Technical Report A62, Helsinki University of Technology, Publications in Computer and Information Science, Espoo, Finland, November 2000.

Convergence of a stochastic semisupervised clustering algorithm

Samuel Kaski
Neural Networks Research Centre
Helsinki University of Technology
P. O. Box 5400, 02015 HUT, Finland

1 Introduction

Consider a clustering task (Sinkkonen and Kaski, 2000) in which the average distortion to be minimized is

$$C = \sum_j \int y_j(\mathbf{x}) D(p(c|\mathbf{x}), \boldsymbol{\psi}_j) p(\mathbf{x}) d\mathbf{x} \ . \tag{1}$$

Here the probability density function (pdf) of the data is $p(\mathbf{x}, c)$, and the data consists of pairs $(\mathbf{x}, c)$. The $y_j(\mathbf{x})$ are cluster membership functions indicating the extent to which the data sample $\mathbf{x}$ belongs to cluster j. The functions fulfill the conditions $0 \leq y_j(\mathbf{x}) \leq 1$ and $\sum_j y_j(\mathbf{x}) = 1$. The distortion measure D is the Kullback-Leibler divergence between the conditional distribution $p(c|\mathbf{x})$ and the model or prototype distribution $\boldsymbol{\psi}_j$. The cluster membership functions are parameterized, and the cost function is to be optimized with respect to the model distributions and the parameters of the membership functions.

The goal of this paper is to prove that a certain stochastic approximation algorithm that minimizes the cost function converges or, more specifically, that the cost function converges (with probability one). The proof utilizes ideas presented in (Bottou, 1998). For a comprehensive treatment on stochastic approximation see (Kushner and Yin, 1997).

The cost function can be written as

$$C(\boldsymbol{\theta}) = E_{\mathbf{z}} Q(\mathbf{z}, \boldsymbol{\theta}) = \int Q(\mathbf{z}, \boldsymbol{\theta}) dP(\mathbf{z}) + C' \ , \tag{2}$$

where

$$Q(\mathbf{z}, \boldsymbol{\theta}) = -\sum_j y_j(\mathbf{x}; \boldsymbol{\theta}_j) \log \psi_{ji} \ , \tag{3}$$

$$\mathbf{z} = (\mathbf{x}, c, r, s) \ , \tag{4}$$

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N) \ , \tag{5}$$

and

$$\psi_{ji} = \frac{\exp(\gamma_{ji})}{\sum_k \exp(\gamma_{jk})} \ . \tag{6}$$

Note that the cost function is non-negative since it originates from the average (non-negative) Kullback-Leibler divergence (1); the constant C' makes the cost function (2) non-negative. The parameters are real-valued, $\boldsymbol{\theta}_j \in \mathbb{R}^n$ and $\boldsymbol{\gamma}_j \in \mathbb{R}^n$. The pdf of the data points $\mathbf{x} \in \mathbb{R}^n$ is denoted by $p(\mathbf{x})$. The c are samples of a multinomial random variable having values c_i, $i = 1, \ldots, N_c$. The function y_j is of the normalized exponential type,

$$y_j(\mathbf{x}; \boldsymbol{\theta}_j) = \frac{\exp f(\mathbf{x}; \boldsymbol{\theta}_j)}{\sum_l \exp f(\mathbf{x}; \boldsymbol{\theta}_l)} \ ,$$

and in this proof the basis functions are of the von Mises-Fisher type, i.e.,

$$f(\mathbf{x};\boldsymbol{\theta}_j) = \kappa \frac{\mathbf{x}^T\boldsymbol{\theta}_l}{\|\boldsymbol{\theta}_l\|} \; . \tag{7}$$

To simplify the notation the value of κ will be set to one without loss of generality.

The distribution of the data is assumed to satisfy

$$\int y_l(\mathbf{x};\boldsymbol{\theta}_l)p(c_i,\mathbf{x})d\mathbf{x} > \epsilon > 0$$

for all l and i, irrespective of the value of $\boldsymbol{\theta}_l$. Since y_l is always positive this condition is not very restrictive.

The values r and s are samples of independent random variables with pdfs uniformly distributed over the interval $(0,1)$. They have been introduced for technical (algorithmic) reasons. The variables r and s do not affect the results: when both are integrated out in (2) the cost function becomes equal to (1).

The stochastic approximation-type algorithm used for minimizing (2) is an iterative procedure in which the step

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \alpha(t)\mathbf{h}(\mathbf{z},\boldsymbol{\theta}) + \mathbf{g}(\mathbf{z},\alpha(t),\boldsymbol{\theta}) \; , \tag{8}$$

is repeated for $t = 1, 2, \ldots$. Here $\alpha(t)$ is a scalar coefficient that fulfills the usual stochastic approximation criteria

$$\sum_t \alpha^2(t) < \infty \text{ and} \sum_t \alpha(t) = \infty \; . \tag{9}$$

The function $\mathbf{h} = (\mathbf{h}^1, \ldots, \mathbf{h}^{2n})^T$ is defined by

$$\begin{aligned} \mathbf{h}^l(\mathbf{z};\boldsymbol{\theta}) &= \sum_j I_j(\mathbf{x},r;\boldsymbol{\theta})I_l(\mathbf{x},s;\boldsymbol{\theta})\frac{\partial f(\mathbf{x};\boldsymbol{\theta}_l)}{\partial\boldsymbol{\theta}_l}\log\frac{\psi_{ji}}{\psi_{li}}, \quad \text{and} \\ \mathbf{h}^{n+l}(\mathbf{z};\boldsymbol{\theta}) &= I_l(\mathbf{x},s;\boldsymbol{\theta})(\boldsymbol{\psi}_l - \mathbf{e}_i) \end{aligned}$$

for all $0 < l \leq N$. Here i is the index of the current sample $c = c_i$ and $\mathbf{e}_i$ is the unit vector in which the ith component is one. The $I_l(\mathbf{x},s;\boldsymbol{\theta})$ is an indicator function that chooses which parameters will be updated. The interval $(0,1)$ is divided into N subintervals having lengths $y_1(\mathbf{x};\boldsymbol{\theta}), \ldots, y_N(\mathbf{x};\boldsymbol{\theta})$, and the value of $I_l(\mathbf{z},s;\boldsymbol{\theta})$ is one if s belongs to the lth interval and zero otherwise.

The function $\mathbf{g}$ normalizes the parameters $\boldsymbol{\theta}_l$. Note that the purpose of this operation is merely to remove excessive degrees of freedom; the value of the cost function is not affected. The function $\mathbf{g}$ is defined by $\mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta}) = (\mathbf{g}^1(\mathbf{z},\alpha;\boldsymbol{\theta}_1), \ldots, \mathbf{g}^N(\mathbf{z},\alpha;\boldsymbol{\theta}_n), 0, \ldots, 0)$, where

$$\mathbf{g}^l(\mathbf{z},\alpha;\boldsymbol{\theta}_l) = \frac{\boldsymbol{\theta}_l - \alpha\mathbf{h}(\mathbf{z};\boldsymbol{\theta})}{\|\boldsymbol{\theta}_l - \alpha\mathbf{h}(\mathbf{z};\boldsymbol{\theta})\|} - (\boldsymbol{\theta}_l - \alpha\mathbf{h}(\mathbf{z};\boldsymbol{\theta}))$$

normalizes the parameter vector $\boldsymbol{\theta}_l$.

The proof consists of four parts. It is shown that (i) the norm of the parameter vector $\boldsymbol{\theta}$ remains bounded; this result is needed in the remaining steps, (ii) the algorithm proceeds, on the average, into the direction that decreases the cost function, (iii) the value of the cost function converges when the algorithm is applied iteratively, and (iv) the gradient of the cost function converges to zero.

2 The norm of $\boldsymbol{\theta}$ remains bounded

It is evident from the algorithm that the norm of $\boldsymbol{\theta}_j$ is constant, equal to one. It remains to be shown that each parameter γ_{lm} remains bounded.

Let us start by considering the change of the square of γ_{lm}. Remember that

$$\gamma_{lm}(t+1) = \gamma_{lm}(t) - \alpha(t) I_l(\mathbf{x}, s; \boldsymbol{\theta})(\psi_{lm}(t) - \delta_{mi}(t)) \ .$$

The change is (the time step t is omitted from the right side for brevity of notation)

$$(\gamma_{lm}(t+1))^2 - (\gamma_{lm}(t))^2 = -I_l(\mathbf{x}, s; \boldsymbol{\theta})[2\alpha\gamma_{lm}(\psi_{lm} - \delta_{mi}) + \alpha^2(\psi_{lm} - \delta_{mi})^2] \ .$$

Denote all the information that was available just before iteration t by $\mathcal{P}(t) = (\mathbf{z}(0), \ldots, \mathbf{z}(t-1), \boldsymbol{\theta}(0), \ldots, \boldsymbol{\theta}(t), \alpha(0), \ldots, \alpha(t))$. The expected change given this information is

$$\begin{aligned}
&E\{(\gamma_{lm}(t+1))^2 - (\gamma_{lm}(t))^2 | \mathcal{P}(t)\} \\
&\quad = -2\alpha\gamma_{lm}(E_{\mathbf{x},s,c}\{I_l(\mathbf{x}, s; \boldsymbol{\theta})\}\psi_{lm} - E_{\mathbf{x},s,c}\{\delta_{mi} I_l(\mathbf{x}, s; \boldsymbol{\theta})\}) \\
&\qquad\quad + \alpha^2 E\{I_l(\mathbf{x}, s; \boldsymbol{\theta})(\psi_{lm} - \delta_{mi})^2 | \mathcal{P}(t)\} \\
&\quad = -2\alpha\gamma_{lm}\left(\psi_{lm}\int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(\mathbf{x})d\mathbf{x} - \int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(c_m, \mathbf{x})d\mathbf{x}\right) \\
&\qquad\quad + \alpha^2 E\{I_l(\mathbf{x}, s; \boldsymbol{\theta})(\psi_{lm} - \delta_{mi})^2 | \mathcal{P}(t)\} \\
&\quad = -2\alpha\gamma_{lm}\int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(\mathbf{x})d\mathbf{x}\left(\psi_{lm} - \frac{\int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(c_m, \mathbf{x})d\mathbf{x}}{\int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(\mathbf{x})d\mathbf{x}}\right) \\
&\qquad\quad + \alpha^2 E\{I_l(\mathbf{x}, s; \boldsymbol{\theta})(\psi_{lm} - \delta_{mi})^2 | \mathcal{P}(t)\} \ . \qquad (10)
\end{aligned}$$

Since it was assumed that the numerator of the fraction in (10) is strictly positive the denominator is as well. Hence the whole fraction is strictly positive,

$$\inf_{\boldsymbol{\theta}} \frac{\int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(c_m, \mathbf{x})d\mathbf{x}}{\int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(\mathbf{x})d\mathbf{x}} \equiv \underline{K} > 0 \ .$$

Let us then use (10) to bound γ_{lm}.

Assume first that the sequence $\gamma_{lm}(t)$ diverges. Then there must be a subsequence that tends either to $-\infty$ or ∞. Assume first the former. There must then exist a divergent subsequence of $\gamma_{lm}(t)$ that is always negative and for which $\psi_{lm}(t) < \underline{K}$. Denote the subsequence by $\gamma(t)$. Then $\psi_{lm} < \int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(c_m, \mathbf{x})d\mathbf{x} / \int y_l(\mathbf{x}; \boldsymbol{\theta}_l)p(\mathbf{x})d\mathbf{x}$, and the first term of (10) is negative. The expectation in the second term is bounded, denote the bound by K. The expected changes in $(\gamma(t))^2$ can then be bounded by

$$E\{(\gamma(t+1))^2 - (\gamma(t))^2 | \mathcal{P}(t)\} \leq \alpha^2(t)K \ .$$

Denote by $\delta(t)$ the sequence for which $\delta(t) = 1$ if $E\{(\gamma(t+1))^2 - (\gamma(t))^2 | \mathcal{P}(t)\} > 0$, and $\delta(t) = 0$ otherwise. Then

$$E\{\delta(t)((\gamma(t+1))^2 - (\gamma(t))^2)\} = E\{\delta(t)E\{(\gamma(t+1))^2 - (\gamma(t))^2 | \mathcal{P}(t)\}\} \leq \alpha^2(t)K \ . \qquad (11)$$

Since $\sum_{t=1}^{\infty} \alpha^2(t)K < \infty$ the sum over t of the expected positive variations (11) converges. Moreover, since the sequence $(\gamma(t))^2$ is non-negative, a quasi-martingale convergence theorem (see Bottou (1998) for details) states that the sequence $(\gamma(t))^2$ converges with probability one. This is a contradiction, and hence there cannot exist (with probability one) any subsequence that would tend towards $-\infty$.

Assume then that there is a subsequence of $\gamma_{lm}(t)$ that tends to ∞. Since $\sum_k \gamma_{lk}(t) = 0$ there must then exist a subsequence of $\gamma_{lq}(t)$ that tends to $-\infty$, for some $q \neq m$. This was proven above to be impossible.

In conclusion, each γ_{lm} is bounded with probability one, and hence the norm of the whole parameter vector $\boldsymbol{\theta}$ is.

3 The mean step of the algorithm essentially reduces the cost function

Let us show that the mean of the step $\mathbf{h}$ is equal to the gradient of the cost function. It is easy to see that $\int I_j(\mathbf{x}, r; \boldsymbol{\theta}) dP(r) = y_j(\mathbf{x}; \boldsymbol{\theta})$, and hence the mean step can be expressed as

$$\begin{aligned} \int \mathbf{h}^l(\mathbf{z}; \boldsymbol{\theta}) dP(\mathbf{z}) &= \int \sum_j y_j(\mathbf{x}; \boldsymbol{\theta}) y_l(\mathbf{x}; \boldsymbol{\theta}) \frac{\partial f(\mathbf{x}; \boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta}_l} \log \frac{\psi_{ji}}{\psi_{li}} dP(\mathbf{x}, c), \quad \text{and} \\ \int \mathbf{h}^{n+l}(\mathbf{z}; \boldsymbol{\theta}) dP(\mathbf{z}) &= \int y_l(\mathbf{x}, \boldsymbol{\theta})(\boldsymbol{\psi}_l - \mathbf{e}_i) dP(\mathbf{x}, c) \;, \end{aligned} \tag{12}$$

for $l \leq N$. It can be shown by taking the derivative that $\partial Q(\mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is equal to the integrand in (12).

Note that the variables r and s do not appear on the right hand side of (12) any more. Since both of the variables integrate to unity the right hand side is equal to $\int \partial Q(\mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta} dP(\mathbf{z})$. Hence, it remains to be shown that

$$\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \int \frac{\partial Q(\mathbf{z}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dP(\mathbf{z}) \;,$$

i.e., that the order of the gradient and the integral can be exchanged.

It follows from a corollary of the Lebesque dominated convergence theorem (see, for example, Rudin, 1976) that if the domain is finite and all functions exist and are uniformly bounded, the order of taking the gradient and integration can be exchanged. The domain is finite as proven in the previous section. The derivatives exist and are uniformly bounded: it follows from the results of the previous section that there exist limits $0 < \psi_0 \leq \psi_{ji} \leq \psi_1 < 1$, and hence $\log \psi_{ji}/\psi_{li}$ is uniformly bounded. Since the rest of the functions in (12) are clearly uniformly bounded, the whole gradient $\partial Q(\mathbf{z}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is as well. Hence the dominated convergence theorem applies.

The word "essentially" in the title of the section means that the other step $\mathbf{g}$ in the algorithm does not change the value of the cost function: its sole purpose is to normalize the vectors $\boldsymbol{\theta}_j$, and as can be seen from (7) the norm does not affect the basis functions f.

4 The value of the cost function converges

Let us denote for brevity the sequence of values of the cost function by

$$u(t) = C(\boldsymbol{\theta}(t)) \;.$$

The task is then to show that this positive sequence converges. It will be proven by first showing that the sum of expected positive variations of $u(t)$ is finite, and then applying the quasi-martingale convergence theorem. The finiteness of the sum is proven by deriving an upper bound for the terms of the sum using Taylor expansions. Because of (9), the high-order terms of the expansions taken with respect to α at $\alpha = 0$ cannot take the sum to infinity, and can hence be neglected. The same holds for expansions taken with respect to the changes of $\boldsymbol{\theta}$, since by (8) the changes become small when α becomes small.

Applying the definition of the algorithm (8) and the Taylor expansion with respect to $\Delta\boldsymbol{\theta}(t) \equiv \boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t)$ at $\boldsymbol{\theta}(t)$, and retaining terms up to the second order yields

$$\begin{aligned}
u(t+1) &= u(t) + \Delta\boldsymbol{\theta}(t)^T \frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}} \\
&\quad + \frac{1}{2}\Delta\boldsymbol{\theta}(t)^T \frac{\partial^2 C(\boldsymbol{\theta}(t))}{\partial^2 \boldsymbol{\theta}} \Delta\boldsymbol{\theta}(t) \\
&\quad + \mathcal{O}(\|\Delta\boldsymbol{\theta}(t)\|^3) \\
&= u(t) + [-\alpha(t)\mathbf{h}(\mathbf{z};\boldsymbol{\theta}(t)) + \mathbf{g}(\mathbf{z},\alpha(t);\boldsymbol{\theta}(t))]^T \frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}} \\
&\quad + \frac{1}{2}[-\alpha(t)\mathbf{h}(\mathbf{z};\boldsymbol{\theta}(t)) + \mathbf{g}(\mathbf{z},\alpha(t);\boldsymbol{\theta}(t))]^T \frac{\partial^2 C(\boldsymbol{\theta}(t))}{\partial^2 \boldsymbol{\theta}} \\
&\quad \times [-\alpha(t)\mathbf{h}(\mathbf{z};\boldsymbol{\theta}(t)) + \mathbf{g}(\mathbf{z},\alpha(t);\boldsymbol{\theta}(t))] + \mathcal{O}(\|\Delta\boldsymbol{\theta}(t)\|^3)
\end{aligned}$$

Note that since all parameters are bounded all their continuous functions, including the second derivatives, are bounded as well[1]. Let us denote $\sup_{\|\mathbf{x}\|=1,\|\mathbf{y}\|=1,\boldsymbol{\theta}} |\mathbf{x}^T(\partial^2 C(\boldsymbol{\theta})/\partial^2\boldsymbol{\theta})\mathbf{y}| = K_1$. Note additionally that $\mathcal{O}(\|\Delta\boldsymbol{\theta}(t)\|^3) = \mathcal{O}(\alpha^3)$ for small α. Hence (the time step t is omitted for brevity of notation),

$$\begin{aligned}
u(t+1) &\leq u(t) + [-\alpha\mathbf{h}(\mathbf{z};\boldsymbol{\theta}) + \mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})]^T \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&\quad + \frac{1}{2}K_1\| - \alpha\mathbf{h}(\mathbf{z};\boldsymbol{\theta}) + \mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})\|^2 + \mathcal{O}(\alpha^3) \\
&= u(t) - \alpha\mathbf{h}(\mathbf{z};\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&\quad + \frac{1}{2}K_1\|\mathbf{h}(\mathbf{z};\boldsymbol{\theta})\|^2\alpha^2 + \frac{1}{2}K_1\|\mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})\|^2 - K_1\mathbf{h}(\mathbf{z};\boldsymbol{\theta})^T\mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})\alpha \\
&\quad + \mathcal{O}(\alpha^3)\ .
\end{aligned}$$

Let us then calculate the average change of u, conditioned on all that is available up to the present state. Let us denote $\sup_{\boldsymbol{\theta}} E_{\mathbf{z}}\{\|\mathbf{h}(\mathbf{z};\boldsymbol{\theta})\|^2\} = K_2 < \infty$, and assume that all expectations are taken with respect to $\mathbf{z}$.

$$\begin{aligned}
E\{u(t+1) - u(t)|\mathcal{P}(t)\} &\leq -\alpha\left\|\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|^2 + E\{\mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\} \\
&\quad + \frac{1}{2}K_1K_2\alpha^2 + \frac{1}{2}K_1E\{\|\mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})\|^2\} \\
&\quad - \alpha K_1E\{\mathbf{h}(\mathbf{z};\boldsymbol{\theta})^T\mathbf{g}(\mathbf{z},\alpha;\boldsymbol{\theta})\} + \mathcal{O}(\alpha^3) \\
&= -\alpha\left\|\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|^2 + \sum_j E\{\mathbf{g}^j(\mathbf{z},\alpha;\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\} \\
&\quad + \frac{1}{2}K_1K_2\alpha^2 + \frac{1}{2}K_1\sum_j E\{\|\mathbf{g}^j(\mathbf{z},\alpha;\boldsymbol{\theta})\|^2\} \\
&\quad - \alpha K_1\sum_j E\{\mathbf{h}^j(\mathbf{z};\boldsymbol{\theta})^T\mathbf{g}^j(\mathbf{z},\alpha;\boldsymbol{\theta})\} + \mathcal{O}(\alpha^3)\ . \quad (13)
\end{aligned}$$

Suitable upper bounds are needed for the terms in (13), i.e., upper bounds having converging sums over t. The first term is clearly non-positive and can be bounded by zero. In bounding the other terms some additional results are needed.

[1] Technically the functions must be continuous within the closure of the domain into which the parameters are bounded, which holds for all functions in this paper.

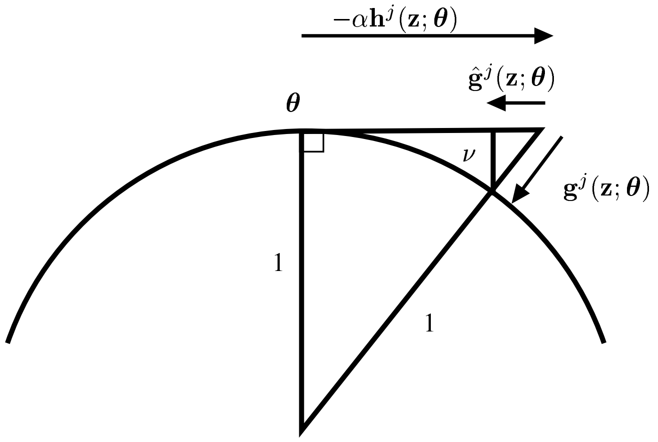

Figure 1: Demonstration of the geometry of the adaptation.

Denote the component of $\mathbf{g}^j(\mathbf{z};\boldsymbol{\theta})$ that is parallel to $\mathbf{h}^j(\mathbf{z};\boldsymbol{\theta})$ by $\hat{\mathbf{g}}^j(\mathbf{z};\boldsymbol{\theta})$ (cf. Fig. 1). The norms of both $\mathbf{g}^j$ and $\hat{\mathbf{g}}^j$ will be needed below (the arguments of the functions will be omitted for brevity). Based on the Pythagorean theorem,

$$1^2 + \alpha^2 ||\mathbf{h}^j||^2 = (1 + ||\mathbf{g}^j||)^2 \ ,$$

from which

$$||\mathbf{g}^j|| = \sqrt{1 + \alpha^2 ||\mathbf{h}^j||^2} - 1 \ . \tag{14}$$

The norm of $\hat{\mathbf{g}}^j$ can then be solved utilizing the two similar triangles in Figure 1, from which $\nu = ||\hat{\mathbf{g}}^j||/(\alpha||\mathbf{h}^j||)$, and again the Pythagorean theorem, this time applied to the smaller triangle. The result is

$$||\hat{\mathbf{g}}^j|| = \frac{\alpha ||\mathbf{h}^j||}{\sqrt{1 + \alpha^2 ||\mathbf{h}^j||^2}} ||\mathbf{g}^j|| = \alpha ||\mathbf{h}^j|| - \frac{\alpha ||\mathbf{h}^j||}{\sqrt{1 + \alpha^2 ||\mathbf{h}^j||^2}} \ . \tag{15}$$

Let us then return to the jth term in the first sum over j in (13). Noting that $\hat{\mathbf{g}}^j$ has the opposite direction as $-\alpha \mathbf{h}^j$, and utilizing (15) yields

$$\begin{aligned} E\left\{ \mathbf{g}^j(\mathbf{z},\alpha;\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\} &= E\left\{ \hat{\mathbf{g}}^j(\mathbf{z},\alpha;\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\} \\ = E\left\{ ||\hat{\mathbf{g}}^j|| \frac{(\mathbf{h}^j)^T}{||\mathbf{h}^j||} \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\} &= \alpha E\left\{ (1 - \frac{1}{\sqrt{1 + \alpha^2 ||\mathbf{h}^j||^2}}) (\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\} \\ &= \alpha E\left\{ (\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\} - \alpha E\left\{ \frac{(\mathbf{h}^j)^T}{\sqrt{1 + \alpha^2 ||\mathbf{h}^j||^2}} \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\} \ , \end{aligned}$$

and by Taylor expansion with respect to α, at $\alpha = 0$, applied to the second term,

$$
\begin{aligned}
E\left\{\mathbf{g}^j(\mathbf{z},\alpha;\boldsymbol{\theta})^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\right\} &= \alpha \left\| \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\|^2 - \alpha E\left\{(\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} + \right.\\
&\quad + \alpha \left[-(1+\alpha^2\|\mathbf{h}^j\|^2)^{-3/2}\alpha\|\mathbf{h}^j\|^2\right]_{\alpha=0} (\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\\
&\left. + \frac{1}{2}\alpha^2 \left[\frac{3}{2}(1+\alpha^2\|\mathbf{h}^j\|^2)^{-5/2}\alpha^2\|\mathbf{h}^j\|^4 - (1+\alpha^2\|\mathbf{h}^j\|^2)^{-3/2}\|\mathbf{h}^j\|^2\right]_{\alpha=0} (\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\right\}\\
&\quad + \mathcal{O}(\alpha^3)\\
&= \alpha \left\| \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} \right\|^2 - \alpha E\left\{(\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} - \frac{1}{2}\alpha^2 \|\mathbf{h}^j\|^2 (\mathbf{h}^j)^T \frac{\partial C(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\right\} + \mathcal{O}(\alpha^3)\\
&= \mathcal{O}(\alpha^3) \ . \qquad (16)
\end{aligned}
$$

Consider then the jth term in the second sum over j in (13). A second-order Taylor expansion with respect to α at $\alpha = 0$ yields

$$
E\{\|\mathbf{g}^j\|\} = E\{\sqrt{1+\alpha^2\|\mathbf{h}^j\|^2} - 1\} = E\{\frac{1}{2}\alpha^2\|\mathbf{h}^j\|^2\} + \mathcal{O}(\alpha^3) \le \frac{1}{2}\alpha^2 K_2 + \mathcal{O}(\alpha^3) \ , \qquad (17)
$$

where (14) was used. Hence, by squaring (14),

$$
\begin{aligned}
E\{\|\mathbf{g}^j\|^2\} &= E\{\alpha^2\|\mathbf{h}^j\|^2\} - 2E\{\sqrt{1+\alpha^2\|\mathbf{h}^j\|^2} - 1\}\\
&= E\{\alpha^2\|\mathbf{h}^j\|^2\} - 2E\{\frac{1}{2}\alpha^2\|\mathbf{h}^j\|^2\} + \mathcal{O}(\alpha^3) = \mathcal{O}(\alpha^3) \ . \qquad (18)
\end{aligned}
$$

In the jth term in the last sum of (13) the expectation can be bounded by

$$
\begin{aligned}
&E\{\mathbf{h}^j(\mathbf{z};\boldsymbol{\theta})^T \mathbf{g}^j(\mathbf{z},\alpha;\boldsymbol{\theta})\} = E\{(\mathbf{h}^j)^T \hat{\mathbf{g}}^j\} = E\{\|\mathbf{h}^j\|\|\hat{\mathbf{g}}^j\|\}\\
&= \alpha E\left\{\left(1 - \frac{1}{\sqrt{1+\alpha^2\|\mathbf{h}\|^2}}\right)\|\mathbf{h}^j\|^2\right\} = \alpha E\{\|\mathbf{h}^j\|^2\} - \alpha E\left\{\frac{\|\mathbf{h}^j\|^2}{\sqrt{1+\alpha^2\|\mathbf{h}^j\|^2}}\right\}\\
&= \frac{1}{2}\alpha^2\|\mathbf{h}^j\|^4 \ , \qquad (19)
\end{aligned}
$$

again by a Taylor expansion.

Let us then collect the results by substituting (16), (18), and (19) into (13), resulting in

$$
E\{u(t+1) - u(t)|\mathcal{P}(t)\} \le \frac{1}{2}K_1 K_2 \alpha^2(t) + \mathcal{O}(\alpha^3(t)) \ . \qquad (20)
$$

The sequence $\delta(t)$ can now be defined in the same way as in Section 2, and the theorem from martingale theory can be used to prove that the sequence $u(t)$ converges with probability one.

5 The norm of the gradient of the cost function converges to zero

Denote for brevity

$$
v(t) = \left\| \frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}} \right\|^2 \ .
$$

From the previous section it is known that the sum over t of the left-hand side of (13) is finite, and the sum of the non-negative terms on the right-hand side is finite as well. Hence,

$$\sum_{t=1}^{\infty} \alpha(t) v(t) < \infty \ . \tag{21}$$

This result will be used below in proving the convergence.

Develop then the variations of v into a Taylor expansion with respect to $\Delta\boldsymbol{\theta}(t) \equiv \boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t)$, at $\boldsymbol{\theta}(t)$:

$$\begin{aligned} v(t+1) &= v(t) + 2\left[\frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}}\right]^T \frac{\partial^2 C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}^2} \Delta\boldsymbol{\theta}(t) \\ &\quad + \frac{1}{2}(\Delta\boldsymbol{\theta}(t))^T \frac{\partial^2 \|\partial C(\boldsymbol{\theta}(t))/\partial \boldsymbol{\theta}\|^2}{\partial \boldsymbol{\theta}^2} \Delta\boldsymbol{\theta}(t) \ . \end{aligned}$$

The elements of $\partial^2 \|\partial C(\boldsymbol{\theta}(t))/\partial \boldsymbol{\theta}\|^2 / \partial \boldsymbol{\theta}^2$ are bounded since they are continuous functions of the bounded parameters. Therefore the last term is bounded by $\|\Delta\boldsymbol{\theta}(t)\|^2 K_4$ for some K_4. Plug then the stochastic approximation formula in, and the following expression is obtained:

$$\begin{aligned} v(t+1) &\leq v(t) + 2\left[\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \frac{\partial^2 C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}(-\alpha\mathbf{h} + \mathbf{g}) + K_4\|-\alpha\mathbf{h} + \mathbf{g}\|^2 \\ &\quad + \mathcal{O}(\alpha^3) \\ &= v(t) - 2\alpha\left[\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \frac{\partial^2 C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\mathbf{h} + 2\left[\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \frac{\partial^2 C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\mathbf{g} \\ &\quad + K_4\|\mathbf{h}\|^2\alpha^2 + K_4\|\mathbf{g}\|^2 - 2\mathbf{h}^T\mathbf{g}\alpha + \mathcal{O}(\alpha^3) \end{aligned}$$

For simplicity of notation, the time t and the arguments of $\mathbf{h}$ and $\mathbf{g}$ were again omitted. Let us then take the expectation, yielding

$$\begin{aligned} E\{v(t+1) - v(t)|\mathcal{P}(t)\} &\leq -2\alpha\left[\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \frac{\partial^2 C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &\quad + 2\left[\frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}}\right]^T \frac{\partial^2 C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}E\{\mathbf{g}\} + K_4 E\{\|\mathbf{h}\|^2\}\alpha^2 \\ &\quad + K_4 E\{\|\mathbf{g}\|^2\} - 2E\{\mathbf{h}^T\mathbf{g}\}\alpha + \mathcal{O}(\alpha^3) \end{aligned}$$

Denote the vertical "slice" of any matrix A containing the columns corresponding to $\boldsymbol{\theta}_j$ by A^j, and bound the second derivatives by K_1 as previously, yielding

$$\begin{aligned} E\{v(t+1) - v(t)|\mathcal{P}(t)\} &\leq 2\alpha K_1 \left\|\frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}}\right\|^2 \\ &\quad + \sum_j \left[\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \left[\frac{\partial^2 C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right]^j E\{\mathbf{g}^j\} \\ &\quad + K_4 E\{\|\mathbf{h}\|^2\}\alpha^2 + K_4 \sum_j E\{\|\mathbf{g}^j\|^2\} \\ &\quad - 2\alpha\sum_j E\{(\mathbf{h}^j)^T\mathbf{g}^j\} + \mathcal{O}(\alpha^3) \ . \end{aligned}$$

Denote

$$\sup_{\boldsymbol{\theta}} \sup_{\|\mathbf{x}\|=1} \left[\frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}}\right]^T \left[\frac{\partial^2 C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}^2}\right]^j \mathbf{x} = K_1^j < \infty \ .$$

Let us then use this bound, (17), (18), and (19) to obtain the final upper bound:

$$E\{v(t+1) - v(t)|\mathcal{P}(t)\} \quad \leq \quad 2\alpha K_1 \left\| \frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|^2 + \frac{1}{2}\alpha^2 K_2 \sum_j K_1^j + K_2 K_4 \alpha^2 + \mathcal{O}(\alpha^3) \; .$$

If $\delta(t)$ is defined as previously, the expected positive changes of v can be bounded by

$$E\{\delta(t)(v(t+1) - v(t))\} = E\{\delta(t)E\{(v(t+1) - v(t)|\mathcal{P}(t))\}\}$$
$$\leq \alpha(t)K_1 \left\| \frac{\partial C(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}} \right\|^2 + \alpha^2 \left(\frac{1}{2} K_2 \sum_j K_1^j + K_2 K_4 \right) + \mathcal{O}(\alpha^3) \; .$$

The sums of all three the terms are finite, and since $v(t)$ is positive the quasi-martingale theorem can be used again to guarantee that $v(t)$ converges with probability one. It then follows from (21) that the gradient must converge to zero.

Acknowledgements

I would like to thank Janne Sinkkonen and Jaakko Peltonen for useful comments on the manuscript.

References

Bottou, L. (1998). On-line learning and stochastic approximations. In Saad, D., editor, *On-line learning in neural networks*, pages 9–42. Cambridge University Press, Cambridge, UK.

Kushner, H. J. and Yin, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer, New York.

Rudin, W. (1976). *Principles of mathematical analysis*. Mc-Graw-Hill, Singapore. Third Ed.

Sinkkonen, J. and Kaski, S. Semisupervised clustering based on conditional distributions in an auxiliary space. Submitted to a journal.