก λ_{min} จะช้าที่สุด ความยาวก้าวสูงสุดถูก

กำหนดจาก λ_{max} เพราะว่าเกรเดียนต์มีขนาดมากที่สุดในทิศทางของเวกเตอร์เฉพาะของ λ_{max} เมื่อความยาวก้าวมากขึ้นการเปลี่ยนแปลงในทิศของเวกเตอร์เฉพาะของ λ_{max} ก็ยิ่งมีผลมากขึ้นกว่าทิศทางอื่น ๆ



รูปที่ 2.4 แสดงลักษณะของคอนทัวร์ที่มีผลต่อการเข้าสู่จุดเหมาะสมที่สุด

เมื่อพิจารณาคอนทัวร์ในกรณีที่มีพารามิเตอร์ 2 ตัว ซึ่งเป็นรูปวงรี โดยแกนหลัก (แกนที่ยาวกว่า) จะอยู่ในแนวเดียวกับเวกเตอร์เฉพาะของ λ_{min} แกนรอง (แกนที่สั้นกว่า) อยู่ในแนวเดียวกับเวกเตอร์เฉพาะของ λ_{max} ดังแสดงในรูปที่ 2.4 ดังนั้น การลู่เข้าสู่จุดเหมาะสมที่สุดให้เร็วควรก้าวไปในทิศทางของเวกเตอร์เฉพาะของ λ_{max} มากกว่าทิศทางของเวกเตอร์เฉพาะของ λ_{min} ในกรณีที่การแผ่ออกของค่าเฉพาะมีค่าสูง คอนทัวร์จะมีลักษณะเป็นวงรีมาก ถ้าค่า β ถูกกำหนด

จาก λ_{max} การก้าวในทิศทางของแกนรองจะไม่ข้ามจุดเหมาะสมที่สุดแน่นอน และ β จะมีผลมากขึ้นเมื่อก้าวไปในทิศของแกนรองมากกกว่าทิศของแกนหลัก เพราะว่า ทำให้ผลต่างของความผิดพลาดกำลังสองเฉลี่ยมากยิ่งขึ้น

พิจารณารูปที่ 2.4 ประกอบกับรูปที่ 2.2 ถ้าจุดเริ่มต้น $\underline{w}_0$ อยู่บนแกนรองของคอนทัวร์ และกำหนดค่า $\beta < 1/\lambda_{max}$ จะทำให้ไม่ก้าวข้ามจุดเหมาะสมที่สุด แต่ละครั้งที่ก้าวก็จะเข้าใกล้มากขึ้นจนกระทั่งถึงจุดเหมาะสมที่สุด ถ้าเลือก $\beta = 1/\lambda_{max}$ จะสามารถเข้าสู่จุดเหมาะสมที่สุดได้ในการก้าวเพียงครั้งเดียว และถ้าเลือก $\beta > 1/\lambda_{max}$ จะทำให้ก้าวข้ามจุดเหมาะสมที่สุดและเกิดการแกว่งไปมา แต่ถ้ายังเลือกให้ $\beta < 2/\lambda_{max}$ ฟังก์ชันเป้าหมายจะลดลงจนถึงจุดเหมาะสมที่สุด ตัวอย่างเช่น เลือกให้ $\beta = \beta_{opt}$ ซึ่งทำให้ก้าวข้ามจุดเหมาะสมที่สุด แต่ยังสามารถเข้าสู่จุดเหมาะสมที่สุดได้เร็วกว่าการก้าวตามทิศทางของแกนหลัก ซึ่งเป็นทิศของเวกเตอร์เฉพาะจาก λ_{min}

2.4 การหาจุดเหมาะสมที่สุดโดยใช้วิธีเกรเดียนต์แบบสโตคาสติกส์

ในหัวข้อ 2.1 ถึง 2.3 เราสามารถหาจุดเหมาะสมที่สุดของการลดระดับตามแนวเกรเดียนต์ได้ เพราะทราบค่าทางสถิติของสัญญาณขาเข้า คือเมตริกซ์อัตสหสัมพันธ์ Φ และเวกเตอร์สหสัมพันธ์ข้าม $\underline{p}$ แต่ค่าทั้งสองจำเป็นต้องใช้การหาค่าเฉลี่ยทั้งชุด (ensemble average) จึงจะได้ค่าที่แท้จริง ซึ่งไม่สามารถทำได้ในการใช้งานส่วนใหญ่ ดังนั้น จึงจำเป็นต้องใช้การหาค่าเฉลี่ยตามเวลา (time average) ซึ่งเป็นเพียงการประมาณแทน ในหัวข้อนี้ กล่าวถึงวิธีเกรเดียนต์แบบสโตคาสติกส์ (Stochastic Gradient) ซึ่งใช้การประมาณดังกล่าวสำหรับปรับพารามิเตอร์

การปรับให้ใช้ค่าเฉลี่ยตามเวลาแทน ถึงแม้จะทำให้สามารถใช้งานได้กว้างขวางกว่า แต่ทำให้ไม่สามารถลู่เข้าจนถึงจุดเหมาะสมที่สุดได้ และยังคงไม่สามารถใช้ในกรณีที่สัญญาณขาเข้ามีค่าทางสถิติเปลี่ยนแปลงตามเวลา (nonstationary) ได้

ในหัวข้อที่ผ่านมา เกรเดียนต์ $\nabla_{\underline{w}}\left\{E\left[e^2(t)\right]\right\}$ มีค่าเป็นไปตามสมการ

$$\nabla_{\underline{w}}\left\{E\left[e^2(t)\right]\right\} = -2E\left[e(t)\underline{x}(t)\right] \tag{2.4.1}$$

การหาค่าของเครื่องหมาย E ซึ่งเป็นค่าเฉลี่ยทั้งชุดไม่สามารถทำได้ ดังนั้น วิธีเกรเดียนต์แบบสโตคาสติกส์จึงไม่ใส่เครื่องหมายนี้ และให้เกรเดียนต์เป็นผลคูณระหว่างค่าผิดพลาด $e(t)$ และ

สัญญาณขาเข้า $\underline{x}(t)$ ซึ่งมีลักษณะเป็นกระบวนการสโตคาสติกส์ ดังนั้น สมการปรับพารามิเตอร์สำหรับวิธีเกรเดียนต์แบบสโตคาสติกส์จะเป็น

$$\underline{w}(t) = \underline{w}(t-1) - \frac{\beta}{2}\nabla_{\underline{w}}\left[e^2(t)\right] \tag{2.4.2}$$

$$= \underline{w}(t-1) + \beta e(t)\underline{x}(t) \tag{2.4.3}$$

$$= \left[I - \beta\underline{x}(t)\underline{x}^T(t)\right]\underline{w}(t-1) + \beta d(t)\underline{x}(t) \tag{2.4.4}$$

นอกจากการใช้ค่าเฉลี่ยตามเวลาแทนแล้ว ข้อแตกต่างจากการลดระดับตามแนวเกรเดียนต์ปกติอีกข้อ คือ การใช้สัญญลักษณ์ของเวลา t แทนสัญญลักษณ์แสดงจำนวนครั้งที่ปรับ j เพราะว่าในวิธีนี้ การคำนวณเกิดขึ้น 1 รอบใน 1 ช่วงเวลา สัญญลักษณ์แสดงจำนวนรอบจึงสามารถแทนได้ด้วยสัญญลักษณ์ของเวลาหรือจำนวนครั้งที่ทำการเก็บข้อมูลเข้ามาในระบบ และใช้สัญญลักษณ์ $\underline{x}(t)\underline{x}^T(t)$ แทนเมตริกซ์ Φ เพราะไม่มีการหาค่าคาดหวัง และใช้ $d(t)\underline{y}(t)$ แทนเวกเตอร์ $\underline{p}$ โดยมีเหตุผลเช่นเดียวกัน

พิจารณา สมการปรับค่าพารามิเตอร์ตัวที่ i ตามวิธีเกรเดียนต์แบบสโตคาสติกส์ ดังนี้

$$w_i(t) = w_i(t-1) + \beta e(t)x_i(t) \tag{2.4.5}$$

จากสมการ (2.4.5) จะเห็นว่า ค่าพารามิเตอร์ในรอบถัดไป เป็นค่าพารามิเตอร์ในรอบปัจจุบันบวกผลคูณของความยาวก้าว β, ค่าผิดพลาด $e(t)$ และสัญญาณขาเข้า $x_i(t)$ ซึ่งเป็นกระบวนการสโตคาสติกส์ ทำให้ผลคูณนี้เป็นกระบวนการสโตคาสติกส์ และเห็นได้ว่า ค่าพารามิเตอร์เป็นผลบวกสะสมตั้งแต่เริ่มต้นของผลคูณนี้ ดังนั้น พารามิเตอร์จึงเป็นกระบวนการสโตคาสติกส์เช่นกัน การพิจารณาค่าของผลคูณหรือพารามิเตอร์จึงต้องใช้ค่าเฉลี่ยมาพิจารณาแทน

ในวิธีนี้ หลักการความตั้งฉาก สามารถแปลความหมายต่างจากที่แสดงในหัวข้อที่ 2.2 ได้ดังนี้ หลักการความตั้งฉากกล่าวว่า เมื่อถึงจุดเหมาะสมที่สุดแล้วผลคูณของค่าผิดพลาดและสัญญาณขาเข้ามีค่าเฉลี่ยเป็นศูนย์ สิ่งที่หลักการเชิงตั้งฉากบอกทางอ้อมเมื่อยังไม่ถึงจุดเหมาะสมที่สุด คือ เครื่องหมายของค่าเฉลี่ยของผลคูณเป็นทิศทางที่จะเคลื่อนไปสู่จุดเหมาะสมที่สุด ซึ่งสามารถอธิบายได้โดยพิจารณาผลคูณระหว่างค่าผิดพลาดและสัญญาณขาเข้า $e(t)x_i(t)$ ส่วนที่มี

ผลของค่าพารามิเตอร์ก็คือ $-w_i(t-1)x_i^2(t)$ (แทนค่า $e(T)$ ด้วย $d(t)-\sum_{k=1}^{N} w_k(t)x_k(t)$ ในสมการที่ (2.4.5) และพิจารณาเฉพาะเทอมที่มีผลของพารามิเตอร์ $w_i(t)$) ถ้าพารามิเตอร์ $w_i(t-1)$ เป็นบวก เทอมนี้จะเป็นลบมากขึ้น ทำให้ค่าเฉลี่ยตามเวลาของพารามิเตอร์ของ $w_i(t)$ น้อยกว่า $w_i(t-1)$ และถ้าพารามิเตอร์ $w_i(t-1)$ เป็นลบ เทอมนี้ก็จะเป็นบวก ค่าเฉลี่ยของ $w_i(t)$ ก็จะมากขึ้น เพื่อให้ผลเฉลี่ยเข้าสู่ศูนย์

การศึกษาถึงการลู่เข้าที่เป็นประโยชน์อันหนึ่งคือ การสมมติให้สัญญาณขาเข้าเป็นกระบวนการสโตคาสติกส์คงที่แบบกว้าง (wide-sense stationary) โดยทราบค่าทางสถิติบางตัว (ฟังก์ชันสหสัมพันธ์ เป็นต้น) ซึ่งคาดว่าเวกเตอร์พารามิเตอร์สามารถลู่เข้าได้จากการหาค่าที่เหมาะสมโดยอาศัยค่าสถิติที่ทราบ ถึงแม้ความเร็วในการลู่เข้า (speed of convergence) ของเงื่อนไขนี้ไม่สามารถใช้ได้เมื่อค่าสถิติของสัญญาณขาเข้าเปลี่ยนแปลงตามเวลา (nonstationary) แต่ก็สามารถชี้ให้เห็นถึงประสิทธิภาพการลู่เข้าได้อย่างดี

เมื่อสัญญาณขาเข้าเป็นกระบวนการสโตคาสติกส์คงที่แบบกว้าง เราสามารถหาค่าคาดหวังของสมการ (2.4.4) ได้เป็น

$$E\left[\underline{w}(t)\right] = E\left[\left(I-\beta\underline{x}(t)\underline{x}^T(t)\right)\underline{w}(t-1)\right] + \beta E\left[d(t)\underline{x}(t)\right] \tag{2.4.6}$$

$$\approx \left(I-\beta\Phi\right)E\left[\underline{w}(t-1)\right] + \beta\underline{p} \tag{2.4.7}$$

ซึ่งใช้การประมาณว่า เวกเตอร์ของพารามิเตอร์และสัญญาณขาเข้าไม่มีสหสัมพันธ์ต่อกันได้ เพราะว่า เวกเตอร์พารามิเตอร์มีการเปลี่ยนแปลงไปแบบช้าๆ ซึ่งเป็นผลจาก β ที่มีค่าน้อย หลังจากประมาณแล้ว สมการที่ (2.4.6) จะเหมือนกับการลดระดับตามแนวเกรเดียนต์ปกติ การลู่เข้าจะเหมือนกับที่แสดงในหัวข้อที่ 2.3 แต่ค่าของเวกเตอร์พารามิเตอร์ไม่ได้ก้าวเข้าสู่จุดเหมาะสมที่สุดจริงๆ เพียงแต่ค่าเฉลี่ยของเวกเตอร์พารามิเตอร์และค่าเฉลี่ยอื่นๆ เท่านั้นที่ไปถึงจุดเหมาะสมที่สุด ซึ่งความจริงเวกเตอร์พารามิเตอร์กระเพื่อมรอบๆ จุดเหมาะสมที่สุด เพราะว่ายังมีเทอมผลคูณในสมการ (2.4.5) ซึ่งเป็นกระบวนการสโตคาสติกส์อยู่ โดยค่า β ยิ่งมากก็ยิ่งทำให้ผลจากเทอมนี้มีมากขึ้น การกระเพื่อมก็จะมากขึ้น

การประมาณที่ใช้จะใกล้เคียงความจริงเมื่อ β มีค่าน้อย โดยปกติ จะต้องน้อยกว่าค่า β_{opt} ที่กล่าวไว้ในหัวข้อที่ 2.3 มาก ซึ่งเป็นผลให้ความเร็วในการลู่เข้าน้อยลงด้วย การเพิ่มความ

ยาวก้าวจะสามารถเพิ่มความเร็วในการลู่เข้า แต่ว่าการกระเพื่อมที่เกิดขึ้นจะมากขึ้นตามไปด้วย ถ้าเพิ่มความยาวก้าวมากขึ้นเรื่อย ๆ เมื่อเริ่มแรกจะสามารถเพิ่มความเร็วในการลู่เข้าได้ แต่ถ้าเพิ่มขึ้นอีก ความเร็วในการลู่เข้าก็จะลดลงและขาดเสถียรภาพในที่สุด ดังนั้น การเลือกความยาวก้าวให้มีค่ามากควรอยู่ในขอบเขตที่ทำให้ความเร็วในการลู่เข้าเพิ่มขึ้น

อาศัยการประมาณนี้ ขั้นตอนวิธีสามารถปรับพารามิเตอร์เพื่อให้เข้าสู่จุดเหมาะสมที่สุดได้ แต่จะมีการกระเพื่อมซึ่งเกิดขึ้นจากการใช้ β เป็นค่าคงตัว ถ้าเปลี่ยน β ให้มีค่าเปลี่ยนตามเวลาจะสามารถลู่เข้าสู่จุดเหมาะสมที่สุดได้ แต่ต้องตั้งเงื่อนไขสำหรับเปลี่ยนขนาดอย่างระมัดระวัง เพราะอาจทำให้ไม่ลู่เข้าได้ ปกติ β ควรมีค่ามากในตอนเริ่ม เพื่อให้สามารถลู่เข้าได้เร็ว เมื่อเข้าใกล้จึงลดขนาดลงเรื่อย ๆ เพื่อให้ก้าวเข้าสู่จุดเหมาะสมที่สุดได้ นั่นคือ β มีค่าเข้าใกล้ศูนย์เมื่อเข้าใกล้จุดเหมาะสมที่สุด การลดค่า β ก่อนอาจทำให้ไม่สามารถไปถึงจุดเหมาะสมที่สุดได้เช่นกัน ดังนั้น การกำหนด β ให้เปลี่ยนตามเวลาเพื่อให้เข้าสู่จุดเหมาะสมที่สุด จึงควรจะเป็นไปตามสมการ

$$\sum_{i=0}^{\infty} \beta(t) = \infty \tag{2.4.8}$$

และ

$$\sum_{i=0}^{\infty} \beta^2(t) < \infty \tag{2.4.9}$$

ถ้าสัญญาณขาเข้าเปลี่ยนแปลงตามเวลา การประมาณนี้จะไม่สามารถทราบได้ว่า สัญญาณเปลี่ยนแปลงไปเช่นไร แต่ถ้าการเปลี่ยนแปลงยังไม่เร็วนัก ขั้นตอนวิธีก็อาจสามารถปรับพารามิเตอร์ให้ลู่เข้าสู่จุดเหมาะสมที่สุดได้ แต่ถ้าการเปลี่ยนแปลงเร็ว ขั้นตอนวิธีจะไม่สามารถปรับพารามิเตอร์ได้ทัน การปรับค่า β ให้เปลี่ยนตามเวลาทำให้ขั้นตอนวิธีลู่เข้าสู่จุดเหมาะสมที่สุดเร็วขึ้นก็อาจช่วยแก้ไขปัญหานี้ได้ดีขึ้น

2.5 การลู่เข้าโดยวิธีเกรเดียนต์แบบสโตคาสติกส์

วิธีเกรเดียนต์แบบสโตคาสติกส์สามารถทำให้การประยุกต์ใช้งานกว้างขวางขึ้น เพราะว่าสามารถใช้ได้โดยไม่จำเป็นต้องทราบค่าทางสถิติของสัญญาณขาเข้า แต่ก็ทำให้ค่าพารามิเตอร์กระเพื่อมรอบ ๆ จุดเหมาะสมที่สุดโดยไม่สามารถเข้าสู่จุดเหมาะสมที่สุดอย่างแท้จริงได้ แต่ผลจากการใช้เกรเดียนต์แบบสโตคาสติกส์แทนยังทำให้พารามิเตอร์ที่ปรับมีลักษณะเป็นกระบวนการสโตคาสติกส์ เป็นผลให้ช่วงความยาวก้าวที่ทำให้ลู่เข้าแคบลงกว่าที่เคยแสดงไว้ในหัวข้อที่ 2.3 โดย

ความยาวก้าวจะต้องทำให้ความแปรปรวนของพารามิเตอร์ลู่เข้าด้วย ในหัวข้อนี้ จะแสดงลักษณะการลู่เข้าของพารามิเตอร์ในวิธีเกรเดียนต์แบบสโตคาสติกส์ และหาช่วงของความยาวก้าวที่ทำให้ลู่เข้าของวิธีนี้

การปรับพารามิเตอร์ในวิธีเกรเดียนต์แบบสโตคาสติกส์ จะเป็นไปตามสมการ

$$\underline{w}(t) = \underline{w}(t-1) + \beta e(t)\underline{x}(t) \tag{2.5.1}$$

$$= \left[I - \beta \underline{x}(t)\underline{x}^T(t)\right]\underline{w}(t-1) + \beta d(t)\underline{x}(t) \tag{2.5.2}$$

2.5.1 ค่าเฉลี่ยของเวกเตอร์พารามิเตอร์

เนื่องจาก ค่าพารามิเตอร์ในสมการที่ (2.5.1) และ (2.5.2) เป็นกระบวนการสโตคาสติกส์ ค่าคาดหวังของสมการทั้งสอง จะเป็น

$$E\left[\underline{w}(t)\right] = E\left[\underline{w}(t-1)\right] + \beta E\left\{\left[d(t) - \underline{x}^T(t)\underline{w}(t-1)\right]\underline{x}(t)\right\} \tag{2.5.3}$$

$$= E\left[\left(I - \beta \underline{x}(t)\underline{x}^T(t)\right)\underline{w}(t-1)\right] + \beta E\left[d(t)\underline{x}(t)\right] \tag{2.5.4}$$

โดยปกติ การหาค่าคาดหวังเทอมทางด้านขวามือทำได้ยาก เนื่องจากเวกเตอร์พารามิเตอร์ขึ้นอยู่กับสัญญาณขาเข้า และเป็นฟังก์ชันไม่เชิงเส้นของเวกเตอร์ในอดีต แต่ถ้าตั้งสมมติฐานให้เวกเตอร์พารามิเตอร์และสัญญาณขาเข้าไม่มีสหสัมพันธ์กัน โดยพิจารณาแยกกันในแต่ละรอบ ซึ่งมีเหตุผลสนับสนุนจากการกำหนดให้ β มีค่าน้อยทำให้เวกเตอร์ $\underline{w}(t)$ ไม่กระเพื่อมห่างจากจุดเหมาะสมที่สุดมากนัก β ยิ่งมีค่าน้อยผลจากสัญญาณขาเข้าต่อเวกเตอร์พารามิเตอร์ก็จะยิ่งน้อยการประมาณจะใกล้เคียงมากขึ้น

เมื่อใช้การประมาณแล้ว สมการ (2.5.4) จะสามารถเขียนใหม่ ได้เป็น

$$E\left[\underline{w}(t)\right] \approx \left(1 - \beta\Phi\right)E\left[\underline{w}(t-1)\right] + \beta \underline{p} \tag{2.5.5}$$

โดยกำหนดให้ $\Phi \equiv E\left[\underline{x}(t)\underline{x}^T(t)\right]$ เป็นเมตริกซ์อัตสหสัมพันธ์ที่มีขนาดเป็น NxN และ $\underline{p} \equiv E\left[d(t)\underline{x}(t)\right]$ เป็นเวกเตอร์สหสัมพันธ์ข้าม (cross-correlation) ที่มีขนาดเป็น Nx1 เมตริกซ์ Φ จะเป็นเมตริกซ์สมมาตร และถ้ากำหนดให้เป็นเมตริกซ์แบบบวกแน่นอน จะสามารถแยกเป็นผลคูณเมตริกซ์ได้ดังนี้

$$\Phi = V\Lambda V^T \tag{2.5.6}$$

เมื่อ Λ เป็นเมตริกซ์แนวทแยงของค่าเฉพาะ แทนด้วยสัญญลักษณ์ได้ดังนี้

$$\Lambda = diag\left[\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_N\right] \tag{2.5.7}$$

และ V เป็นเมตริกซ์ที่ประกอบด้วยเวกเตอร์เฉพาะ $\underline{v}_i$ ที่ได้จากค่าเฉพาะ λ_i ของเมตริกซ์ Φ ที่คอลัมน์ i เวกเตอร์ $\underline{v}_i$ จะมีคุณสมบัติตามที่กล่าวไว้ในสมการ (2.3.7) ซึ่งทำให้ $V^{-1} = V^T$

เราจะกำหนดเวกเตอร์ผลต่าง $\underline{q}(t)$ และทำการแปลงให้ไปอยู่ในทิศทางของเวกเตอร์เฉพาะ เพื่อให้เทอมที่เป็นผลคูณข้าม (cross term) มีค่าเป็นศูนย์ ดังนี้

$$\underline{q}(t) \equiv \underline{w}(t) - \underline{w}_{opt} \tag{2.5.8}$$

$$\underline{\tilde{q}}(t) \equiv V^T \underline{q}(t) \tag{2.5.9}$$

โดยที่ $\underline{\tilde{q}}(t)$ เป็นเวกเตอร์ที่แปลงแล้ว เวกเตอร์สัญญาณขาเข้า $\underline{x}(t)$ จะแปลงในลักษณะเดียวกัน ดังนี้

$$\underline{\tilde{x}}(t) = V^T \underline{x}(t) \tag{2.5.10}$$

และเวกเตอร์ $\underline{w}_{opt}$ เป็นเวกเตอร์พารามิเตอร์ที่จุดเหมาะสมที่สุด ดังแสดงในหัวข้อที่ 2.1

$$\underline{w}_{opt} = \Phi^{-1}\underline{p} \tag{2.5.11}$$

เช่นเดียวกัน ถ้าต้องการแปลงส่วนประกอบที่ i ของเวกเตอร์ สามารถทำได้โดยการคูณเวกเตอร์เดิมที่มีส่วนประกอบนั้นด้วยเวกเตอร์เฉพาะ $\underline{v}_i^T$ ทางด้านหน้า ส่วนประกอบของเวกเตอร์ที่แปลงไปแล้วจะไม่มีสหสัมพันธ์ต่อกัน ผลจากการแปลงจะได้ว่า

$$E\left[\tilde{\underline{x}}(t)\tilde{\underline{x}}^T(t)\right] = \Lambda \tag{2.5.12}$$

จากสมการ (2.5.4) คูณ V^T เข้าทางด้านซ้ายทั้งสองข้างของสมการ ได้เป็น

$$E\left[\tilde{\underline{q}}(t)\right] \approx \left[I - \beta\Lambda\right]E\left[\tilde{\underline{q}}(t-1)\right] \tag{2.5.13}$$

เนื่องจาก Λ เป็นเมตริกซ์แนวทแยง (diagonal matrix) สมการ (2.5.13) จะสามารถเขียนเป็นสมการของส่วนประกอบในเวกเตอร์ $\tilde{\underline{q}}(t)$ ได้ดังนี้

$$E\left[\tilde{q}_i(t)\right] \approx \left(1 - \beta\lambda_i\right)^t \tilde{q}_i(0) \tag{2.5.14}$$

เมื่อ $\tilde{q}_i(t)$ เป็นส่วนประกอบที่ i ของ $\tilde{\underline{q}}(t)$

ค่าเฉลี่ยของเวกเตอร์ผลต่าง $E\left[\underline{q}(t)\right] = VE\left[\tilde{\underline{q}}(t)\right]$ จะลู่เข้าสู่เวกเตอร์ศูนย์ในแบบเอ็กซ์โปเนนเชียล (exponential) ตามค่าเฉพาะทั้ง N โมดคูณด้วยความยาวก้าว β ซึ่งจะถูกเลือกให้อยู่ระหว่าง 0 ถึง $2/\lambda_{\max}$ โดย $\lambda_{\max}$ เป็นค่าเฉพาะมากที่สุดของเมตริกซ์อัตสหสัมพันธ์ Φ จากสมการ (2.5.14) ค่าคงตัวทางเวลา (time constant) ที่เกิดจากค่าเฉพาะที่ i (เวลาที่ $E\left[q_i(t)\right]$ ใช้ในการลดลงเป็น $\frac{1}{e}q_i(0)$) จะเป็น

$$\tau_i \approx \frac{1}{\ln\left(1 - \beta\lambda_i\right)} \approx \frac{1}{\beta\lambda_i} \tag{2.5.15}$$

ดังนั้น เมื่อ β มีค่าน้อย สมการ (2.5.5) ซึ่งแสดงค่าเฉลี่ยของเวกเตอร์พารามิเตอร์ สามารถเขียนใหม่ได้ ดังนี้

$$E\left[\underline{w}(t)\right] \approx \underline{w}_{opt} + V\left(I - \beta\Lambda\right)^t \tilde{\underline{q}}(0) \tag{2.5.16}$$

ค่าเฉลี่ยเวกเตอร์พารามิเตอร์ $E[\underline{w}(t)]$ จะลู่เข้าหาเวกเตอร์ $\underline{w}_{opt}$ ดังที่กล่าวไว้ในหัวข้อที่ 2.3 ดังนั้น เทอมที่ 2 ทางด้านขวามือของสมการ (2.5.16) จะลู่เข้าสู่ศูนย์ โดยที่ β มีค่าน้อยและความเร็วในการลู่เข้าถูกกำหนดด้วยค่าเฉพาะน้อยที่สุด $\lambda_{\min}$ ซึ่งทำให้เกิดค่าคงตัวทางเวลามากที่สุด

2.5.2 ความผิดพลาดกำลังสองเฉลี่ย

ความผิดพลาดกำลังสองเฉลี่ย สามารถเขียนในรูปของเวกเตอร์พารามิเตอร์ $\underline{w}(t)$ ได้เป็น

$$\varepsilon(t) = E\left[d^2(t)\right] - 2\underline{w}^T(t)\underline{p} + \underline{w}^T(t)\Phi\underline{w}(t)$$

$$= E\left[d^2(t)\right] - \underline{w}_{opt}{}^T\Phi\underline{w}_{opt} + 2\underline{w}^T(t)\left[\Phi\underline{w}_{opt} - \underline{p}\right] + \left[\underline{w}(t) - \underline{w}_{opt}\right]^T\Phi\left[\underline{w}(t) - \underline{w}_{opt}\right] \qquad (2.5.17)$$

โดยที่ฟังก์ชัน $\varepsilon(t)$ เป็นความผิดพลาดกำลังสองเฉลี่ย ซึ่งเป็นฟังก์ชันเป้าหมายของการหาจุดเหมาะสมที่สุด โดยสมการ (2.5.17) ใช้สมมติฐานว่า เวกเตอร์พารามิเตอร์ไม่ขึ้นกับสัญญาณขาเข้า และจากสมการ (2.5.11) $\underline{w}_{opt} = \Phi^{-1}\underline{p}$ เราสามารถเขียนสมการ (2.5.17) ได้เป็น

$$\varepsilon(t) = \varepsilon_{\min} + \underline{q}^T(t)\Phi\underline{q}(t) \qquad (2.5.18)$$

โดยที่

$$\varepsilon_{\min} = E\left[d^2(t)\right] - \underline{p}^T\Phi\underline{p} \qquad (2.5.19)$$

ซึ่งเป็นความผิดพลาดกำลังสองเฉลี่ยต่ำสุด ถ้าเราหาค่าเฉลี่ยของสมการ (2.5.17) แล้วใช้การแปลงเวกเตอร์ไปอยู่ในแนวของเวกเตอร์เฉพาะในสมการ (2.5.8) และ (2.5.9) จะได้สมการความผิดพลาดกำลังสองเฉลี่ยที่เกินจากค่าต่ำสุด $\varepsilon_{ex}(t)$ จากการปรับพารามิเตอร์เป็น

$$\varepsilon_{ex}(t) = E\left[\underline{\tilde{q}}(t)\Lambda\underline{\tilde{q}}(t)\right] = \sum_{k=1}^{N}\lambda_k E\left[\tilde{q}_k^2(t)\right] \qquad (2.5.20)$$

ถ้าไม่หาค่าคาดหวังทางด้านขวา สมการจะเหมือนในหัวข้อที่ 2.3 ซึ่งเป็นความผิดพลาดกำลังสองเฉลี่ยของการลดระดับตามแนวเกรเดียนต์ตามปกติ แต่ในกรณีของวิธีเกรเดียนต์แบบสโตคาสติกส์ เราต้องหาค่าเฉลี่ยทั้งชุดของพารามิเตอร์ สำหรับการหาค่า $\varepsilon_{ex}(t)$ ในรูปของค่าสถิติอันดับสองของสัญญาณขาเข้า เราจะกำหนดเมตริกซ์ความแปรปรวนของเวกเตอร์ผลต่างซึ่งได้แปลงไปอยู่ในแนวของเวกเตอร์เฉพาะแล้ว ดังนี้

$$Q(t) \equiv E\left[\underline{\tilde{q}}(t)\underline{\tilde{q}}^T(t)\right] \tag{2.5.21}$$

โดยใช้นิยามนี้ สมการ (2.5.20) สามารถเขียนใหม่ได้เป็น

$$\varepsilon_{ex} = trace\left\{\Lambda Q(t)\right\} \tag{2.5.22}$$

การแปลงสมการให้อยู่ในรูปของ $Q(t)$ เราจะใช้เวกเตอร์พารามิเตอร์ที่จุดเหมาะสมที่สุด $\underline{w}_{opt}$ ลบจากสองข้างของสมการ (2.5.2) และแปลงเวกเตอร์ของสมการ (2.5.8) ถึง (2.5.10) ให้ไปอยู่ในแนวเวกเตอร์เฉพาะ ได้เป็น

$$\begin{aligned}\underline{\tilde{q}}(t) &= \underline{\tilde{q}}(t-1) + \beta e(t)\underline{\tilde{x}}(t) \\ &= \left[I - \beta\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right]\underline{\tilde{q}}(t-1) + \beta e_{opt}(t)\underline{\tilde{x}}(t)\end{aligned} \tag{2.5.23}$$

โดยที่ความผิดพลาดที่เกิดจากเวกเตอร์พารามิเตอร์ที่จุดเหมาะสมที่สุด $e_{opt}(t)$ เป็น

$$e_{opt}(t) \equiv d(t) - \underline{w}_{opt}^T\,\underline{x}(t) \tag{2.5.24}$$

แล้วทำการแปลงสมการ (2.5.23) เพื่อให้อยู่ในรูปของเมตริกซ์ความแปรปรวนของเวกเตอร์ผลต่าง $Q(t) = \left[\underline{w}(t) - \underline{w}_{opt}\right]\left[\underline{w}(t) - \underline{w}_{opt}\right]^T$ ซึ่งจะได้เป็น

$$\begin{aligned}\underline{\tilde{q}}(t)\underline{\tilde{q}}^T(t) &= \left[I - \beta\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right]\underline{\tilde{q}}(t-1)\underline{\tilde{q}}^T(t-1)\left[I - \beta\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right] \\ &\quad + 2\beta\left[I - \beta\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}(t)\right)\right]\left(\underline{\tilde{q}}(t-1)\underline{\tilde{x}}^T(t)\right)e_{opt}(t) + \beta^2 e_{opt}^2(t)\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\end{aligned} \tag{2.5.25}$$

สมการนี้จะต้องหาค่าคาดหวังเพื่อให้อยู่ในรูปของเมตริกซ์ความแปรปรวนของเวกเตอร์ผลต่าง $Q(t)$ และต้องใช้การประมาณค่าหลายอย่างเพื่อให้ได้ผลลัพธ์ของการเฉลี่ยนี้ เมื่อใช้สมมติฐานว่า เวกเตอร์พารามิเตอร์และสัญญาณขาเข้าไม่มีสหสัมพันธ์กัน จะได้ค่าประมาณของเทอมแรกด้านขวามือในสมการ (2.5.25) เป็น

$$
\begin{aligned}
&E\left\{\left(I-\beta\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right)\underline{\tilde{q}}(t-1)\underline{\tilde{q}}^T(t-1)\left(I-\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right)\right\} \\
&\qquad\approx Q(t-1)\left(I-2\beta\Lambda\right)+\beta^2 E\left[\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)Q(t-1)\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right] \qquad (2.5.26)
\end{aligned}
$$

โดยที่ $\tilde{x}_i(t)$ เป็นส่วนประกอบที่ i ของเวกเตอร์ $\underline{\tilde{x}}(t)$ และ $\tilde{q}_i(t)$ เป็นส่วนประกอบที่ i ของเวกเตอร์ $\underline{\tilde{q}}(t)$ ค่าประมาณที่ l,m ของเมตริกซ์ $\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)$ ในเทอมสุดท้ายของสมการ (2.5.28) เป็น

$$
\begin{aligned}
&\left\{E\left[\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)Q(t-1)\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right]\right\}_{l,m} \\
&\qquad= E\left(\left[\tilde{x}_l\underline{\tilde{x}}^T(t)\right]Q\left[\tilde{x}_m\underline{\tilde{x}}(t)\right]\right) \\
&\qquad= \sum_{k=1}^{N}\sum_{i=1}^{N}E\left[\tilde{q}_k\tilde{q}_i\right]E\left[\tilde{x}_l\tilde{x}_k\tilde{x}_i\tilde{x}_m\right] \\
&\qquad\approx \sum_{i=1}^{N}E\left[\tilde{q}_i^2\right]E\left[\tilde{x}_l^2\right]E\left[\tilde{x}_i^2\right]\delta_{lm} \\
&\qquad= \sum_{i=1}^{N}E\left[\tilde{q}_i^2\right]\lambda_l\lambda_i\delta_{lm} \\
&\qquad= \delta_{lm}\lambda_l\, trace\left[\Lambda Q(t-1)\right] \qquad (2.5.27)
\end{aligned}
$$

เมื่อ $\delta_{lm}=\begin{cases}1; & x_i=x_l,x_m;\ \lambda_i=\lambda_l,\lambda_m \\ 0; & others\end{cases} \qquad 1\le i\le N$

ค่าสถิติอันดับสี่ของเทอมสุดท้ายในสมการ (2.5.26) จะถูกประมาณด้วยค่าสถิติอันดับสองเมตริกซ์ที่ได้ในสมการ (2.5.27) เป็นเมตริกซ์แนวทแยง ซึ่งค่าที่อยู่นอกแนวทแยงถูกประมาณเป็นศูนย์ เพราะว่าสหสัมพันธ์ระหว่างคู่ส่วนประกอบของเวกเตอร์ผลต่างที่ถูกแปลงแล้วคือ $\tilde{q}_k$ และ $\tilde{q}_l$ มีค่าน้อยมาก เช่นเดียวกับ ค่านอกแนวทแยงของเมตริกซ์ค่าสถิติอันดับสี่ของสัญญาณขาเข้า $\tilde{x}_i, 1\le i\le N$ ที่ไม่มีสหสัมพันธ์กัน

การหาเทอมที่ 2 ทางด้านขวามือของสมการ (2.5.25) ทำได้โดยใช้สมมติฐานว่า เวกเตอร์พารามิเตอร์ไม่ขึ้นกับสัญญาณขาเข้าเช่นกัน ซึ่งได้เป็น

$$E\left\{2\beta\left(I-\beta\left(\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right)\right)\left[\underline{\tilde{q}}(t-1)\underline{\tilde{x}}^T(t)\right]e_{opt}(t)\right\}$$
$$\approx 2\beta E\left[\underline{\tilde{q}}(t-1)\right]\left\{E\left[e_{opt}(t)\underline{\tilde{x}}^T(t)\right]-\beta E\left[\left(\underline{\tilde{x}}^T(t)\right)\right]-\beta E\left[\left(\underline{\tilde{x}}^T(t)\underline{\tilde{x}}(t)\right)e_{opt}(t)\underline{\tilde{x}}^T(t)\right]\right\} \tag{2.5.28}$$

เนื่องจาก $E[\underline{w}(t)]$ ลู่เข้าหา $\underline{w}_{opt}$ และเป็นไปตามสมการ (2.5.7) และ (2.5.8) ดังนั้นค่า $E\left[\underline{\tilde{q}}(t-1)\right]$ จะลู่เข้าหาเวกเตอร์ศูนย์ หรือมีค่าน้อยมากเมื่อ t มีค่ามาก และหลักการเชิงตั้งฉากได้แสดงว่า

$$E\left[e_{opt}(t)\underline{\tilde{x}}(t)\right]=\underline{0} \tag{2.5.29}$$

ดังนั้น เทอมที่แสดงในสมการ (2.5.28) ควรจะมีค่าน้อยมากเมื่อเทียบกับเทอมที่แสดงในสมการ (2.5.27) จึงประมาณเทอมนี้เป็นศูนย์ ส่วนเทอมสุดท้ายในสมการ (2.5.25) จะมีค่าเป็น

$$E\left[\beta^2 e_{opt}^2(t)\underline{\tilde{x}}(t)\underline{\tilde{x}}^T(t)\right]\approx\beta^2\varepsilon_{\min}\Lambda \tag{2.5.30}$$

ซึ่งเป็นการประมาณเนื่องจากการใช้สมมติฐานว่าเวกเตอร์พารามิเตอร์และสัญญาณขาเข้าไม่มีสหสัมพันธ์กัน

เมื่อรวมทุกเทอมในสมการ (2.5.27) ถึง (2.5.30) จะได้เป็น

$$Q(t)\approx Q(t-1)\left(I-2\beta\Lambda\right)+\beta^2\Lambda trace\left[\Lambda Q(t-1)\right]+\beta^2\varepsilon_{\min}\Lambda \tag{2.5.31}$$

เมตริกซ์ $Q(t)$ ในสมการ (2.5.31) เป็นเมตริกซ์แนวทแยง และค่านอกแนวทแยงเป็นศูนย์ เราสามารถกำหนดเวกเตอร์ $\underline{s}(t)$ ซึ่งมีส่วนประกอบที่ i เป็น $E\left[\tilde{q}_i^2(t)\right]$ (ความแปรปรวนของผลต่างพารามิเตอร์จากจุดเหมาะสมที่สุด) ในสมการ (2.5.34) ได้เป็น

$$\underline{s}(t)\approx A\underline{s}(t-1)+\beta^2\varepsilon_{\min}\underline{\lambda} \tag{2.5.32}$$

เมื่อ $\underline{\lambda}$ เป็นเวกเตอร์ค่าเฉพาะที่มีขนาดเท่ากับ N ส่วนค่าในเมตริกซ์ A ซึ่งมีขนาด NxN จะมีองค์ประกอบ (element) ที่ l,m เป็น

$$A_{lm} = \left(1 - 2\beta\lambda_l\right)\delta_{lm} + \beta^2 \lambda_l \lambda_m \tag{2.5.33}$$

โดย δ_{lm} มีค่าเป็น 1 เมื่อค่าเฉพาะ $\lambda = \lambda_l, \lambda_m$ ถ้าไม่ใช่จะมีค่าเป็น 0 เราสามารถเขียน A ในรูปสมการของเมตริกซ์ ได้เป็น

$$A = \beta^2 \underline{\lambda}\underline{\lambda}^T + \left(I - 2\beta\Lambda\right) \tag{2.5.34}$$

สมการ (2.5.20) ได้แสดงความสัมพันธ์ระหว่างความผิดพลาดที่มากเกินจากค่าต่ำสุด $\left(\varepsilon_{ex}\right)$ และเวกเตอร์ $\underline{s}(t)$ ส่วนสมการ (2.5.32) แสดงลักษณะพลวัตของเวกเตอร์ $\underline{s}(T)$ ซึ่งขึ้นกับค่าเฉพาะของเมตริกซ์ A โดยใช้สัญญลักษณ์ μ_i, $1 \le i \le N$ แทน เนื่องจาก A เป็นเมตริกซ์สมมาตร และ β มีค่าน้อยพอที่จะทำให้องค์ประกอบที่อยู่ใน A มีค่าเป็นบวก ดังนั้น A จะสามารถแยกเป็นผลคูณเมตริกซ์ได้ในแบบเดียวกับเมตริกซ์อัตสหสัมพันธ์ Φ ดังนี้

$$A = UMU^T \tag{2.5.35}$$

เมื่อ M เป็นเมตริกซ์แนวทแยงสำหรับค่าเฉพาะ μ_i; $1 \le i \le N$ ของ A และคอลัมน์ที่ i ของ U เป็นเวกเตอร์ตั้งฉากปกติ (orthonormal vector) $\underline{u}_i$; $1 \le i \le N$ ที่ได้จากค่าเฉพาะ μ_i และทำการแปลงเวกเตอร์ $\underline{s}(t)$ ให้ไปอยู่ในแนวเวกเตอร์เฉพาะ $\underline{u}_i$; $1 \le i \le N$ ดังนี้

$$\underline{\tilde{s}}(t) \equiv U^T \underline{s}(t)$$

และใช้แทนเวกเตอร์ $\underline{s}(T)$ ในสมการ (2.5.32) เพื่อให้เทอมคูณกันระหว่างส่วนประกอบมีค่าเป็นศูนย์ จะได้สมการเป็น

$$\underline{\tilde{s}}(t) = M\underline{\tilde{s}}(t-1) + \beta^2 \varepsilon_{\min} U^T \underline{\lambda} \tag{2.5.36}$$

สมการ (2.5.36) นี้ (ได้แสดงไว้แตกต่างจากต้นฉบับ) สามารถแปลความหมายได้เช่นเดียวกับสมการค่าเฉลี่ยของเวกเตอร์ผลต่างในสมการ (2.5.13) คือ ในการก้าวแต่ละครั้ง ส่วน

ประกอบของ $\underline{s}(t)$ จะถูกปรับไปในทิศทางของเวกเตอร์เฉพาะ $\underline{u}_i$ ของ A ขนาดที่ถูกปรับในแต่ละทิศทางหาได้จากค่าเฉพาะ μ_i ของเวกเตอร์เฉพาะนั้น และความผิดพลาดกำลังสองเฉลี่ยที่ได้จากสมการ (2.5.20) และ (2.5.32) จะลดลงตามผลรวมถ่วงน้ำหนัก (weighted sum) ของส่วนประกอบในเวกเตอร์ $\underline{s}$

เนื่องจากเมตริกซ์ A มีองค์ประกอบเป็นจำนวนจริงและเป็นเมตริกซ์สมมาตร ค่าเฉพาะของ A จึงเป็นจำนวนจริง คำตอบของสมการ (2.5.36) หรือสมการ (2.5.35) (ที่ยังไม่แปลงเวกเตอร์) จะลู่เข้า ก็ต่อเมื่อ $|\mu_i| < 1;\ 1 \le i \le N$ และเนื่องจากค่าเฉพาะ μ_i ขึ้นกับความยาวก้าว β ดังนั้น ค่า β ที่ทำให้ความแปรปรวนของพารามิเตอร์ลู่เข้า ก็คือค่าที่ทำให้ทุกค่าเฉพาะของ A น้อยกว่า 1 นั่นเอง ซึ่งความผิดพลาดกำลังสองเฉลี่ยในสมการ (2.5.20) จะลู่เข้าด้วย ลองพิจารณาผลบวกขององค์ประกอบในแถวที่ l ของ A ซึ่งจะได้เป็น (แสดงไว้แตกต่างจากต้นฉบับ)

$$\sum_{m=1}^{N}\left[\left(1-2\beta\lambda_l\right)+\beta^2\lambda_l\lambda_m\right] = 1-\beta\lambda_l\left(2-\beta\sum_{m=1}^{N}\lambda_m\right) \tag{2.5.37}$$

ผลบวกนี้จะมีค่าน้อยกว่า 1 เมื่อ

$$0 < \beta < \frac{2}{\sum_{m=1}^{N}\lambda_m} = \frac{2}{N\phi_0} \tag{2.5.38}$$

โดยที่ $\phi_0 = \lambda_{av}$ เป็นค่าเฉลี่ยของค่าเฉพาะในเมตริกซ์ A เงื่อนไขนี้จะทำให้องค์ประกอบของเมตริกซ์ A ตามสมการ (2.5.34) เป็นบวกทั้งหมด เมตริกซ์ที่มีองค์ประกอบเป็นบวกทั้งหมดและผลรวมของแต่ละแถวน้อยกว่า 1 จะมีค่าเฉพาะที่มีขนาดน้อยกว่า 1 ดังนั้น เงื่อนไขในสมการ (2.5.38) เป็นขอบเขตของค่า β ที่ทำให้ความแปรปรวนของพารามิเตอร์ลู่เข้า และเงื่อนไขนี้จะจำกัดและถูกต้องกว่าในสมการ (2.3.11) ซึ่งเป็นขอบเขตสำหรับการลู่เข้าของค่าเฉลี่ยของเวกเตอร์พารามิเตอร์ เหตุผลที่ทำให้สมการ (2.5.38) จำกัดกว่าสมการ (2.3.11) เพราะว่า ในสมการ (2.3.11) ไม่ได้คิดถึงค่าความแปรปรวนของพารามิเตอร์ด้วย จากสมการ (2.5.38) และ (2.3.11) ยังทำให้รู้ว่า มีค่า β ช่วงหนึ่งที่ทำให้ค่าเฉลี่ยของพารามิเตอร์ลู่เข้าแต่ค่าความแปรปรวนไม่ลู่เข้า

2.5.3 ค่าลู่เข้าของความผิดพลาดกำลังสองเฉลี่ยและความแปรปรวนของพารามิเตอร์

เมื่อแสดงลักษณะของความผิดพลาดกำลังสองเฉลี่ยในรูปฟังก์ชันของเวลาแล้ว การคำนวณค่าลู่เข้าของความแปรปรวนพารามิเตอร์ และความผิดพลาดกำลังสองเฉลี่ย ทำได้โดยแทนเมตริกซ์ A ในสมการ (2.5.34) ลงในสมการ (2.5.32) และใช้ค่า ε_{ex} จากสมการ (2.5.20) จะได้

$$\underline{\tilde{s}}(t) \approx \left(I - 2\beta\Lambda\right)\underline{s}(t-1) + \beta^2\left[\varepsilon_{ex}(t) + \varepsilon_{\min}\right]\underline{\lambda} \tag{2.5.39}$$

ถ้าค่า β ที่เลือกเป็นไปตามเงื่อนไขในสมการ (2.5.38) แล้วค่า $\lim_{t\to\infty}\underline{s}(t) = \lim_{t\to\infty}\underline{s}(t-1)$ ถ้าหาลิมิต (limit) ทั้งสองข้างของสมการ (2.5.39) โดยให้ $t \to \infty$ จะได้เป็น

$$\lim_{t\to\infty} 2\beta\Lambda\underline{s}(t) \approx \beta^2\left[\lim_{t\to\infty}\varepsilon_{ex}(t) + \varepsilon_{\min}\right]\underline{\lambda} \tag{2.5.40}$$

หรือ

$$\lim_{t\to\infty}\underline{s}(t) = \frac{\beta}{2}\underline{1}\left[\lim_{t\to\infty}\varepsilon_{ex}(t) + \varepsilon_{\min}\right] \tag{2.5.41}$$

เมื่อ $\underline{1}$ เป็นเวกเตอร์ที่มีส่วนประกอบทุกตัวเป็น 1 เมื่อคูณสมการ (2.5.41) ทางด้านหน้าด้วย $\underline{\lambda}^T$ จะได้

$$\lim_{t\to\infty}\varepsilon_{ex}(t) = \frac{\beta}{2}\left(\sum_{j=1}^{N}\lambda_j\right)\left[\lim_{t\to\infty}\varepsilon_{ex}(t) + \varepsilon_{\min}\right] \tag{2.5.42}$$

แทน $N\phi_0$ ด้วย $\sum_{j=1}^{N}\lambda_j$ และแก้สมการหาค่า $\varepsilon_{ex}(\infty)$ จะได้สมการแสดงค่าลู่เข้าของความผิดพลาดกำลังสองเฉลี่ยที่เกินจากค่าต่ำสุด ซึ่งเกิดจากการปรับค่าพารามิเตอร์เป็น

$$\lim_{t\to\infty}\varepsilon_{ex}(t) = \frac{\beta\varepsilon_{\min}N\phi_0}{2 - \beta N\phi_0} \tag{2.5.43}$$

และแทนสมการ (2.5.43) ในสมการ (2.5.41) จะได้สมการแสดงค่าลู่เข้าของความแปรปรวนของพารามิเตอร์ เป็น

$$\lim_{t \to \infty} E\left[\tilde{q}_i^2(t)\right] = \frac{\beta \varepsilon_{\min}}{2 - \beta N \phi_0} \tag{2.5.44}$$

สมการ (2.5.15) และ (2.5.44) แสดงให้เห็นว่า เมื่อค่า β มีค่ามากขึ้น เวลาที่ใช้ในการลู่เข้าของระบบจะลดลง แต่ค่าของความผิดพลาดกำลังสองเฉลี่ยและความแปรปรวนของพารามิเตอร์จะเพิ่มขึ้น ในกรณีของสัญญาณขาเข้ามีค่าสถิติไม่เปลี่ยนตามเวลา (stationary) จะต้องกำหนดให้ค่า β มีค่ามากในตอนเริ่มต้นเพื่อให้ลู่เข้าได้เร็วและลดลงในเวลาต่อมาเพื่อให้ความผิดพลาดกำลังสองเฉลี่ยต่ำสุด ส่วนกรณีที่สัญญาณขาเข้ามีการเปลี่ยนแปลงตามเวลา (nonstationary) ต้องให้ความยาวก้าวในตอนเริ่มต้นมีค่ามากทำให้สามารถลดความผิดพลาดกำลังสองเฉลี่ยได้เร็ว และลดค่าความยาวลงตามลำดับเพื่อให้ขั้นตอนวิธีติดตามการเปลี่ยนแปลงที่น้อยลงและทำให้การกระเพื่อมน้อยลงด้วย

2.6 สรุป

ในบทนี้ ได้กล่าวถึงรายละเอียดของการลดระดับตามแนวเกรเดียนต์ เพื่อหาขอบเขตของความยาวก้าว ในกรณีที่สัญญาณขาเข้ามีลักษณะเป็นกระบวนการสโตคาสติกส์โดยใช้ความผิดพลาดกำลังสองเฉลี่ยเป็นฟังก์ชันเป้าหมาย ในกรณีที่ไม่ทราบค่าทางสถิติของสัญญาณขาเข้า การหาจุดเหมาะสมที่สุดจะใช้วิธีเกรเดียนต์แบบสโตคาสติกส์ เป็นผลให้พารามิเตอร์ที่ถูกปรับในระบบมีลักษณะเป็นกระบวนการสโตคาสติกส์ ค่าพารามิเตอร์จะไม่สามารถก้าวเข้าสู่จุดเหมาะสมที่สุดเมื่อใช้ความยาวก้าวเป็นค่าคงตัว แต่จะเกิดการกระเพื่อมรอบจุดเหมาะสมที่สุด การกระเพื่อมนี้เกิดขึ้นมากน้อยตามค่าความยาวก้าวที่ใช้ และผลของการใช้วิธีเกรเดียนต์แบบสโตคาสติกส์ยังทำให้ขอบเขตของความยาวก้าวแคบลง ขอบเขตซึ่งปกติถูกกำหนดด้วยค่าเฉพาะสูงสุดของเมตริกซ์อัตสหสัมพันธ์ของสัญญาณขาเข้า จะถูกจำกัดด้วยผลรวมของค่าเฉพาะทุกค่า ซึ่งความยาวก้าวที่อยู่ระหว่างช่วงทั้งสองนี้ ทำให้ค่าเฉลี่ยพารามิเตอร์ลู่เข้าแต่ความแปรปรวนไม่ลู่เข้า

สำหรับบทต่อไป จะกล่าวถึงขั้นตอนการปรับอัตราการเรียนรู้ที่ได้นำมาทดสอบในวิทยานิพนธ์ฉบับนี้ ค่าอัตราการเรียนรู้ของข่ายงานระบบประสาทจะถูกใช้เป็นความยาวก้าวในการปรับพารามิเตอร์ของระบบ ดังนั้น เนื้อหาที่กล่าวถึงในบทนี้ จึงสามารถนำไปใช้พิจารณาหรือวิเคราะห์อัตราการเรียนรู้สำหรับการประยุกต์ใช้งานต่าง ๆ ได้