

Multiplicative Updates for Nonnegative Quadratic Programming

Fei Sha* Yuanqing Lin[†] Lawrence. K. Saul[‡] Daniel. D. Lee[§]

August 10, 2006

Abstract

Many problems in neural computation and statistical learning involve optimizations with nonnegativity constraints. In this paper, we study convex problems in quadratic programming where the optimization is confined to an axis-aligned region in the nonnegative orthant. For these problems, we derive multiplicative updates that improve the value of the objective function at each iteration and converge monotonically to the global minimum. The updates have a simple closed form and do not involve any heuristics or free parameters that must be tuned to ensure convergence. Despite their simplicity, they differ strikingly in form from other multiplicative updates used in machine learning. We provide complete proofs of convergence for these updates and describe their application to problems in signal processing and pattern recognition.

*387 Soda Hall, Computer Science Division, University of California, Berkeley, CA 94720-1776, USA (feisha@cs.berkeley.edu)

[†]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA (linyuanq@seas.upenn.edu)

[‡]Department of Computer Science and Engineering, University of California (San Diego), La Jolla, CA 92093-0404, USA (saul@cs.ucsd.edu)

[§]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA (ddlee@seas.upenn.edu)

1 Introduction

Many problems in neural computation and statistical learning involve optimizations with nonnegativity constraints. Examples include large margin classification by support vector machines (Vapnik, 1998), density estimation in Bayesian networks (Bauer et al., 1997), dimensionality reduction by nonnegative matrix factorization (Lee and Seung, 1999), and acoustic echo cancellation (Lin et al., 2004). The optimizations for these problems cannot be solved in closed form; thus, iterative learning rules are required that converge in the limit to actual solutions.

The simplest such learning rule is gradient descent. Minimizing an objective function $F(\mathbf{v})$ by gradient descent involves the additive update:

$$v_i \leftarrow v_i - \eta(\partial F/\partial v_i) \quad , \quad (1)$$

where $\eta > 0$ is a positive learning rate, and all the elements of the parameter vector $v = (v_1, v_2, \dots, v_N)$ are updated in parallel. Gradient descent is not particularly well suited to constrained optimizations, however, because the additive update in eq. (1) can lead to violations of the constraints. A simple extension enforces the nonnegativity constraints:

$$v_i \leftarrow \max(v_i - \eta(\partial F/\partial v_i), 0) \quad . \quad (2)$$

The update rule in eq.(2) is a special instance of gradient projection methods (Bertsekas, 1999; Serafini et al., 2005). The nonnegativity constraints are enforced by projecting the gradient-based updates in eq. (1) onto the convex feasible set—namely, the nonnegative orthant $v_i \geq 0$. The projected gradient updates also depend on a learning rate parameter η .

For optimizations with nonnegativity constraints, an equally simple but more appropriate learn-

ing rule involves the so-called Exponentiated Gradient (EG) (Kivinen and Warmuth, 1997):

$$v_i \leftarrow v_i e^{-\eta(\partial F/\partial v_i)} . \quad (3)$$

Eq. (3) is an example of a multiplicative update. Because the elements of the exponentiated gradient are always positive, this update naturally enforces the nonnegativity constraints on v_i . By taking the logarithm of both sides of eq. (3), we can view the EG update as an additive update¹ in the log domain:

$$\log v_i \leftarrow \log v_i - \eta(\partial F/\partial v_i) . \quad (4)$$

Multiplicative updates such as EG typically lead to faster convergence than additive updates (Kivinen and Warmuth, 1997) if the solution \mathbf{v}^* of the optimization problem is sparse, containing a large number of zero elements. Note, moreover, that sparse solutions are more likely to arise in problems with nonnegativity constraints because in these problems minima can emerge at $v_i^* = 0$ without the precise vanishing of the partial derivative $(\partial F/\partial v_i)|_{\mathbf{v}^*}$ (as would be required in an unconstrained optimization).

The EG update in eq. (3)—like gradient descent in eq. (1) and projected gradient descent in eq. (2)—depends on the explicit introduction of a learning rate $\eta > 0$. The size of the learning rate must be chosen to avoid divergent oscillations (if η is too large) and unacceptably slow convergence (if η is too small). The necessity of choosing a learning rate can be viewed as a consequence of the generality of these learning rules; they do not assume or exploit any structure in the objective function $F(\mathbf{v})$ beyond the fact that it is differentiable.

Not surprisingly, many objective functions in machine learning have structure that can be exploited in their optimizations—and in particular, by multiplicative updates. Such updates need not involve learning rates, and they may also involve intuitions rather different from the connection

¹This update differs slightly from gradient descent in the variable $u_i = \log v_i$, which would involve the partial derivative $\partial F/\partial u_i = v_i(\partial F/\partial v_i)$ as opposed to what appears in eq. (4).

between EG and gradient descent in eq. (3–4). For example, the Expectation-Maximization (EM) algorithm for latent variable models (Dempster et al., 1977) and the generalized iterative scaling (GIS) algorithm for logistic regression (Darroch and Ratcliff, 1972) can be viewed as multiplicative updates (Saul et al., 2003), but unlike the EG update, they can not be cast as simple variants of gradient descent in the log domain.

In this paper, we derive multiplicative updates for convex problems in quadratic programming where the optimization is confined to an axis-aligned region in the nonnegative orthant. Our multiplicative updates have the property that they improve the value of the objective function at each iteration and converge monotonically to the global minimum. Despite their simplicity, they differ strikingly in form from other multiplicative updates used in statistical learning, including EG, EM, and GIS. This paper provides a complete derivation and proof of convergence for the multiplicative updates, originally described in previous work (Sha et al., 2003a,b) The proof techniques should be of general interest to researchers in neural computation and statistical learning faced with problems in constrained optimization.

The basic problem that we study in this paper is quadratic programming with nonnegativity constraints:

$$\begin{aligned} \text{minimize} \quad & F(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T \mathbf{A} \mathbf{v} + \mathbf{b}^T \mathbf{v} \\ \text{subject to} \quad & \mathbf{v} \geq \mathbf{0} . \end{aligned} \tag{5}$$

The constraint indicates that the variable \mathbf{v} is confined to the nonnegative orthant. We assume that the matrix \mathbf{A} is symmetric and strictly positive definite, so that the objective function $F(\mathbf{v})$ in eq. (5) is bounded below and its optimization is convex. In particular, it has one global minimum and no local minima.

Monotonically convergent multiplicative updates for minimizing eq. (5) were previously developed for the special case of nonnegative matrix factorization (NMF) (Lee and Seung, 2001). In this setting, the matrix elements of \mathbf{A} are nonnegative and the vector elements of \mathbf{b} are negative. The updates for NMF are derived from an auxiliary function similar to the one used in EM algorithms.

They take the simple, elementwise multiplicative form:

$$v_i \leftarrow \left[\frac{|b_i|}{(\mathbf{A}\mathbf{v})_i} \right] v_i, \tag{6}$$

which is guaranteed to preserve the nonnegativity constraints on \mathbf{v} . The validity of these updates for NMF hinges on the assumption that the matrix elements of \mathbf{A} are nonnegative: otherwise, the denominator in eq. (6) could become negative, leading to a violation of the nonnegativity constraints on \mathbf{v} .

In this paper, we generalize the multiplicative updates in eq. (6) to a wider range of problems in NQP. Our updates assume only that the matrix \mathbf{A} is positive semidefinite: in particular, it may have negative elements off the diagonal, and the vector \mathbf{b} may have both positive and negative elements. Despite the greater generality of our updates, they retain a simple, elementwise multiplicative form. The multiplicative factors in the updates involve only two matrix-vector multiplications and reduce to eq. (6) for the special case of NMF. The updates can also be extended in a straightforward way to the more general problem of NQP with upper bound constraints on the variable $\mathbf{v} \leq \ell$. Under these additional constraints, the variable \mathbf{v} is restricted to an axis-aligned “box” in the nonnegative orthant with opposing vertices at the origin and the nonnegative vector ℓ .

We prove that our multiplicative updates converge monotonically to the global minimum of the objective function for NQP. The proof relies on constructing an auxiliary function, as in earlier proofs for EM and NMF algorithms (Dempster et al., 1977; Lee and Seung, 2001). In general, monotonic improvement in an auxiliary function only suffices to establish convergence to a local stationary point, not necessarily a global minimum. For our updates, however, we are able to prove global convergence by exploiting the particular structure of their fixed points as well as the convexity of the objective function.

The rest of this paper is organized as follows. In section 2, we present the multiplicative updates and develop some simple intuitions behind their form. The updates are then derived more formally

and their convergence properties established in section 3, which completes the proofs sketched in earlier papers (Sha et al., 2003a,b). In section 4, we briefly describe some applications to problems in signal processing (Lin et al., 2004) and pattern recognition (Cristianini and Shawe-Taylor, 2000). Finally, in section 5, we conclude by summarizing the main advantages of our approach.

2 Algorithm

We begin by presenting the multiplicative updates for the basic problem of NQP in eq. (5). Some simple intuitions behind the updates are developed by analyzing the Kuhn-Karesh-Tucker (KKT) conditions for this problem. We then extend the multiplicative updates to handle the more general problem of NQP with additional upper bound constraints $\mathbf{v} \leq \ell$.

2.1 Updates for NQP

The multiplicative updates for NQP are expressed in terms of the positive and negative components of the matrix \mathbf{A} . In particular, let \mathbf{A}^+ and \mathbf{A}^- denote the *nonnegative* matrices with elements:

$$A_{ij}^+ = \begin{cases} A_{ij} & \text{if } A_{ij} > 0, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad A_{ij}^- = \begin{cases} |A_{ij}| & \text{if } A_{ij} < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

It follows that $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$. In terms of these nonnegative matrices, the objective function in (5) can be decomposed as the combination of three terms, which we write as

$$F(\mathbf{v}) = F_a(\mathbf{v}) + F_b(\mathbf{v}) - F_c(\mathbf{v}) \quad (8)$$

for reasons that will become clear shortly. We use the first and third terms in eq. (8) to “split” the quadratic piece of $F(\mathbf{v})$, and the second term to capture the linear piece:

$$\begin{aligned} F_a(\mathbf{v}) &= \frac{1}{2}\mathbf{v}^T \mathbf{A}^+ \mathbf{v}, \\ F_b(\mathbf{v}) &= \mathbf{b}^T \mathbf{v}, \\ F_c(\mathbf{v}) &= \frac{1}{2}\mathbf{v}^T \mathbf{A}^- \mathbf{v}. \end{aligned} \tag{9}$$

The decomposition (8) follows trivially from the definitions in eqs. (7) and (9). The gradient of $F(\mathbf{v})$ can be similarly decomposed in terms of contributions from these three pieces. We have chosen our notation in eq. (9) so that $b_i = \partial F_b / \partial v_i$; for the quadratic terms in the objective function, we define the corresponding derivatives:

$$a_i = \frac{\partial F_a}{\partial v_i} = (\mathbf{A}^+ \mathbf{v})_i, \tag{10}$$

$$c_i = \frac{\partial F_c}{\partial v_i} = (\mathbf{A}^- \mathbf{v})_i. \tag{11}$$

Note that the partial derivatives in eqs. (10) and (11) are guaranteed to be nonnegative when evaluated at vectors \mathbf{v} in the nonnegative orthant. The multiplicative updates are expressed in terms of these partial derivatives as:

$$v_i \leftarrow \left[\frac{-b_i + \sqrt{b_i^2 + 4a_i c_i}}{2a_i} \right] v_i. \tag{12}$$

Note that these updates reduce to the special case of eq. (6) for NMF when the matrix \mathbf{A} has no negative elements.

The updates in eq. (12) are meant to be applied in parallel to all the elements of \mathbf{v} . They are remarkably simple to implement and notable for their absence of free parameters or heuristic criteria that must be tuned to ensure convergence. Since $a_i \geq 0$ and $c_i \geq 0$, it follows that the multiplicative prefactor in eq. (12) is always nonnegative; thus, the optimization remains confined to the feasible

region for NQP. As we show in section 3, moreover, these updates are guaranteed to decrease the value of $F(\mathbf{v})$ at each iteration.

There is a close link between the sign of the partial derivative $\partial F/\partial v_i$ and the effect of the multiplicative update on v_i . In particular, using the fact that $\partial F/\partial v_i = a_i + b_i - c_i$, it is easy to show that the update decreases v_i if $\partial F/\partial v_i > 0$ and increases v_i if $\partial F/\partial v_i < 0$. Thus, the multiplicative update in eq. (12) moves each element v_i in an opposite direction to its partial derivative.

2.2 Fixed Points

Further intuition for the updates in eq. (12) can be gained by examining their fixed points. Let m_i denote the multiplicative prefactor inside the brackets on the right-hand-side of eq. (12). Fixed points of the updates occur when either (i) $v_i = 0$ or (ii) $m_i = 1$. What does the latter condition imply? Note that the expression for m_i is simply the quadratic formula for the larger root of the polynomial $p(m) = a_i m^2 + b_i m - c_i$. Thus $m_i = 1$ implies that $a_i + b_i - c_i = 0$. From the definitions in eqs. (8–11), moreover, it follows that $\partial F/\partial v_i = 0$. Thus, the two criteria for fixed points can be restated as (i) $v_i = 0$ or (ii) $\partial F/\partial v_i = 0$. These are consistent with the Karush-Kuhn-Tucker (KKT) conditions for the NQP problem in eq. (5), as we now show.

Let λ_i denote the Lagrange multiplier used to enforce the nonnegativity constraint on v_i . The KKT conditions are given by:

$$\mathbf{A}\mathbf{v} + \mathbf{b} = \boldsymbol{\lambda}, \quad \mathbf{v} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \circ \mathbf{v} = \mathbf{0}, \quad (13)$$

in which \circ stands for elementwise vector multiplication. A necessary and sufficient condition for \mathbf{v} to solve eq. (5) is that there exists a vector $\boldsymbol{\lambda}$ such that \mathbf{v} and $\boldsymbol{\lambda}$ satisfy this system. It follows from eq. (13) that the gradient of $F(\mathbf{v})$ at its minimum is nonnegative: $\nabla F = \mathbf{A}\mathbf{v} + \mathbf{b} \geq \mathbf{0}$. Moreover, for inactive constraints (corresponding to elements of the minimizer that are strictly positive), the corresponding partial derivatives of the objective function must vanish: $\partial F/\partial v_i = 0$ if $v_i > 0$.

Thus, the KKT conditions imply that (i) $v_i = 0$ or (ii) $\partial F/\partial v_i = 0$, and any solution satisfying the KKT conditions corresponds to a fixed point of the multiplicative updates, though not vice versa.

2.3 Upper Bound Constraints

The multiplicative updates in eq. (12) can also be extended to incorporate upper bound constraints of the form $\mathbf{v} \leq \ell$. A simple way of enforcing such constraints is to clip the output of the updates in eq. (12):

$$v_i \leftarrow \min \left\{ \ell_i, \left[\frac{-b_i + \sqrt{b_i^2 + 4a_i c_i}}{2a_i} \right] v_i \right\} . \quad (14)$$

As we show in the next section, this clipped update is also guaranteed to decrease the objective function $F(\mathbf{v})$ in eq. (5) if it results in a change of v_i .

3 Convergence Analysis

In this section, we prove that the multiplicative updates in eq. (12) converge monotonically to the global minimum of the objective function $F(\mathbf{v})$. Our proof is based on the derivation of an auxiliary function which provides an upper bound on the objective function. Similar techniques have been used to establish the convergence of many algorithms in statistical learning (e.g., the Expectation-Maximization algorithm (Dempster et al., 1977) for maximum likelihood estimation) and nonnegative matrix factorization (Lee and Seung, 2001). The proof is composed of two parts. We first show that the multiplicative updates monotonically decrease the objective function $F(\mathbf{v})$. Then we show that the updates converge to the global minimum. We assume throughout the paper that the matrix \mathbf{A} is positive definite such that the objective function is convex. (Though Theorem 2 does not depend on this assumption, convexity is used to establish the stronger convergence results that follow.)

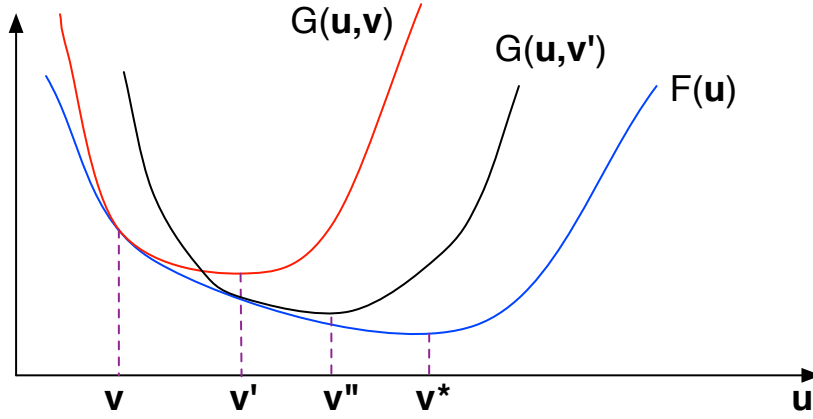


Figure 1: Using an auxiliary function $G(\mathbf{u}, \mathbf{v})$ to minimize an objective function $F(\mathbf{v})$. The auxiliary function is constructed around the current estimate of the minimizer; the next estimate is found by minimizing the auxiliary function, which provides an upper bound on the objective function. The procedure is iterated until it converges to a stationary point (generally, a local minimum) of the objective function.

3.1 Monotonic Convergence

An auxiliary function $G(\mathbf{u}, \mathbf{v})$ for the objective function in eq. (5) has two crucial properties: (i) $F(\mathbf{u}) \leq G(\mathbf{u}, \mathbf{v})$ and (ii) $F(\mathbf{v}) = G(\mathbf{v}, \mathbf{v})$ for all positive vectors \mathbf{u} and \mathbf{v} . From such an auxiliary function, we can derive the update rule $\mathbf{v}' = \arg \min_{\mathbf{u}} G(\mathbf{u}, \mathbf{v})$ which never increases (and generally decreases) the objective function $F(\mathbf{v})$:

$$F(\mathbf{v}') \leq G(\mathbf{v}', \mathbf{v}) \leq G(\mathbf{v}, \mathbf{v}) = F(\mathbf{v}) . \quad (15)$$

By iterating this update, we obtain a series of values of \mathbf{v} that improve the objective function. Figure 3.1 graphically illustrates how the auxiliary function $G(\mathbf{u}, \mathbf{v})$ is used to compute a minimum of the objective function $F(\mathbf{v})$ at $\mathbf{v} = \mathbf{v}^*$.

To derive an auxiliary function for NQP, we first decompose the objective function $F(\mathbf{v})$ in eq. (5) into three terms as in eqs. (8–9) and then derive the upper bounds for each of them separately. The following two lemmas establish the bounds relevant to the quadratic terms $F_a(\mathbf{u})$

and $F_c(\mathbf{u})$.

Lemma 1 *Let \mathbf{A}^+ denote the matrix composed of the positive elements of the matrix \mathbf{A} , as defined in eq. (7). Then for all positive vectors \mathbf{u} and \mathbf{v} , the quadratic form $F_a(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T \mathbf{A}^+ \mathbf{u}$ satisfies the following inequality:*

$$F_a(\mathbf{u}) \leq \frac{1}{2} \sum_i \frac{(\mathbf{A}^+ \mathbf{v})_i}{v_i} u_i^2 . \quad (16)$$

Proof. Let δ_{ij} denote the Kronecker delta function, and let \mathbf{K} be the diagonal matrix with elements

$$K_{ij} = \delta_{ij} \frac{(\mathbf{A}^+ \mathbf{v})_i}{v_i} . \quad (17)$$

Since $F_a(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T \mathbf{A}^+ \mathbf{u}$, the inequality in eq. (16) is equivalent to the statement that the matrix $(\mathbf{K} - \mathbf{A}^+)$ is positive semidefinite. Consider the matrix \mathbf{M} whose elements

$$M_{ij} = v_i(K_{ij} - A_{ij}^+)v_j \quad (18)$$

are obtained by rescaling componentwise the elements of $(\mathbf{K} - \mathbf{A}^+)$. Thus, $(\mathbf{K} - \mathbf{A}^+)$ is positive semidefinite if \mathbf{M} is positive semidefinite. We note that for all vectors \mathbf{u} :

$$\mathbf{u}^T \mathbf{M} \mathbf{u} = \sum_{ij} u_i v_i (K_{ij} - A_{ij}^+) v_j u_j \quad (19)$$

$$= \sum_{ij} \delta_{ij} (\mathbf{A}^+ \mathbf{v})_i u_i u_j v_j - \sum_{ij} A_{ij}^+ v_i v_j u_i u_j \quad (20)$$

$$= \sum_{ij} A_{ij}^+ v_i v_j u_i^2 - \sum_{ij} A_{ij}^+ v_i v_j u_i u_j \quad (21)$$

$$= \frac{1}{2} \sum_{ij} A_{ij}^+ v_i v_j (u_i - u_j)^2 \geq 0 . \quad (22)$$

Thus, $(\mathbf{K} - \mathbf{A}^+)$ is positive semidefinite, proving the bound in eq. (16). An alternative proof that $(\mathbf{K} - \mathbf{A}^+)$ is semidefinite positive can also be made by appealing to the Frobenius-Perron Theorem (Lee and Seung, 2001). \square

For the terms related to the negative elements in the matrix \mathbf{A} , we have following result.

Lemma 2 Let \mathbf{A}^- denote the matrix composed of the negative elements of the matrix \mathbf{A} , as defined in eq. (7). Then for all positive vectors \mathbf{u} and \mathbf{v} , the quadratic form $F_c(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T \mathbf{A}^- \mathbf{u}$ satisfies the following inequality:

$$-F_c(\mathbf{u}) \leq -\frac{1}{2} \sum_{ij} A_{ij}^- v_i v_j \left(1 + \log \frac{u_i u_j}{v_i v_j} \right) . \quad (23)$$

Proof. To prove this bound, we use the simple inequality: $z \geq 1 + \log z$. Substituting $z = u_i u_j / (v_i v_j)$ into this inequality gives:

$$u_i u_j \geq v_i v_j \left(1 + \log \frac{u_i u_j}{v_i v_j} \right) . \quad (24)$$

Substituting the above inequality into $F_c(\mathbf{u}) = \frac{1}{2} \sum_{ij} A_{ij}^- u_i u_j$ and noting the negative sign, we arrive at the bound in eq. (23). \square

Combining lemmas 1 and 2, and noting that $F_b(\mathbf{u}) = \sum_i b_i u_i$, we have proved the following theorem.

Theorem 1 Define a function $G(\mathbf{u}, \mathbf{v})$ on positive vectors \mathbf{u} and \mathbf{v} by:

$$G(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \sum_i \frac{(\mathbf{A}^+ \mathbf{v})_i}{v_i} u_i^2 - \frac{1}{2} \sum_{ij} A_{ij}^- v_i v_j \left(1 + \log \frac{u_i u_j}{v_i v_j} \right) + \sum_i b_i u_i \quad (25)$$

Then $G(\mathbf{u}, \mathbf{v})$ is an auxiliary function for the function $F(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T \mathbf{A} \mathbf{v} + \mathbf{b}^T \mathbf{v}$, satisfying $F(\mathbf{u}) \leq G(\mathbf{u}, \mathbf{v})$ and $F(\mathbf{v}) = G(\mathbf{v}, \mathbf{v})$.

As explained previously, a new estimate that improves the objective function $F(\mathbf{v})$ at its current estimate \mathbf{v} is obtained by minimizing the auxiliary function $G(\mathbf{u}, \mathbf{v})$ with respect to its first argument \mathbf{u} , as shown by following theorem.

Theorem 2 Given a positive vector \mathbf{v} and a mapping $\mathbf{v}' = \mathcal{M}(\mathbf{v})$ such that $\mathbf{v}' = \arg \min_{\mathbf{u}} G(\mathbf{u}, \mathbf{v})$, we have,

$$F(\mathbf{v}') \leq F(\mathbf{v}). \quad (26)$$

Moreover, if $\mathbf{v}' \neq \mathbf{v}$, then the inequality holds strictly. Therefore the objective function is strictly decreased unless at the fixed point of the mapping $\mathcal{M}(\mathbf{v})$, where $\mathbf{v} = \mathcal{M}(\mathbf{v})$. The mapping $\mathcal{M}(v)$ takes the form of eq. (12) if \mathbf{v} is constrained only to be nonnegative and takes the form of eq. (14) of \mathbf{v} is box-constrained.

Proof. The inequality in eq. (26) is a direct result from the definition of the auxiliary function and its relation to the objective function. The derivation in eq. (15) is reproduced here for easy reference:

$$F(\mathbf{v}') \leq G(\mathbf{v}', \mathbf{v}) \leq G(\mathbf{v}, \mathbf{v}) = F(\mathbf{v}) . \quad (27)$$

To show that the objective function is *strictly* decreased if the new estimate \mathbf{v}' is not the same as the old estimate \mathbf{v} , we must also show that the auxiliary function is strictly decreased: namely, if $\mathbf{v}' \neq \mathbf{v}$, then $G(\mathbf{v}', \mathbf{v}) < G(\mathbf{v}, \mathbf{v})$. This can be proved by further examining the properties of the auxiliary function.

We begin by showing that $G(\mathbf{u}, \mathbf{v})$ is the sum of strictly convex functions of \mathbf{u} . For a strictly convex function, the minimizer is unique, and the minimum is strictly less than any other values of the function. We reorganize the expression of the auxiliary function $G(\mathbf{u}, \mathbf{v})$ given by eq. (25) such that there are no interaction terms among the variables u_i :

$$G(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \sum_i \frac{(\mathbf{A}^+ \mathbf{v})_i}{v_i} u_i^2 - \sum_i (\mathbf{A}^- \mathbf{v})_i v_i \log \frac{u_i}{v_i} + \sum_i b_i u_i - \frac{1}{2} \sum_{ij} A_{ij}^- v_i v_j.$$

We identify the auxiliary function with $G(\mathbf{u}, \mathbf{v}) = \sum_i G_i(u_i) - \frac{1}{2} \mathbf{v}^T \mathbf{A}^- \mathbf{v}$, where $G_i(u_i)$ is a single-variable function of u_i :

$$G_i(u_i) = \frac{1}{2} \frac{(\mathbf{A}^+ \mathbf{v})_i}{v_i} u_i^2 - (\mathbf{A}^- \mathbf{v})_i v_i \log \frac{u_i}{v_i} + b_i u_i \quad (28)$$

Note that the minimizer of $G(\mathbf{u}, \mathbf{v})$ can be easily found by minimizing each $G_i(u_i)$ separately, namely, $v'_i = \arg \min_{u_i} G_i(u_i)$. Moreover, we will show that $G_i(u_i)$ is strictly convex in u_i . To see

this, we examine its second derivative with respect to u_i :

$$G_i''(u_i) = \frac{(\mathbf{A}^+\mathbf{v})_i}{v_i} + \frac{(\mathbf{A}^-\mathbf{v})_i}{u_i^2}v_i. \quad (29)$$

For a positive vector \mathbf{v} , $(\mathbf{A}^+\mathbf{v})_i$ and $(\mathbf{A}^-\mathbf{v})_i$ cannot be simultaneously equal to zero. Otherwise, the i -th row of \mathbf{A} is all-zero, contradicting our assumption that \mathbf{A} is strictly convex. This implies that $G_i''(u_i)$ is strictly positive and $G_i(u_i)$ is strictly convex in u_i .

The theorem 2 follows directly from the above observation. In particular, if v_i is not a minimizer of $G_i(u_i)$, then $v'_i \neq v_i$ and $G_i(v'_i) < G_i(u_i)$. Since the auxiliary function $G(\mathbf{u}, \mathbf{v})$ is the sum of all the individual terms $G_i(u_i)$ plus a term independent of \mathbf{u} , we have shown that $G(\mathbf{v}', \mathbf{v})$ is strictly less than $G(\mathbf{v}, \mathbf{v})$ if $\mathbf{v}' \neq \mathbf{v}$. This leads to $F(\mathbf{v}') < F(\mathbf{v})$.

As explained previously, the minimizer \mathbf{v}' can be computed by finding the minimizer of each individual term $G_i(u_i)$. Computing the derivative of $G_i(u_i)$ with respect to u_i , setting it to zero, and solving for u_i leads to the multiplicative updates in eq. (12). Minimizing $G_i(u_i)$ subject to box constraints $u_i \in [0, \ell_i]$ leads to the clipped multiplicative updates in eq. (14). \square

3.2 Global Convergence

The multiplicative updates define a mapping \mathcal{M} from the current estimate \mathbf{v} of the minimizer to a new estimate \mathbf{v}' . By iteration, the updates generate a sequence of estimates $\{\mathbf{v}^1, \mathbf{v}^2, \dots\}$, satisfying $\mathbf{v}^{k+1} = \mathcal{M}(\mathbf{v}^k)$. The sequence monotonically improves the objective function $F(\mathbf{v})$. Since the sequence $\{F(\mathbf{v}^1), F(\mathbf{v}^2), \dots\}$ is monotonically decreasing and is bounded below by the global minimum value of $F(\mathbf{v})$, the sequence converges to some value when k is taken to the limit of infinity. While establishing monotonic convergence of the sequence, however, the above observation does not rule out the possibility that the sequence converges to spurious fixed points of the iterative procedure $\mathbf{v}^{k+1} = \mathcal{M}(\mathbf{v}^k)$ that are *not* the global minimizer of the objective function. In this section, we prove that the multiplicative updates do indeed converge to the global minimizer

and attain the global minimum of the objective function. (Note: the technical details of this section are not necessary for understanding how to derive or implement the multiplicative updates.)

3.2.1 Outline of the Proof

Our proof relies on a detailed investigation of the fixed points of the mapping \mathcal{M} defined by the multiplicative updates. In what follows, we distinguish between the “spurious” fixed points of \mathcal{M} that violate the KKT conditions versus the unique fixed point of \mathcal{M} that satisfies the KKT conditions and attains the global minimum value of $F(\mathbf{v})$. The basic idea of the proof is to rule out both the possibility that the multiplicative updates converge to a spurious fixed point, as well as the possibility that they lead to oscillations among two or more fixed points.

Our proof consists of three stages. First, we show that any accumulation point of the sequence $\{\mathbf{v}^1, \mathbf{v}^2, \dots\}$ must be a fixed point of the multiplicative updates, either a spurious fixed point or the global minimizer. Such a result is considerably weaker than global convergence to the minimizer. Second, we show that there do not exist convergent subsequences \mathcal{S} of the mapping \mathcal{M} with spurious fixed points as accumulation points. In particular, we show that if such a sequence \mathcal{S} converges to a spurious fixed point, then it must have a subsequence converging to a different fixed point, yielding a contradiction. Therefore, the accumulation point of any convergent subsequence must be the global minimizer. Third, we strengthen the result on subsequence convergence and show that the sequence $\{\mathbf{v}^1, \mathbf{v}^2, \dots\}$ converges to the global minimizer.

Our proof starts from Zangwill’s Convergence Theorem (Zangwill, 1969), a well-known convergence result for general iterative methods, but our final result does not follow simply from this general framework. We review Zangwill’s Convergence Theorem in appendix A. The application of this theorem in our setting yields the weaker result in the first step of our proof—namely, convergence to a fixed point of the multiplicative updates. As explained in the appendix, however, Zangwill’s Convergence Theorem does not exclude the possibility of convergence to spurious fixed points. We derive our stronger result of global convergence by exploiting the particular structure

of the objective function and the multiplicative update rules for NQP. A key step (Lemma 4) in our proof is to analyze the mapping \mathcal{M} on sequences that are in the vicinity of spurious fixed points. Our analysis appeals repeatedly to the specific properties of the objective function and the mapping induced by the multiplicative updates.

The following notation and preliminary observations will be useful. We use $\{\mathbf{v}^k\}_1^\infty$ to denote the sequence $\{\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^k, \dots\}$ and $\{F(\mathbf{v}^k)\}_1^\infty$ to denote the corresponding sequence $\{F(\mathbf{v}^1), F(\mathbf{v}^2), \dots, F(\mathbf{v}^k), \dots\}$. We assume that the matrix \mathbf{A} in eq. (5) is strictly positive definite so that the objective function has a unique global minimum. From this it also follows that the matrix \mathbf{A}^+ in eq. (7) has strictly positive elements along the diagonal.

3.2.2 Positivity

Our proof will repeatedly invoke the observation that for a strictly positive vector \mathbf{v} , the multiplicative updates in eq. (12) yield a strictly positive vector $\mathbf{v}' = \mathcal{M}(\mathbf{v})$. There is one exception to this rule which we address here. Starting from a strictly positive vector \mathbf{v} , the multiplicative updates will set an element $v'_i = 0$ directly to zero in the case that $b_i \geq 0$ and the i th row of the matrix \mathbf{A} has no negative elements. It is easy to verify in this case, however, that the global minimizer \mathbf{v}^* of the objective function does have a zero element at $\mathbf{v}_i^* = 0$. Once an element is zeroed by the multiplicative updates, it remains zero under successive updates. In effect, when this happens, the original problem in NQP reduces to a smaller problem—of dimensionality equal to the number of non-trivial modes in the original system. Without loss of generality, therefore, we will assume in what follows that any trivial degrees of freedom have already been removed from the problem. More specifically, we will assume that the i th row of the matrix \mathbf{A} has one or more negative elements whenever $b_i \geq 0$, and that consequently, a strictly positive vector \mathbf{v} is always mapped to a strictly positive vector $\mathbf{v}' = \mathcal{M}(\mathbf{v})$.

3.2.3 Accumulation Points of $\{\mathbf{v}^k\}_1^\infty$

The following lemma is a direct result of Zangwill's Convergence Theorem, as reviewed in the appendix A. It establishes the link between the accumulation points of $\{\mathbf{v}^k\}_1^\infty$ and the fixed points of \mathcal{M} .

Lemma 3 *Given a point \mathbf{v}^1 , suppose the update rule in eq. (12) generates a sequence $\{\mathbf{v}^k\}_1^\infty$, then either the algorithm terminates at a fixed point of \mathcal{M} or the accumulation point of any convergent subsequence in $\{\mathbf{v}^k\}_1^\infty$ is a fixed point of \mathcal{M} .*

Proof. If there is a $k \geq 1$ such that \mathbf{v}^k is a fixed point of \mathcal{M} , then the update rule terminates. Therefore, we consider the case that an infinite sequence is generated and show how to apply Zangwill's Convergence Theorem.

Let \mathcal{M} be the update procedure in Zangwill's Convergence Theorem. We first verify that the sequence $\{\mathbf{v}^k\}_1^\infty$ generated by \mathcal{M} is in a compact set. Because $\{F(\mathbf{v}^k)\}_1^\infty$ is a monotonically decreasing sequence, it follows that for all k :

$$\mathbf{v}^k \in \Omega = \{\mathbf{v} | F(\mathbf{v}) \leq F(\mathbf{v}^1)\}. \quad (30)$$

Note that the set Ω is compact because it defines an ellipsoid confined to the positive orthant.

We define the desired set \mathbf{S} be the collection of all the fixed points of \mathcal{M} . If $\mathbf{v} \notin \mathbf{S}$, then from Theorem 2, we have that $F(\mathcal{M}(\mathbf{v})) < F(\mathbf{v})$. On the other hand, if $\mathbf{v} \in \mathbf{S}$, then we have that $F(\mathcal{M}(\mathbf{v})) = F(\mathbf{v})$. This shows that the mapping \mathcal{M} maintains strict monotonicity of the objective function outside of the desired set.

The last condition to verify is that \mathcal{M} is closed at \mathbf{v} if \mathbf{v} is not in the desired set. Note that \mathcal{M} is continuous if $\mathbf{v} \neq \mathbf{0}$. Therefore, if the origin $\mathbf{v} = \mathbf{0}$ is a fixed point of \mathcal{M} , then \mathcal{M} is closed outside the desired set.

If the origin is not a fixed point of \mathcal{M} , then it cannot be the global minimizer. Moreover, we can choose the initial estimate \mathbf{v}^1 such that $F(\mathbf{v}^1) < F(\mathbf{0})$. With this choice, it follows from the

monotonicity of \mathcal{M} that the origin is not contained in Ω and that \mathcal{M} is continuous on Ω .

Either way we have shown that \mathcal{M} is closed on a proper domain. Therefore, we can apply Zangwill's Convergence Theorem to the mapping \mathcal{M} restricted on Ω : the limit of any convergent subsequence in $\{\mathbf{v}^k\}_1^\infty$ is in the desired set, or equivalently, a fixed point of \mathcal{M} . \square

REMARK. It is easy to check whether the global minimizer occurs at the origin with value $F(\mathbf{0}) = 0$. In particular, if all the elements of \mathbf{b} are nonnegative, then the origin is the global minimizer. On the other hand, if there is a nonnegative element of \mathbf{b} , then we can choose the initial estimate \mathbf{v}^1 such that $F(\mathbf{v}^1) < F(\mathbf{0})$. For example, suppose $b_k < 0$. Then we can choose \mathbf{v}^1 such its k^{th} element is σ and all other elements are τ . A positive σ and τ can be found such that $F(\mathbf{v}^1) < 0$ by noting:

$$\begin{aligned} F(\mathbf{v}^1) &= \frac{1}{2} \sum_{i,j \neq k} A_{ij} \tau^2 + \sum_{i \neq k} A_{ik} \tau \sigma + \sum_{i \neq k} b_i \tau + \frac{1}{2} A_{kk} \sigma^2 + b_k \sigma \\ &\leq \frac{1}{2} \sum_{ij} |A_{ij}| \tau^2 + \left(\sum_i |A_{ik}| \sigma + \sum_i |b_i| \right) \tau + \left(\frac{1}{2} A_{kk} \sigma + b_k \right) \sigma. \end{aligned} \quad (31)$$

Note that if we choose a positive $\sigma < -2b_k/A_{kk}$, we can always find a positive τ such that $F(\mathbf{v}^1) < 0$ because the rightmost term of the inequality in eq. (31) is negative and the left and middle terms vanish as $\tau \rightarrow 0^+$.

3.2.4 Properties of the Fixed Points

As stated in Section 2.2, the minimizer of $F(\mathbf{v})$ satisfies the KKT conditions and corresponds to a fixed point of the mapping \mathcal{M} defined by the multiplicative update rule in eq. (12). The mapping \mathcal{M} , however, also has fixed points that do not satisfy the KKT conditions. We refer to these as spurious fixed points.

Lemma 3 states that any convergent subsequence in $\{\mathbf{v}^k\}_1^\infty$ must have a fixed point of \mathcal{M} as its accumulation point. To prove that the multiplicative updates converge to the global minimizer, we will show that spurious fixed points cannot be accumulation points. Our strategy is to demonstrate

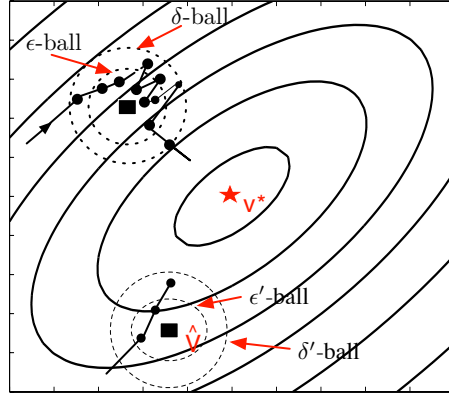


Figure 2: Fixed points of the multiplicative updates: the global minimizer \mathbf{v}^* , indicated by a star, and spurious fixed points $\hat{\mathbf{v}}$ and $\hat{\mathbf{v}}'$, indicated by squares. Contour lines of the objective function are shown as ellipses. A hypothetical sequence $\{\mathbf{v}^k\}_1^\infty$ with a subsequence converging to the spurious fixed point $\hat{\mathbf{v}}$ is represented by solid lines connecting small black circles. The δ -ball around the spurious fixed point $\hat{\mathbf{v}}$ does not intersect the δ' -ball around the other spurious fixed point $\hat{\mathbf{v}}'$.

that any subsequence \mathcal{S} converging to a spurious fixed point must itself have a subsequence “running away” from the fixed point. The idea of the proof is shown schematically in Figure 2. The star in the center of the figure denotes the global minimizer \mathbf{v}^* . Black squares denote spurious fixed points $\hat{\mathbf{v}}$ and $\hat{\mathbf{v}}'$. The figure also shows a hypothetical subsequence that converges to the spurious fixed point $\hat{\mathbf{v}}$.

At a high level, the proof (by contradiction) is as follows. Suppose that there exists a convergent subsequence as shown in the figure. Then we can draw a very small ϵ -ball around the spurious fixed point $\hat{\mathbf{v}}$ containing an infinite number of elements of the subsequence. We will show that under the mapping \mathcal{M} , the subsequence must have an infinite number of successors that are outside the ϵ -ball yet inside a δ -ball where $\delta > \epsilon$. This bounded successor sequence must have a subsequence converging to an accumulation point, which by Lemma 3, must also be a fixed point. However, we can choose the δ -ball to be sufficiently small such that the annulus between the ϵ -ball and δ -ball contains no other fixed points. This yields a contradiction.

More formally, we begin by proving the following lemma:

Lemma 4 Let \mathbf{v}^1 denote a positive initial vector satisfying $F(\mathbf{v}^1) < F(\mathbf{0})$. Suppose that the sequence $\mathbf{v}^{k+1} = \mathcal{M}(\mathbf{v}^k)$ generated by the iterative update in eq. (12) has a subsequence that converges to the spurious fixed point $\hat{\mathbf{v}}$. Then there exists an $\epsilon > 0$ and a $\delta > 0$ such that, for every $\mathbf{v} \neq \hat{\mathbf{v}}$ such that $\|\mathbf{v} - \hat{\mathbf{v}}\| < \epsilon$, there exists an integer $p \geq 1$ such that $\epsilon < \|\mathcal{M}^p(\mathbf{v}) - \hat{\mathbf{v}}\| < \delta$, where $\mathcal{M}^p(\mathbf{v})$ is p times composition of \mathcal{M} applied to \mathbf{v} : $\underbrace{\mathcal{M} \circ \mathcal{M} \cdots \circ \mathcal{M}}_p(\mathbf{v})$.

Proof. If $\hat{\mathbf{v}}$ is a fixed point, then either $\hat{v}_i \neq 0$ or $\hat{v}_i = 0$ for any i . If the former is true, as shown in Section 2.2, it follows that $(\partial F/\partial v_i)|_{\hat{\mathbf{v}}} = 0$. When $\hat{v}_i = 0$, then either $(\partial F/\partial v_i)|_{\hat{\mathbf{v}}} \geq 0$ or $(\partial F/\partial v_i)|_{\hat{\mathbf{v}}} < 0$. If $\hat{\mathbf{v}}$ is a spurious fixed point that violate the KKT conditions, then there exists at least one i such that,

$$\hat{v}_i = 0 \quad \text{and} \quad \left. \frac{\partial F}{\partial v_i} \right|_{\hat{\mathbf{v}}} < 0$$

Let $\epsilon_{\hat{\mathbf{v}}}$ be a small ball centered at $\hat{\mathbf{v}}$ with radius of ϵ , namely, $\epsilon_{\hat{\mathbf{v}}} = \{\mathbf{v} \mid \|\mathbf{v} - \hat{\mathbf{v}}\| < \epsilon\}$. By continuity, there exists an ϵ such that $(\partial F/\partial v_i) < 0$ for all $\mathbf{v} \in \epsilon_{\hat{\mathbf{v}}}$.

Let Γ be the image of $\epsilon_{\hat{\mathbf{v}}}$ under the mapping \mathcal{M} . Since \mathcal{M} is a continuous mapping, we can find a minimum ball $\delta_{\hat{\mathbf{v}}} = \{\mathbf{v} \mid \|\mathbf{v} - \hat{\mathbf{v}}\| < \delta\}$ to encircle Γ . We claim that the ϵ and δ satisfy the lemma.

As observed in Section 2.2, the multiplicative update increases v_i if $\partial F/\partial v_i$ is negative. Consider the sequence $\{\mathcal{M}(\mathbf{v}), \mathcal{M} \circ \mathcal{M}(\mathbf{v}), \dots, \mathcal{M}^k(\mathbf{v}), \dots\}$. The i -th component of the sequence is monotonically increasing until the condition $(\partial F/\partial v_i)$ becomes nonnegative. This happens only if the element of the sequence is outside of the $\epsilon_{\hat{\mathbf{v}}}$ ball. Thus for every $\mathbf{v} \in \epsilon_{\hat{\mathbf{v}}}$, the update will push v_i to larger and larger values until it escapes from the ball. Let p be the smallest integer such that

$$\mathcal{M}^{p-1}(\mathbf{v}) \in \epsilon_{\hat{\mathbf{v}}} \quad \text{and} \quad \mathcal{M}^p(\mathbf{v}) \notin \epsilon_{\hat{\mathbf{v}}}$$

By construction of the $\delta_{\hat{\mathbf{v}}}$ ball, \mathcal{M}^p is inside the ball since Γ is the image of the $\epsilon_{\hat{\mathbf{v}}}$ under the mapping \mathcal{M} and $\mathcal{M}^{p-1}(\mathbf{v}) \in \epsilon_{\hat{\mathbf{v}}}$. Therefore, $\mathcal{M}^p(\mathbf{v}) \in \Gamma \subset \delta_{\hat{\mathbf{v}}}$. \square

The size of the δ -ball depends on the spurious fixed point $\hat{\mathbf{v}}$ and ϵ . Can $\delta_{\hat{\mathbf{v}}}$ contain another

spurious fixed point $\hat{\mathbf{v}}'$? The following lemma shows that we can choose ϵ sufficiently small such that the ball $\delta_{\hat{\mathbf{v}}}$ contains no other fixed points.

Lemma 5 *If $\hat{\mathbf{v}}$ and $\hat{\mathbf{v}}'$ are two different spurious fixed points, let ϵ and δ be the radii for $\hat{\mathbf{v}}$ such that Lemma 4 holds and ϵ' and δ' be the radii for $\hat{\mathbf{v}}'$. There exists an $\epsilon_0 > 0$ such that $\max(\epsilon, \epsilon') \leq \epsilon_0$ and $\delta_{\hat{\mathbf{v}}} \cap \delta_{\hat{\mathbf{v}}'}$ is empty.*

Proof. It suffices to show that $\mathcal{M}(\mathbf{v})$ becomes arbitrarily close to $\hat{\mathbf{v}}$ as \mathbf{v} approaches $\hat{\mathbf{v}}$. Since \mathcal{M} is a bounded and continuous mapping, the image Γ of $\epsilon_{\hat{\mathbf{v}}}$ under \mathcal{M} becomes arbitrarily small as $\epsilon \rightarrow 0$. Note that $\hat{\mathbf{v}}$ is a fixed point of \mathcal{M} , we have $\hat{\mathbf{v}} \in \Gamma$. Therefore, the δ -ball centered at $\hat{\mathbf{v}}$ can be made arbitrarily small as \mathbf{v} approaches $\hat{\mathbf{v}}$.

Let us choose ϵ sufficiently small such that δ is less than the half of the distance between $\hat{\mathbf{v}}$ and $\hat{\mathbf{v}}'$ and choose ϵ' likewise, then the intersection of $\delta_{\hat{\mathbf{v}}}$ and $\delta_{\hat{\mathbf{v}}'}$ is empty. This is illustrated in Figure 2. \square

Because the spurious fixed points are separated by their δ -balls, we will show that the existence of a subsequence converging to a spurious fixed point leads to a contradiction. This observation leads to the following theorem.

Theorem 3 *If the matrix \mathbf{A} is strictly positive definite, then no convergent subsequence of $\{\mathbf{v}^k\}_1^\infty$ can have a spurious fixed point of the multiplicative updates as an accumulation point.*

Proof. The number of spurious fixed points is finite and bounded above by the number of ways of choosing zero elements of \mathbf{v} . Let $\Delta > 0$ be the minimum pairwise distance between these fixed points. By Lemma 4, we can choose a small ϵ such that the radius of the δ -ball for any spurious fixed point is less than $\Delta/2$. With this choice, the δ -balls at different spurious fixed points are non-overlapping.

Suppose there is a convergent subsequence $\{\mathbf{v}^k\} \subset \{\mathbf{v}^k\}_1^\infty$ such that $\mathbf{v}^k \rightarrow \hat{\mathbf{v}}$ with $k \in \mathcal{K}$, where \mathcal{K} is an index subset and $\hat{\mathbf{v}}$ is a spurious fixed point. Without loss of generality, we assume the whole subsequence is contained in the ϵ -ball of $\hat{\mathbf{v}}$.

For each element \mathbf{v}^k of the subsequence, by Lemma 4, there exists an integer p_k such that $\mathcal{M}^{p_k}(\mathbf{v}^k)$ is outside of the ϵ -ball yet inside the δ -ball. Consider the “successor” sequence $\mathcal{M}^{p_k}(\mathbf{v}^k)$ with $k \in \mathcal{K}$, schematically shown in Figure 2 as the black circles between the ϵ -ball and the δ -ball. The infinite successor sequence is bounded between the ϵ -ball and δ -ball and therefore must have a convergent subsequence. By Lemma 3, the accumulation point of this subsequence must be a fixed point of \mathcal{M} . However, this leads to contradiction. On one hand, the subsequence is outside of the ϵ -ball of $\hat{\mathbf{v}}$ so $\hat{\mathbf{v}}$ is not the accumulation point. On the other hand, the subsequence is inside the δ -ball of $\hat{\mathbf{v}}$: therefore, it cannot have any other spurious fixed point $\hat{\mathbf{v}}'$ as its accumulation point, because we have shown that all pairs of fixed points are separated by their respective δ -balls. Therefore, the accumulation point of the subsequence cannot be a fixed point. Thus, we arrive at a contradiction, showing that spurious fixed points cannot be accumulation points of any convergent subsequence. \square

3.2.5 Convergence to the Global Minimizer

We have shown that the only possible accumulation point of $\{\mathbf{v}^k\}_1^\infty$ is the global minimizer—namely, the fixed point of \mathcal{M} that satisfies the KKT conditions. We now show the sequence $\{\mathbf{v}^k\}_1^\infty$ itself does indeed converge to the global minimizer.

Theorem 4 *Suppose that the origin is not the global minimizer and that we choose a positive initial vector \mathbf{v}^1 such that $F(\mathbf{v}^1) < 0$. Then the sequence $\{\mathbf{v}^k\}_1^\infty$ converges to the global minimizer \mathbf{v}^* , and $\{F(\mathbf{v}^k)\}_1^\infty$ converges to the optimal value $F^* = F(\mathbf{v}^*)$.*

Proof. As shown in eq. (30), the infinite sequence $\{\mathbf{v}^k\}_1^\infty$ is a bounded set; therefore it must have an accumulation point. By the preceding theorem, the accumulation point of any convergent subsequence of $\{\mathbf{v}^k\}_1^\infty$ cannot be a spurious fixed point; thus, any convergent subsequence must converge to the fixed point that satisfies the KKT conditions—namely, the global minimizer. By monotonicity, it immediately follows that $\{F(\mathbf{v}^k)\}_1^\infty$ converges to the optimal value F^* of the

objective function.

Since $F^* < F(\hat{\mathbf{v}})$ for all spurious fixed points $\hat{\mathbf{v}}$, we can find an $\epsilon^* > 0$ such that the set

$$\Omega^* = \{\mathbf{v} | F(\mathbf{v}) \leq F^* + \epsilon^*\}$$

contains no spurious fixed points of \mathcal{M} . Moreover, since $\{F(\mathbf{v}^k)\}_1^\infty$ is a monotonically decreasing sequence converging to F^* , there exists a k_0 such that $\mathbf{v}^k \in \Omega^*$ for all $k \geq k_0$.

We now prove the theorem by contradiction. Suppose $\{\mathbf{v}^k\}_1^\infty$ does not converge to the minimizer \mathbf{v}^* . Then, there exists an $\epsilon > 0$ such that the set

$$\Psi = \{\mathbf{v}^k : \|\mathbf{v}^k - \mathbf{v}^*\| > \epsilon\}$$

has an infinite number of elements. In other words, there must be a subsequence of $\{\mathbf{v}^k\}_1^\infty$ in which every element has distance at least ϵ from the minimizer. Moreover, the intersection of Ψ and Ω^* must have an infinite number of elements. Note that by construction, Ω^* contains no fixed points other than the global minimizer, and Ψ does not contain global minimizer. Thus there are no fixed points in $\Psi \cap \Omega^*$. The infinite set $\Psi \cap \Omega^*$, however, is bounded, and therefore must have an accumulation point; by Lemma 3, this accumulation point must be a fixed point. This yields a contradiction. Hence, the set Ψ cannot have an infinite number of elements, and the sequence $\{\mathbf{v}^k\}_1^\infty$ must converge to the global minimizer. \square

4 Applications

In this section, we sketch two real-world applications of the multiplicative updates to problems in signal processing and pattern recognition. Recently, the updates have also been applied by astrophysicists to estimate the mass distribution of a gravitational lens and the positions of the sources from combined strong and weak lensing data (Diego et al., 2005 (submitted)).

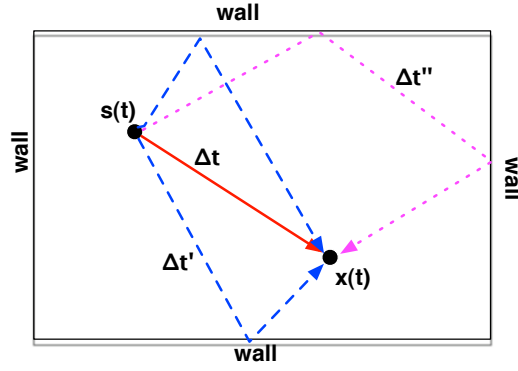


Figure 3: Estimating time delay from signals and their echos in reverberant environments.

4.1 Acoustic Time Delay Estimation

In a reverberant acoustic environment, microphone recordings capture echos reflected from objects such as walls and furniture in addition to the signals that propagate directly from a sound source. As illustrated in Figure 3, the signal at the microphone $x(t)$ can be modeled as a linear combination of the source signal $s(t)$ at different delay times: $s(t - \Delta t)$, $s(t - \Delta t')$ and $s(t - \Delta t'')$. Given a received signal $x(t)$ and its source signal $s(t)$, how can we identify in $x(t)$ all the time-delayed components of $s(t)$? To this end, we consider the model (Lin et al., 2004):

$$x(t) = \sum_i^N \alpha_i s(t - \Delta t_i) \quad \text{with} \quad \alpha_i \geq 0, \quad (32)$$

in which $\{\Delta t_i\}_{i=1}^N$ are all possible time delays (discretized to some finite resolution) and $\{\alpha_i\}$ are the relative amplitudes (or attenuations) of the time-delayed components. The nonnegativity constraints in (32) incorporate the assumption that only the amplitudes of acoustic waves are affected by reflections while the phases are retained (Allen and Berkley, 1979). Within this model, the time-delayed components of $s(t)$ can be identified by computing the amplitudes α_i that best reconstruct

$x(t)$. The reconstruction with least-squares-error is obtained from the nonnegative deconvolution:

$$\alpha^* = \arg \min_{\alpha_i \geq 0} \frac{1}{2} |x(t) - \sum_i^N \alpha_i s(t - \Delta t_i)|^2 . \quad (33)$$

Nonzero weights α_i^* in the least squares reconstruction are interpreted as indicating time-delayed components in the received signal with delays Δt_i .

It is convenient to rewrite this optimization in the frequency domain. Let $\tilde{x}(f)$ and $\tilde{s}(f)$ denote the Fourier transforms of $x(t)$ and $s(t)$, respectively. Also, define the positive semidefinite matrix K_{ij} and the real-valued coefficients c_i by:

$$K_{ij} = \sum_f |\tilde{s}(f)|^2 e^{j2\pi f(\Delta t_j - \Delta t_i)}, \quad (34)$$

$$c_i = \sum_f \tilde{s}^*(f) \tilde{x}(f) e^{j2\pi f \Delta t_i}, \quad (35)$$

where the sums are over positive and negative frequencies. In terms of the matrix K_{ij} and coefficients c_i , the optimization in eq. (33) can be rewritten as:

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \sum_{ij} \alpha_i K_{ij} \alpha_j - \sum_i c_i \alpha_i \\ & \text{subject to} \quad \alpha_i \geq 0 \end{aligned} \quad (36)$$

This has the same form as the NQP problem in eq. (5) and can be solved by the multiplicative updates in eq. (12). Note that K_{ij} defines a Toeplitz matrix if the possible time delays Δt_i are linearly spaced. Using fast Fourier transforms, the Toeplitz structure of K_{ij} can be exploited for much faster matrix-vector operations per multiplicative update.

Figure 4 shows the convergence of the multiplicative updates for a problem in acoustic time delay estimation. The source signal $s(t)$ in this example was a 30 ms window of speech, and the

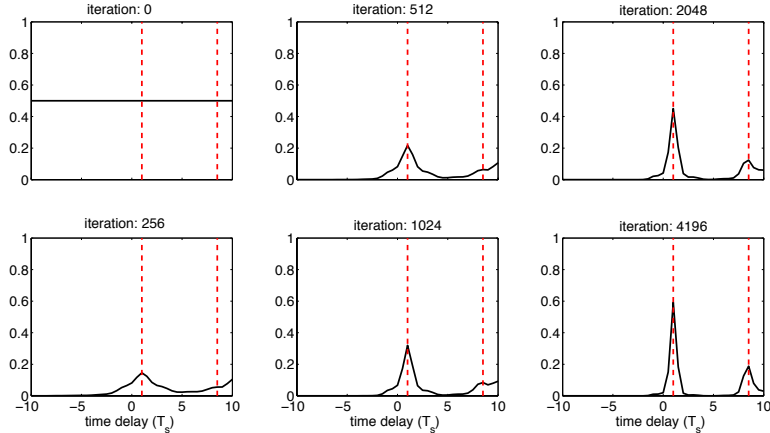


Figure 4: Convergence of the multiplicative updates for acoustic time delay estimation.

received signal $x(t)$ was given by:

$$x(t) = s(t - T_s) + 0.5 s(t - 8.5T_s),$$

where T_s was the sampling period. The vertical axes measure the estimated amplitudes α_i after different numbers of iterations of the multiplicative updates; the horizontal axes measure the time delays Δt_i in the units of T_s . The vertical dashed lines indicate the delays at T_s and $8.5T_s$. The figure shows that as the number of iterations is increased, the time delays are accurately predicted by the peaks in the estimated amplitudes α_i .

4.2 Large Margin Classification

Large margin classifiers have been applied successfully to many problems in machine learning and statistical pattern recognition (Cristianini and Shawe-Taylor, 2000; Vapnik, 1998). These classifiers use hyperplanes as decision boundaries to separate positively and negatively labelled examples represented by multidimensional vectors. Generally speaking, the hyperplanes are chosen to maximize the minimum distance (known as the margin) from any labeled example to the decision boundary; see figure 5.

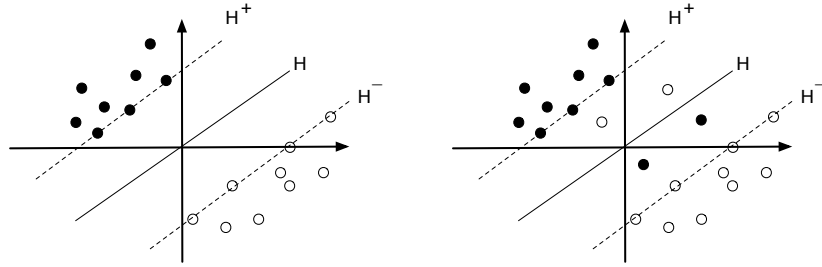


Figure 5: Large margin classifiers. Positively and negatively labeled examples are indicated by filled and hollowed circles, respectively. (a) A linearly separable data set. The support vectors for the maximum margin hyperplane H lie on two hyperplanes H^+ and H^- parallel to the decision boundary. Large margin classifiers maximize the distance between these hyperplanes. (b) A linearly inseparable data set. The support vectors in this case also include examples lying between H^+ and H^- that cannot be classified correctly.

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denote a data set of N labeled “training” examples with binary class labels $y_i = \pm 1$. The simplest case, shown in figure 5a, is that the two classes are linearly separable by a hyperplane that passes through the origin. Let \mathbf{w}^* denote the hyperplane’s normal vector; the classification rule, given by $y = \text{sgn}(\mathbf{w}^{*\text{T}}\mathbf{x})$, labels examples based on whether they lie above or below the hyperplane. The maximum margin hyperplane is computed by solving the constrained optimization:

$$\begin{aligned} & \text{minimize} && \frac{1}{2}\mathbf{w}^{\text{T}}\mathbf{w} \\ & \text{subject to} && y_i\mathbf{w}^{\text{T}}\mathbf{x}_i \geq 1, \quad i = 1, 2, \dots, N \end{aligned} \quad (37)$$

The constraints in this optimization ensure that all the training examples are correctly labelled by the classifier’s decision rule. While a potentially infinite number of hyperplanes satisfy these constraints, the classifier with minimal $\|\mathbf{w}\|$ (and thus maximal margin) has provably small error rates (Vapnik, 1998) on unseen examples. The optimization problem in eq. (37) is a convex quadratic programming problem in the vector \mathbf{w} . Its dual formulation is:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\text{T}} \mathbf{x}_j - \sum_i \alpha_i \\ & \text{subject to} && \alpha_i \geq 0, \quad i = 1, 2, \dots, N \end{aligned} \quad (38)$$

Let \mathbf{A} denote the positive semidefinite matrix with elements $y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$, and let \mathbf{e} denote the column vector of all ones. The objective function for the dual in eq. (38) can then be written as the NQP problem:

$$L(\alpha) = \frac{1}{2} \alpha^\top \mathbf{A} \alpha - \mathbf{e}^\top \alpha, \quad (39)$$

and the multiplicative updates in eq. (12) can be used to find the optimal α^* . Labeled examples that correspond to active constraints in eq. (37) are called support vectors. The normal vector \mathbf{w}^* is completely determined by support vectors since the solution to the primal problem, eq. (37), is given by:

$$\mathbf{w}^* = \sum_i y_i \alpha_i^* \mathbf{x}_i. \quad (40)$$

For non-support vectors, the inequalities are strictly satisfied in eq. (37), and their corresponding Lagrange multipliers vanish (that is, $\alpha_i^* = 0$).

Figure 6 illustrates the convergence of the multiplicative updates for large margin classification of handwritten digits (Schölkopf et al., 1997). The plots show the estimated support vector coefficients α_i after different numbers of iterations of the multiplicative updates. The horizontal axes in these plots index the coefficients α_i of the $N = 1389$ training examples, while the vertical axes show their values. For ease of visualization, the training examples were ordered so that support vectors appear to the left and non-support vectors, to the right. The coefficients were uniformly initialized as $\alpha_i = 1$. Note that the non-support vector coefficients are quickly attenuated to zero.

Multiplicative updates can also be used to train large margin classifiers when the labelled examples are not linearly separable, as shown in Figure 5b. In this case, the constraints in eq. (37) cannot all be simultaneously satisfied, and some of them must be relaxed. One simple relaxation is to permit some slack in the constraints, but to penalize the degree of slack as measured by the

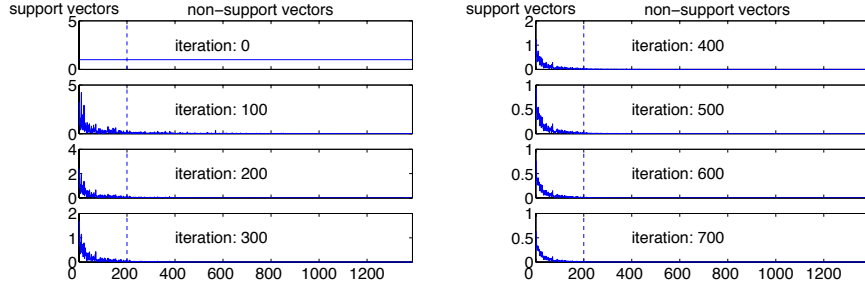


Figure 6: Convergence of the multiplicative updates in eq. (12) for a large margin classifier distinguishing handwritten digits (2s versus 3s). The coefficients corresponding to non-support vectors are quickly attenuated to zero.

ℓ_1 -norm:

$$\begin{aligned}
 & \text{minimize} && \mathbf{w}^T \mathbf{w} + C \sum_i \xi_i \\
 & \text{subject to} && y_i \mathbf{w}^T \mathbf{x}_i \geq 1 - \xi_i \\
 & && \xi_i \geq 0.
 \end{aligned} \tag{41}$$

The parameter C balances the slackness penalty versus the large margin criterion; the resulting classifiers are known as soft margin classifiers. The dual of this optimization is an NQP problem with the same quadratic form as the linearly separable case, but with box constraints:

$$\begin{aligned}
 & \text{minimize} && \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{A} \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\
 & \text{subject to} && 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N,
 \end{aligned} \tag{42}$$

The clipped multiplicative updates in eq. (14) can be used to perform this optimization for soft margin classifiers.

Another way of handling linear inseparability is to embed the data in a high dimensional non-linear feature space, then to construct the maximum margin hyperplane in feature space. The nonlinear mapping is performed implicitly by specifying a kernel function that computes the inner product in feature space. The optimization for the maximum margin hyperplane in feature space has the same form as eq. (38), except that the original Gram matrix with elements $\mathbf{x}_i^T \mathbf{x}_j$ is replaced by the kernel matrix of inner products in feature space (Cristianini and Shawe-Taylor, 2000; Vap-

nik, 1998). The use of multiplicative updates for large margin classification of linearly inseparable data sets is discussed further in (Sha et al., 2003b,a).

A large number of algorithms have been investigated for nonnegative quadratic programming in SVMs. Among them, there are many criteria that could be compared, such as speed of convergence, memory requirements, and ease of implementation. The main utility of the multiplicative updates appears to lie in their ease of implementation. The updates are very well suited for applications involving small to moderately sized data sets, where computation time is not a primary concern, and where the simple, parallel form of the updates makes them easy to implement in high-level languages such as MATLAB.

The most popular methods for training SVMs—so-called “subset” methods—take a fundamentally different approach to NQP. In contrast to the parallel form of the multiplicative updates, subset methods split the variables at each iteration into two sets: a *fixed* set in which the variables are held constant, and a *working* set in which the variables are optimized by an internal subroutine. At the end of each iteration, a heuristic is used to transfer variables between the two sets and improve the objective function.

Two subset methods have been widely used for training SVMs. The first is the method of sequential minimal optimization (SMO) (Platt, 1999), which updates only two coefficients of the weight vector per iteration. In this case, there exists an analytical solution for the updates, so that one avoids the expense of an iterative optimization within each iteration of the main loop. SMO enforces the sum and box constraints for soft margin classifiers. If the sum constraint is lifted, then it is possible to update the coefficients of the weight vector sequentially, one at a time, with an adaptive learning rate that ensures monotonic convergence. This *coordinate descent* approach is also known as the Kernel Adatron (Friess et al., 1998). SMO and Kernel Adatron are among the most viable methods for training SVMs on large data sets, and experiments have shown that they converge much faster than the multiplicative updates (Sha et al., 2003b). Nevertheless, for simplicity and ease of implementation, we believe that the multiplicative updates provide an

attractive starting point for experimenting with large margin classifiers.

5 Summary and Discussion

In this paper, we have described multiplicative updates for solving convex problems in NQP. The updates are distinguished by their simplicity in both form and computation. We showed that the updates lead to monotonic improvement in the objective function for NQP and converge to global minima. The updates can be viewed as generalizations of the iterative rules previously developed for nonnegative matrix factorization (Lee and Seung, 2001). They have a strikingly different form than other additive and multiplicative updates used in statistical learning.

We can also compare the multiplicative updates to interior point methods (Wright, 1997) that have been studied for NQP. These methods start from an interior point of the feasible region, then iteratively update the current estimate of the minimizer along particular search directions. There are many ways to determine the search directions—for example, using Newton’s method to solve the equations characterizing primal and dual optimality, or approximating the original optimization problem by a simpler subproblem inside the trust region of the current estimate. The resulting updates take an elementwise additive form, stepping along a particular search direction in the nonnegative orthant. The step size is chosen to ensure that the search remains in the feasible region while making progress toward the minimizer. If the search direction corresponds to the negative gradient of the objective function $F(\mathbf{v})$, then the updates reduce to steepest descent. Most of the computational effort in interior point methods is devoted to deriving search directions and maintaining the feasibility of the updated estimates.

The multiplicative updates are similar to trust region methods in spirit. Instead of constructing an ellipsoidal trust region centered at the current estimate of the minimizer, however, we have derived the updates from a nonlinear yet analytically tractable auxiliary function. Optimizing the auxiliary function guarantees the improvement of the objective function in the nonnegative orthant,

which can be viewed as the trust region for each update. The search direction derived from the auxiliary function is very simple to compute, as opposed to that of many interior-point methods. It remains an open question to quantify more precisely the rate of convergence of the multiplicative updates.

Though not as well theoretically characterized as traditional methods for NQP, the multiplicative updates have nevertheless proven extremely useful in practice. In this paper, we have described our own use of the updates for acoustic echo cancellation and large margin classification. In the meanwhile, others have applied the updates to NQP problems that arise in the analysis of astrophysical data (Diego et al., 2005 (submitted)). We are hopeful that more applications will continue to emerge in other areas of neural computation and statistical learning.

Acknowledgements

This work was supported by NSF Award 0238323.

References

- Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65:943–950, 1979.
- E. Bauer, D. Koller, and Y. Singer. Update rules for parameter estimation in Bayesian networks. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in AI*, pages 3–13, Providence, RI, 1997. Morgan Kaufmann.
- D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, 1999.
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

- J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480, 1972.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–37, 1977.
- J. M. Diego, M. Tegmark, P. Protopapas, and H. B. Sandvik. Combined reconstruction of weak and strong lensing data with WSLAP, 2005 (submitted). Available at <http://www.arxiv.org/abs/astro-ph/0509103>.
- T. Friess, N. Cristianini, and C. Campbell. The Kernel Adatron algorithm: a fast and simple learning procedure for support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 188–196, 1998.
- J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.
- D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562, Cambridge, MA, 2001. MIT Press.
- D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.
- Y. Lin, D. D. Lee, and L. K. Saul. Nonnegative deconvolution for time of arrival estimation. In *Proceedings of the International Conference of Speech, Acoustics, and Signal Processing (ICASSP-2004)*, volume 2, pages 377–380, Montreal, Canada, 2004.
- J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185–208, Cambridge, MA, 1999. MIT Press.

- L. K. Saul, F. Sha, and D. D. Lee. Statistical signal processing with nonnegativity constraints. In *Proceedings of the Eighth European Conference on Speech Communication and Technology*, volume 2, pages 1001–1004, Geneva, Switzerland, 2003.
- B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, 1997.
- T. Serafini, G. Zanghirati, and L. Zanni. Gradient projection methods for quadratic programs and applications in training support vector machines. *Optimization Methods and Software*, 20(20):353–378, 2005.
- F. Sha, L. K. Saul, and D. D. Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural and Information Processing Systems*, volume 15, pages 897–904, Cambridge, MA, 2003a. MIT Press.
- F. Sha, L. K. Saul, and D. D. Lee. Multiplicative updates for large margin classifiers. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory (COLT-03)*, pages 188–202, Washington D.C., 2003b.
- V. Vapnik. *Statistical Learning Theory*. Wiley, N.Y., 1998.
- S. J. Wright. *Primal-dual interior Point Methods*. SIAM, Philadelphia, PA, 1997.
- W. J. Zangwill. *Nonlinear Programming: a unified approach*. Prentice-Hall, Englewood Cliffs, N.J., 1969.

A Zangwill's Convergence Theorem

Zangwill's Convergence Theorem enumerates the conditions for global convergence of general iterative procedures. The theorem is presented in its most general form in (Zangwill, 1969); for our purposes here, we specifically state the theorem in the context of optimization. Let $F(\mathbf{v})$ denote the objective function to be minimized by an iterative update rule $\mathbf{v}^{k+1} = \mathcal{M}(\mathbf{v}^k)$, where the domain and range of the mapping $\mathcal{M} : \mathcal{V} \rightarrow \mathcal{V}$ lie in the feasible set. Suppose that we apply the mapping \mathcal{M} to generate a sequence of parameters $\{\mathbf{v}^k\}_{k=1}^{\infty}$. Our goal is to examine whether this sequence converges to a point in a "desired" set $\mathcal{V}^* \subset \mathcal{V}$. Assume that the objective function $F(\mathbf{v})$ and the mapping \mathcal{M} are continuous, and that the following conditions are met:

- (i) All points \mathbf{v}^k are in a compact set that is a subset of \mathcal{V} .
- (ii) If $\mathbf{v} \notin \mathcal{V}^*$, then the update leads to a strict reduction in the objective function. That is,
$$F(\mathcal{M}(\mathbf{v})) < F(\mathbf{v}).$$
- (iii) If $\mathbf{v} \in \mathcal{V}^*$, then either $\mathcal{M}(\mathbf{v}) = \mathbf{v}$ or $F(\mathcal{M}(\mathbf{v})) \leq F(\mathbf{v})$.

Zangwill's Convergence Theorem states that under these conditions, either the sequence $\{\mathbf{v}^k\}_{k=1}^{\infty}$ stops at a point in the set \mathcal{V}^* , or all accumulation points of the sequence are in the set \mathcal{V}^* .

The theorem can be used to analyze the convergence of iterative update rules by verifying that these three conditions hold for particular "desired" sets. For the multiplicative updates in eq. (12), the theorem implies convergence to a fixed point $\mathbf{v} = mm(\mathbf{v})$. It does not, however, imply, convergence to the unique fixed point that is the global minimizer of the objective function. In particular, if we constrain \mathcal{V}^* to contain only the global minimizer, then condition (ii), which stipulates that $F(\mathbf{v})$ *strictly* decreases under the mapping \mathcal{M} for all $\mathbf{v} \notin \mathcal{V}^*$, is clearly violated due to the existence of "spurious" fixed points. The proof of global convergence thus requires the extra machinery of section 3.2.