When is there a free matrix lunch?

Manfred K. Warmuth
University of California - Santa Cruz

The “no-free lunch theorems” essentially say that for any two algorithms A and B, there are “as many” targets (or priors over targets) for which A has lower expected loss than B as vice-versa. This can be made precise for certain loss functions [WM97]. This note concerns itself with cases where seemingly harder matrix versions of the algorithms have the same on-line loss bounds as the corresponding vector versions. So it seems that you get a free “matrix lunch” (Our title is however not meant to imply that we have a technical refutation of the no-free lunch theorems).

The simplest case of this phenomenon occurs in the so-called *expert setting.* We have n experts. In each trial the algorithm proposes a probability vector $\mathbf{w}^t$ over the n experts, receives a loss vector $\boldsymbol{\ell}^t \in [0,1]^n$ for the experts and incurs an expected loss $\mathbf{w}^t \cdot \boldsymbol{\ell}^t$. The Weighted Majority or Hedge algorithm uses exponential weights $w_i^t \sim w_i^1 e^{-\eta \sum_{t=1}^{t-1} \ell_i^t}$ and has the following expected loss bound [FS97]:

$$\sum_{t=1}^{T} \mathbf{w}^t \cdot \boldsymbol{\ell}^t \;\le\; \ell^* + \sqrt{2\ell^* \ln n} + \ln n,$$

when η is tuned as a function of n and the best loss $\ell^* = \inf_i \sum_{t=1}^{T} \ell_i^t$.

Recently a matrix version of this algorithm has been developed in parallel by a number of researchers [WK06b,AK07]. Now the experts are outer products or *dyads* $\mathbf{u}\mathbf{u}^\top$, where $\mathbf{u}$ is a unit vector in $\mathbb{R}^n$, and there are continuously many such dyads (one for each pair $\pm\mathbf{u}$). The uncertainty of the algorithm about which dyad is good is expressed as a mixture of dyads (or density matrix) $\mathbf{W}^t$.

The loss vector $\boldsymbol{\ell}^t$ at trial t is replaced by a covariance matrix $\mathbf{L}^t$ whose eigenvalues must lie in the interval [0,1]. The symmetric matrix $\mathbf{L}^t$ specifies the loss for all dyads $\mathbf{u}\mathbf{u}^\top$ at trial t via the formula $(\mathbf{u}\mathbf{u}^\top) \bullet \mathbf{L}^t = \mathbf{u}^\top \mathbf{L}^t \mathbf{u}$ which is the variance at trial t in direction $\mathbf{u}$. The algorithm is charged by the expected variance $\mathbf{W}^t \bullet \mathbf{L}^t = \sum_{i,j} W_{i,j}^t L_{i,j}^t$ and its total expected variance is bounded as

$$\sum_{t=1}^{T} \mathbf{W}^t \bullet \mathbf{L}^t \;\le\; L^* + \sqrt{2L^* \ln n} + \ln n,$$

when η is tuned as a function of n and $L^* = \inf_{\mathbf{u}} \sum_{t=1}^{T} \mathbf{u}^\top \mathbf{L}^t \mathbf{u}$. The algorithm that achieves this is a matrix generalization of the weighted majority algorithm defined using the matrix exponential (See [WK06b,AK07] for details).

Curiously enough, if the initial density matrix is uniform and we let $\mathbf{L}^t = \mathrm{diag}(\boldsymbol{\ell}^t)$, then the matrix version of the algorithm and bound specialize to the original vector version. That is, the original Weighted Majority algorithm is retained as a special case when all instance/loss matrices $\mathbf{L}^t$ have the identity matrix as an eigensystem. The same phenomenon happens for linear regression w.r.t. square loss: The bounds proven for matrix version of the exponentiated gradient algorithm [TRW05] are identical to the original bounds [KW97] when the instances are diagonal matrices. The same happens for Boosting [TRW05] and the matrix version of the Winnow algorithm [War07]. There is even a Bayes rule for density matrices that has the standard Bayes rule and its bounds as a special case when the priors and data likelihoods are diagonal matrices [WK06a].

Note that this phenomenon is also really puzzling from an information theoretic point of view. It takes $\ln n$ nats to encode the identity of one of the n experts. However it should take more than $\ln n$ nats to encode an arbitrary direction/dyad in n dimensions. **What properties are required for an on-line algorithm and its bound for solving a problem with symmetric matrix instances so that the worst-case of the bound is attained when the instances are diagonal (i.e. have the identity matrix as an eigensystem), thus allowing us to get the matrix case for free?** So far, all cases where this has been shown is for algorithms that are motivated by some kind of relative entropy regularized optimization problem. Does this phenomenon occur for other families of updates as well that are motivated by different divergences?

Ideally, we would like the characterization be defined i.t.o. some spectral invariance of the regularization and loss function (see e.g. [WV05] for an example where a family of updates is characterized by a notion of rotation invariance).

Acknowledgement: Thanks to Dima Kuzmin for helpful discussions.

References

[AK07] S. Arora and S. Kale. A combinatorial primal-dual approach to semidefinite programs. In *Proc. 39th Annual ACM Symposium on Theory of Computing*. ACM, 2007. To appear.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[KW97] J. Kivinen and M. K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, January 1997.

[TRW05] K. Tsuda, G. Rätsch, and M. K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projections. *Journal of Machine Learning Research*, 6:995–1018, June 2005.

[War07] Manfred K. Warmuth. Winnowing subspaces. Unpublished manuscript, February 2007.

[WK06a] Manfred Warmuth and Dima Kuzmin. A Bayesian probability calculus for density matrices. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence UAI06)*, 2006.

[WK06b] Manfred K. Warmuth and Dima Kuzmin. Online variance minimization. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT 06)*, Pittsburg, June 2006. Springer.

[WM97] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transaction for Evolutionary Computation*, 1(67), 1997.

[WV05] M.K. Warmuth and S.V.N. Vishwanathan. Leaving the span. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT 05)*, Bertinoro, Italy, June 2005. Springer. Journal version: http://www.cse.ucsc.edu/ manfred/pubs/span.pdf.