

- [13] A. Jagota and M.K. Warmuth. Continuous and discrete time nonlinear gradient descent: relative loss bounds and convergenc. In *International Symposium on Artificial Intelligence and Mathematic*, 1998.
- [14] M.I. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8:1409–1431, 1995.
- [15] Jyrki Kivinen and Manfred K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, January 1997.
- [16] Jyrki Kivinen and Manfred K. Warmuth. Relative loss bounds for multidimensional regression problems. In *Advances in Neural Information Processing Systems 10*, 1997.
- [17] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V Vapnik. Comparison of learning algorithms for handwritten digit recognition. In *International Conference on Artificial Neural Networks*, pages 53–60. World Scientific, 1995.
- [18] X. L. Meng and D. B. Rubin. Recent extensions of the EM algorithm (with discussion). In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics, 4*. Oxfod: Clarendon Press, 1992.
- [19] B.C. Peters and H.F. Walker. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal of Applied Mathematics*, 35:362–378, 1978.
- [20] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2), 1984.
- [21] H. Robbins and S. Monro. A stochastic approximation model. *Annals of Mathematical Statistics*, 22, 1951.
- [22] Y. Singer and M.K. Warmuth. Training algorithms for hidden Markov models using entropy based distance functions. In *Advances in Neural Information Processing Systems 6*, 1996.
- [23] C.F.J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.
- [24] L. Xu and M.I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151, 1996.

Appendix

Throughout the appendix the expectations and variances of random variables are taken with respect to a Gaussian distribution with parameters \mathbf{C}, μ (or $\tilde{\mathbf{C}}, \tilde{\mu}$). We list several algebraic properties that are used in the paper to derive the updates.

1. $\text{tr}(\mathbf{AB}) = \sum_{p=1}^n \sum_{q=1}^m \mathbf{A}_{p,q} \mathbf{B}_{q,p} = \sum_{q=1}^m \sum_{p=1}^n \mathbf{B}_{q,p} \mathbf{A}_{p,q} = \text{tr}(\mathbf{BA})$.

2. Let A be a square matrix. Then, $|\mathbf{A}| = \sum_p \text{cof}_{p,q} \mathbf{A}_{p,q}$ and $\mathbf{A}^{-1} = \frac{\text{cof}_{q,p}}{|\mathbf{A}|}$, where $\text{cof}_{p,q}$ is the cofactor of $\mathbf{A}_{p,q}$ in \mathbf{A} . For a symmetric matrix A , $\text{cof}_{p,q} = \text{cof}_{q,p}$ and we get that

$$\frac{d |\mathbf{A}|}{d \mathbf{A}_{p,q}} = \text{cof}_{p,q} = \text{cof}_{q,p} = |\mathbf{A}| (\mathbf{A}^{-1})_{p,q}.$$

Thus $\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}| \mathbf{A}^{-1}$ and, similarly, $\frac{\partial |\mathbf{A}^{-1}|}{\partial \mathbf{A}} = -|\mathbf{A}^{-1}| \mathbf{A}^{-1}$.

3. From (2) we get $\frac{\partial \ln |\mathbf{A}|}{\partial \mathbf{A}} = \mathbf{A}^{-1}$.

4. From (2) and (3) we have $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ and thus $\frac{\partial \ln |\mathbf{A}|}{\partial \mathbf{A}^{-1}} = -\frac{\partial \ln |\mathbf{A}^{-1}|}{\partial \mathbf{A}^{-1}} = -\mathbf{A}$.

5.

$$\begin{aligned} \mathbf{C} &= E((\mathbf{x} - \mu)(\mathbf{x} - \mu)^T) = E(\mathbf{x}\mathbf{x}^T) - E(\mathbf{x}\mu^T) - E(\mu\mathbf{x}^T) + E(\mu\mu^T) \\ &= E(\mathbf{x}\mathbf{x}^T) - E(\mu\mu^T) \end{aligned}$$

6.

$$\begin{aligned} \tilde{E}((\mathbf{x} - \tilde{\mu})^T \tilde{\mathbf{C}}^{-1} (\mathbf{x} - \tilde{\mu})) &= \tilde{E}(\text{tr}((\mathbf{x} - \tilde{\mu})^T \tilde{\mathbf{C}}^{-1} (\mathbf{x} - \tilde{\mu}))) \\ &= \tilde{E}(\text{tr}(\tilde{\mathbf{C}}^{-1} (\mathbf{x} - \tilde{\mu})(\mathbf{x} - \tilde{\mu})^T)) \quad (\text{Using (1)}) \\ &= \text{tr}(\tilde{\mathbf{C}}^{-1} \tilde{E}(\mathbf{x} - \tilde{\mu})(\mathbf{x} - \tilde{\mu})^T) \\ &= \text{tr}(\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{C}}) = \text{tr}(\mathbf{I}) = d \end{aligned}$$

7.

$$\begin{aligned} \tilde{E}((\mathbf{x} - \mu)^T \mathbf{C}^{-1} (\mathbf{x} - \mu)) &= \tilde{E}(\text{tr}((\mathbf{x} - \mu)^T \mathbf{C}^{-1} (\mathbf{x} - \mu))) \\ &= \text{tr}(\mathbf{C}^{-1} \tilde{E}(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T) \\ &= \text{tr}(\mathbf{C}^{-1} \tilde{E}(\mathbf{x}\mathbf{x}^T - \mathbf{x}\mu^T - \mu\mathbf{x}^T + \mu\mu^T)) \\ &= \text{tr}(\mathbf{C}^{-1} (\tilde{\mathbf{C}} + \tilde{\mu}\tilde{\mu}^T - \tilde{\mu}\mu^T - \mu\tilde{\mu}^T + \mu\mu^T)) \quad (\text{Using (5)}) \\ &= \text{tr}(\mathbf{C}^{-1} \tilde{\mathbf{C}}) + \text{tr}(\mathbf{C}^{-1} (\tilde{\mu} - \mu)(\tilde{\mu} - \mu)^T) \\ &= \text{tr}(\mathbf{C}^{-1} \tilde{\mathbf{C}}) + (\tilde{\mu} - \mu)^T \mathbf{C}^{-1} (\tilde{\mu} - \mu) \end{aligned}$$

8. $\frac{\partial \text{tr}(\mathbf{A}\mathbf{B})}{\partial \mathbf{B}} = \frac{\partial \sum_p \sum_q A_{p,q} B_{q,p}}{\partial \mathbf{B}} = \mathbf{A}^T$

9. Let $\mathbf{I}_{p,q}$ denotes the matrix that has a one in position (p, q) and is zero otherwise and let $\mathbf{A}_{*,p}$ ($\mathbf{A}_{p,*}$) denote the p th column (row) of the matrix \mathbf{A} . Then,

$$\mathbf{A}\mathbf{I}_{p,q}\mathbf{B} = \text{col}(\mathbf{A}, p) \cdot \text{row}(\mathbf{B}, q) = \mathbf{A}_{*,p}\mathbf{B}_{q,*} \quad \text{and} \quad \mathbf{x}^T \mathbf{I}_{p,q} \mathbf{x} = [\mathbf{x}^T \mathbf{x}]_{p,q} = \mathbf{x}_p \mathbf{x}_q$$

10. Let $\mathbf{0}$ denote the zero matrix. Then,

$$\mathbf{0} = \frac{d \mathbf{I}}{d \mathbf{C}_{p,q}} = \frac{d \mathbf{C}^{-1} \mathbf{C}}{d \mathbf{C}_{p,q}} = \frac{d \mathbf{C}^{-1}}{d \mathbf{C}_{p,q}} \mathbf{C} + \mathbf{C}^{-1} \frac{d \mathbf{C}}{d \mathbf{C}_{p,q}}$$

This implies that $\frac{d \mathbf{C}^{-1}}{d \mathbf{C}_{p,q}} = -\mathbf{C}^{-1} \mathbf{I}_{p,q} \mathbf{C}^{-1} = -\mathbf{C}_{*,p}^{-1} \mathbf{C}_{q,*}^{-1}$.

11. $\frac{\partial \mathbf{C}^{-1}}{\partial \mathbf{C}} = -\mathbf{C}^{-1} \mathbf{C}^{-1}$.

12. $\frac{\partial \mathbf{z}^T \mathbf{A} \mathbf{z}}{\partial \mathbf{A}} = \frac{\partial \text{tr}(\mathbf{z}\mathbf{z}^T \mathbf{A})}{\partial \mathbf{A}} = (\mathbf{z}\mathbf{z}^T)^T = \mathbf{z}\mathbf{z}^T$

13. $\frac{d \mathbf{x}^T \mathbf{B} \mathbf{x}}{d \mathbf{y}} = \mathbf{x}^T \frac{d \mathbf{B}}{d \mathbf{y}} \mathbf{x}$

14. From above we get,

$$\begin{aligned} \frac{d (\mathbf{x} - \mu)^T \mathbf{C}^{-1} (\mathbf{x} - \mu)}{d \mathbf{C}_{p,q}} &= -(\mathbf{x} - \mu)^T \mathbf{C}^{-1} \mathbf{I}_{p,q} \mathbf{C}^{-1} (\mathbf{x} - \mu) \\ &= -(\mathbf{C}^{-1} (\mathbf{x} - \mu))^T \mathbf{I}_{p,q} \mathbf{C}^{-1} (\mathbf{x} - \mu) \\ &= -(\mathbf{C}^{-1} (\mathbf{x} - \mu))_p (\mathbf{C}^{-1} (\mathbf{x} - \mu))_q \end{aligned}$$

Hence

$$\begin{aligned}\frac{\partial(\mathbf{x} - \mu)^T \mathbf{C}^{-1}(\mathbf{x} - \mu)}{\partial \mathbf{C}} &= -(\mathbf{C}^{-1}(\mathbf{x} - \mu))(\mathbf{C}^{-1}(\mathbf{x} - \mu))^T \\ &= -\mathbf{C}^{-1}(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \mathbf{C}^{-1}\end{aligned}$$

15. Equality (14) imply that $\frac{\partial(\mathbf{x}-\mu)^T \mathbf{C}^{-1}(\mathbf{x}-\mu)}{\partial \mathbf{C}^{-1}} = (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T$.

16. Let \mathbf{A} and \mathbf{B} be symmetric matrices. Then,

$$\frac{\partial}{\partial \mathbf{B}_{i,j}^{-1}} \sum_{p,q} \mathbf{A}_{p,q} \mathbf{B}_{q,p} = \sum_{p,q} \mathbf{A}_{p,q} \frac{\partial \mathbf{B}_{p,q}}{\partial \mathbf{B}_{i,j}^{-1}} = -\sum_{p,q} \mathbf{A}_{p,q} \mathbf{B}_{i,p} \mathbf{B}_{q,j} = -\mathbf{B}_{i,*} \mathbf{A} \mathbf{B}_{*,j}$$

Hence, $\frac{\partial \text{tr}(\mathbf{A} \mathbf{B})}{\partial \mathbf{B}^{-1}} = -\mathbf{B} \mathbf{A} \mathbf{B}$.