Supplementary material for: Distributed Optimization of Deeply Nested Systems

Miguel Á. Carreira-Perpiñán and Weiran Wang
Electrical Engineering and Computer Science, University of California, Merced
{mcarreira-perpinan, wwang5}@ucmerced.edu

February 22, 2014

Abstract

We give theorem statements and proofs for some of the claims in the paper. See also Carreira-Perpiñán and Wang (2012).

1 Definitions

Consider a regression problem of mapping inputs $\mathbf{x}$ to outputs $\mathbf{y}$ (both high-dimensional) with a deep net $\mathbf{f}(\mathbf{x})$ given a dataset of N pairs $(\mathbf{x}_n, \mathbf{y}_n)$. We define the *nested objective function* to learn a deep net with K hidden layers, like that in fig. 1 of the paper, as (to simplify notation, we ignore bias parameters and assume each hidden layer has H units):

$$E_1(\mathbf{W}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n;\mathbf{W})\|^2 \qquad \mathbf{f}(\mathbf{x};\mathbf{W}) = \mathbf{f}_{K+1}(\dots\mathbf{f}_2(\mathbf{f}_1(\mathbf{x};\mathbf{W}_1);\mathbf{W}_2)\dots;\mathbf{W}_{K+1}) \tag{1}$$

where each layer function has the form $\mathbf{f}_k(\mathbf{x};\mathbf{W}_k) = \sigma(\mathbf{W}_k\mathbf{x})$, i.e., a linear mapping followed by a squashing nonlinearity ($\sigma(t)$ applies a scalar function, such as the sigmoid $1/(1+e^{-t})$, elementwise to a vector argument, with output in $[0,1]$).

In the *method of auxiliary coordinates (MAC)*, we introduce one auxiliary variable per data point and per hidden unit (so $\mathbf{Z} = (\mathbf{Z}_1,\dots,\mathbf{Z}_N)$, with $\mathbf{z}_n = (\mathbf{z}_{1,n},\dots,\mathbf{z}_{K,n})$) and define the following equality-constrained optimization problem:

$$E(\mathbf{W},\mathbf{Z}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_{K+1}(\mathbf{z}_{K,n};\mathbf{W}_{K+1})\|^2 \text{ s.t. } \left.\begin{cases} \mathbf{z}_{K,n} = \mathbf{f}_K(\mathbf{z}_{K-1,n};\mathbf{W}_K) \\ \dots \\ \mathbf{z}_{1,n} = \mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1) \end{cases}\right\} n = 1,\dots,N. \tag{2}$$

Sometimes, for notational convenience (in particular in theorem 3.3), we will write the constraints for the nth point as a single vector constraint $\mathbf{z}_n - \mathbf{F}(\mathbf{z}_n,\mathbf{W};\mathbf{x}_n) = \mathbf{0}$ (with an obvious definition for $\mathbf{F}$). We will also call Ω the feasible set of the MAC-constrained problem, i.e.,

$$\Omega = \{(\mathbf{W},\mathbf{Z})\colon\ \mathbf{z}_n = \mathbf{F}(\mathbf{z}_n,\mathbf{W};\mathbf{x}_n),\ n = 1,\dots,N\}. \tag{3}$$

To solve the constrained problem (2) using the quadratic-penalty (QP) method (Nocedal and Wright, 2006), we optimize the following function over $(\mathbf{W},\mathbf{Z})$ for fixed $\mu > 0$ and drive $\mu \to \infty$:

$$E_Q(\mathbf{W},\mathbf{Z};\mu) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_{K+1}(\mathbf{z}_{K,n};\mathbf{W}_{K+1})\|^2 + \frac{\mu}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{z}_{k,n} - \mathbf{f}_k(\mathbf{z}_{k-1,n};\mathbf{W}_k)\|^2. \tag{4}$$

2 Equivalence of the MAC and nested formulations

First, we give a theorem that holds under very general assumptions. In particular, it does not require the functions to be smooth, it holds for any loss function beyond the least-squares one, and it holds if the nested problem is itself subject to constraints.

Theorem 2.1. *The nested problem* (1) *and the MAC-constrained problem* (2) *are equivalent in the sense that their minimizers are in a one-to-one correspondence.*

Proof. Let us prove that any minimizer of the nested problem is associated with a unique minimizer of the MAC-constrained problem ($\Rightarrow$), and vice versa ($\Leftarrow$). Recall the following definitions (Nocedal and Wright, 2006): (i) For an unconstrained minimization problem $\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x})$, $\mathbf{x}^* \in \mathbb{R}^n$ is a local minimizer if there exists a nonempty neighborhood $\mathcal{N} \subset \mathbb{R}^n$ of $\mathbf{x}^*$ such that $F(\mathbf{x}^*) \leq F(\mathbf{x})\ \forall \mathbf{x} \in \mathcal{N}$. (ii) For a constrained minimization problem $\min F(\mathbf{x})$ s.t. $\mathbf{x} \in \Omega \subset \mathbb{R}^n$, $\mathbf{x}^* \in \mathbb{R}^n$ is a local minimizer if $\mathbf{x}^* \in \Omega$ and there exists a nonempty neighborhood $\mathcal{N} \subset \mathbb{R}^n$ of $\mathbf{x}^*$ such that $F(\mathbf{x}^*) \leq F(\mathbf{x})\ \forall \mathbf{x} \in \mathcal{N} \cap \Omega$.

Define the "forward-propagation" function $\mathbf{g}(\mathbf{W})$ as the result of mapping $\mathbf{z}_{1,n} = \mathbf{f}_1(\mathbf{x}_n; \mathbf{W}_1), \dots, \mathbf{z}_{K,n} = \mathbf{f}_K(\mathbf{z}_{K-1,n}; \mathbf{W}_K)$ for $n = 1, \dots, N$. This maps each $\mathbf{W}$ to a unique $\mathbf{Z}$, and satisfies $\mathbf{f}_{K+1}(\mathbf{z}_{K,n}; \mathbf{W}_{K+1}) = \mathbf{f}_{K+1}(\dots \mathbf{f}_2(\mathbf{f}_1(\mathbf{x}_n; \mathbf{W}_1); \mathbf{W}_2) \dots ; \mathbf{W}_{K+1}) = \mathbf{f}(\mathbf{x}_n; \mathbf{W})$ for $n = 1, \dots, N$, and therefore that $E_1(\mathbf{W}) = E(\mathbf{W}, \mathbf{g}(\mathbf{W}))$ for any $\mathbf{W}$.

($\Rightarrow$) Let $\mathbf{W}^*$ be a local minimizer of the nested problem (1). Then, there exists a nonempty neighborhood $\mathcal{N}$ of $\mathbf{W}^*$ such that $E_1(\mathbf{W}^*) \leq E_1(\mathbf{W})\ \forall \mathbf{W} \in \mathcal{N}$. Let $\mathbf{Z}^* = \mathbf{g}(\mathbf{W}^*)$ and call $\mathcal{M} = \{(\mathbf{W}, \mathbf{Z})\colon\ \mathbf{W} \in \mathcal{N} \text{ and } \mathbf{Z} = \mathbf{g}(\mathbf{W})\}$, which is a nonempty neighborhood of $(\mathbf{W}^*, \mathbf{Z}^*)$ in $(\mathbf{W}, \mathbf{Z})$-space. Now, for any $(\mathbf{W}, \mathbf{Z}) \in \mathcal{M} \cap \mathcal{N}$ we have that $E(\mathbf{W}, \mathbf{Z}) = E(\mathbf{W}, \mathbf{g}(\mathbf{W})) = E_1(\mathbf{W}) \geq E_1(\mathbf{W}^*) = E(\mathbf{W}^*, \mathbf{g}(\mathbf{W}^*)) = E(\mathbf{W}^*, \mathbf{Z}^*)$. Hence $(\mathbf{W}^*, \mathbf{Z}^*)$ is a local minimizer of the MAC-constrained problem.

($\Leftarrow$) Let $(\mathbf{W}^*, \mathbf{Z}^*)$ be a local minimizer of the MAC-constrained problem (2). Then, there exists a nonempty neighborhood $\mathcal{M}$ of $(\mathbf{W}^*, \mathbf{Z}^*)$ such that $E(\mathbf{W}^*, \mathbf{Z}^*) \leq E(\mathbf{W}, \mathbf{Z})\ \forall (\mathbf{W}, \mathbf{Z}) \in \mathcal{M} \cap \Omega$. Note that $(\mathbf{W}, \mathbf{Z}) \in \mathcal{M} \cap \Omega \Rightarrow \mathbf{Z} = \mathbf{g}(\mathbf{W}) \Rightarrow E(\mathbf{W}, \mathbf{Z}) = E_1(\mathbf{W})$, and this applies in particular to $(\mathbf{W}^*, \mathbf{Z}^*)$ (which, being a solution, is feasible and thus belongs to $\mathcal{M} \cap \Omega$). Calling $\mathcal{N} = \{\mathbf{W}\colon\ (\mathbf{W}, \mathbf{Z}) \in \mathcal{M} \cap \Omega\}$, we have that $\forall \mathbf{W} \in \mathcal{N}\colon\ E_1(\mathbf{W}) = E(\mathbf{W}, \mathbf{g}(\mathbf{W})) = E(\mathbf{W}, \mathbf{Z}) \geq E(\mathbf{W}^*, \mathbf{Z}^*) = E(\mathbf{W}^*, \mathbf{g}(\mathbf{W}^*)) = E_1(\mathbf{W}^*)$. Hence $\mathbf{W}^*$ is a local minimizer of the nested problem.

Finally, one can see that the proof holds if the nested problem uses a loss function that is not the least-squares one, and if the nested problem is itself subject to constraints. □

Obviously, the theorem holds if we exchange $\geq$ with $>$ everywhere (thus exchanging non-strict with strict minimizers), and if we exchange "min" with "max" (hence the maximizers of the MAC and nested formulations are in a one-to-one correspondence as well). Figure 1 illustrates the theorem. Essentially, the nested objective function $E_1(\mathbf{W})$ stretches along the manifold defined by $(\mathbf{W}, \mathbf{Z} = \mathbf{g}(\mathbf{W}))$ preserving the minimizers and maximizers. The projection on $\mathbf{W}$-space of the part of $E(\mathbf{W}, \mathbf{Z})$ that sits on top of that manifold recovers $E_1(\mathbf{W})$.

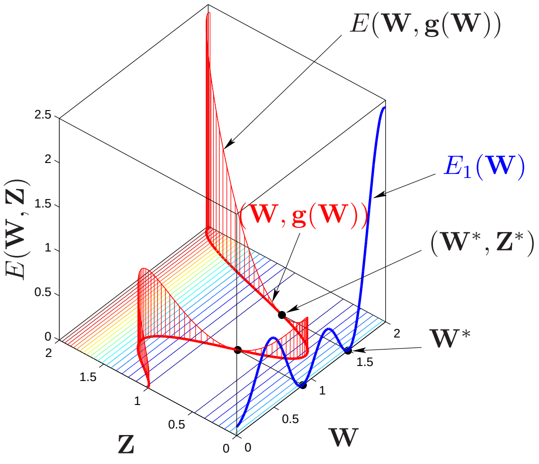

Figure 1: Illustration of the equivalence between the nested and MAC-constrained problems (see the proof of theorem 2.1). The MAC objective function $E(\mathbf{W}, \mathbf{Z})$ is shown with contour lines in the $(\mathbf{W}, \mathbf{Z})$-space, and with the vertical red lines on the feasible set $(\mathbf{W}, \mathbf{g}(\mathbf{W}))$. The nested objective function $E_1(\mathbf{W})$ is shown in blue. Corresponding minima for both problems, $\mathbf{W}^*$ and $(\mathbf{W}^*, \mathbf{Z}^*)$, are indicated.

2.1 KKT conditions

We now show that the first-order necessary (Karush-Kuhn-Tucker, KKT) conditions of both problems (nested and MAC-constrained) have the same stationary points. For simplicity and clarity of exposition, we give a proof for the special case of $K = 1$. The proof for $K > 1$ layers follows analogously. We assume the functions $\mathbf{f}_1$ and $\mathbf{f}_2$ have continuous first derivatives w.r.t. both its input and its weights. $\mathbf{J}_{\mathbf{f}_2}(\cdot;\mathbf{W}_2)$ indicates the Jacobian of $\mathbf{f}_2$ w.r.t. its input. To simplify notation, we sometimes omit the dependence on the weights; for example, we write $\mathbf{f}_2(\mathbf{f}_1(\mathbf{x};\mathbf{W}_1);\mathbf{W}_2)$ as $\mathbf{f}_2(\mathbf{f}_1(\mathbf{x}))$, and $\mathbf{J}_{\mathbf{f}_2}(\cdot;\mathbf{W}_2)$ as $\mathbf{J}_{\mathbf{f}_2}(\cdot)$.

Theorem 2.2. *The KKT conditions for the nested problem* (1) *and the MAC-constrained problem* (2) *are equivalent.*

Proof. The nested problem for a nested function $\mathbf{f}_2(\mathbf{f}_1(\mathbf{x}))$ is:

$$\min_{\mathbf{W}_1,\mathbf{W}_2} E_1(\mathbf{W}_1,\mathbf{W}_2) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_2(\mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1);\mathbf{W}_2)\|^2.$$

Then we have the stationary point equation (first-order necessary conditions for a minimizer):

$$\frac{\partial E_1}{\partial \mathbf{W}_1} = -\sum_{n=1}^{N} \frac{\partial \mathbf{f}_1^T}{\partial \mathbf{W}_1}(\mathbf{x}_n)\, \mathbf{J}_{\mathbf{f}_2}(\mathbf{f}_1(\mathbf{x}_n))^T (\mathbf{y}_n - \mathbf{f}_2(\mathbf{f}_1(\mathbf{x}_n))) = \mathbf{0} \tag{5}$$

$$\frac{\partial E_1}{\partial \mathbf{W}_2} = -\sum_{n=1}^{N} \frac{\partial \mathbf{f}_2^T}{\partial \mathbf{W}_2}(\mathbf{f}_1(\mathbf{x}_n))\, (\mathbf{y}_n - \mathbf{f}_2(\mathbf{f}_1(\mathbf{x}_n))) = \mathbf{0} \tag{6}$$

which is satisfied by all the minima, maxima and saddle points.

The MAC-constrained problem is

$$\min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{Z}} E(\mathbf{W}_1,\mathbf{W}_2,\mathbf{Z}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_2(\mathbf{z}_n;\mathbf{W}_2)\|^2 \text{ s.t. } \mathbf{z}_n = \mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1),\ n = 1,\dots,N,$$

with Lagrangian

$$\mathcal{L}(\mathbf{W}_1,\mathbf{W}_2,\mathbf{Z},\boldsymbol{\lambda}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}_2(\mathbf{z}_n;\mathbf{W}_2)\|^2 - \sum_{n=1}^{N} \boldsymbol{\lambda}_n^T(\mathbf{z}_n - \mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1))$$

and KKT conditions

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{W}_1} = \sum_{n=1}^{N} \frac{\partial \mathbf{f}_1^T}{\partial \mathbf{W}_1}(\mathbf{x}_n)\, \boldsymbol{\lambda}_n = \mathbf{0} \tag{7}$$

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{W}_2} = -\sum_{n=1}^{N} \frac{\partial \mathbf{f}_2^T}{\partial \mathbf{W}_2}(\mathbf{f}_1(\mathbf{x}_n))\, (\mathbf{y}_n - \mathbf{f}_2(\mathbf{z}_n)) = \mathbf{0} \tag{8}$$

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{z}_n} = -\mathbf{J}_{\mathbf{f}_2}(\mathbf{z}_n)^T(\mathbf{y}_n - \mathbf{f}_2(\mathbf{z}_n)) - \boldsymbol{\lambda}_n = \mathbf{0},\ n = 1,\dots,N \tag{9}$$

$$\mathbf{z}_n = \mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1),\ n = 1,\dots,N. \tag{10}$$

Substituting $\boldsymbol{\lambda}_n$ from eq. (9) and $\mathbf{z}_n$ from eq. (10):

$$\boldsymbol{\lambda}_n = -\mathbf{J}_{\mathbf{f}_2}(\mathbf{z}_n)^T(\mathbf{y}_n - \mathbf{f}_2(\mathbf{z}_n)),\ n = 1,\dots,N \tag{9'}$$

$$\mathbf{z}_n = \mathbf{f}_1(\mathbf{x}_n;\mathbf{W}_1),\ n = 1,\dots,N \tag{10'}$$

into eqs. (7)–(8) we recover eqs. (5)–(6), thus a KKT point of the constrained problem is a stationary point of the nested problem. Conversely, given a stationary point $(\mathbf{W}_1,\mathbf{W}_2)$ of the nested problem, and defining $\boldsymbol{\lambda}_n$ and $\mathbf{z}_n$ as in eqs. (9')–(10'), then $(\mathbf{W}_1,\mathbf{W}_2,\mathbf{Z},\boldsymbol{\lambda})$ satisfies eqs. (7)–(10) and so is a KKT point of the constrained problem. Hence, there is a one-to-one correspondence between the stationary points of the nested problem and the KKT points of the MAC-constrained problem. □

From theorems 2.1 and 2.2, it follows that the minimizers, maximizers and saddle points of the nested problem are in one-to-one correspondence with the respective minimizers, maximizers and saddle points of the MAC-constrained problem.

3 Convergence of the quadratic-penalty method for MAC

Let us first give convergence conditions for the general equality-constrained minimization problem:

$$\min f(\mathbf{x}) \text{ s.t. } c_i(\mathbf{x}) = 0,\ i = 1,\dots,m \tag{11}$$

and the quadratic-penalty (QP) function

$$Q(\mathbf{x};\mu) = f(\mathbf{x}) + \frac{\mu}{2}\sum_{i=1}^{m} c_i^2(\mathbf{x}) \tag{12}$$

with penalty parameter $\mu > 0$. Given a positive increasing sequence $(\mu_k) \rightarrow \infty$, a nonnegative sequence $(\tau_k) \rightarrow 0$, and a starting point $\mathbf{x}_0$, the QP method finds an approximate minimizer $\mathbf{x}_k$ of $Q(\mathbf{x};\mu_k)$ for $k = 1,2,\dots$, so that the iterate $\mathbf{x}_k$ satisfies $\|\nabla_{\mathbf{x}} Q(\mathbf{x}_k;\mu_k)\| \leq \tau_k$. Given this algorithm, we have the following theorems:

Theorem 3.1 (Nocedal and Wright, 2006, Th. 17.1). *Suppose that $(\mu_k) \rightarrow \infty$ and $(\tau_k) \rightarrow 0$. If each $\mathbf{x}_k$ is the exact global minimizer of $Q(\mathbf{x};\mu_k)$, then every limit point $\mathbf{x}^*$ of the sequence $(\mathbf{x}_k)$ is a global solution of the problem* (11).

Theorem 3.2 (Nocedal and Wright, 2006, Th. 17.2). *Suppose that $(\mu_k) \rightarrow \infty$ and $(\tau_k) \rightarrow 0$, and that $\mathbf{x}^*$ is a limit point of $(\mathbf{x}_k)$. Then $\mathbf{x}^*$ is a stationary point of the function $\sum_{i=1}^m c_i^2(\mathbf{x})$. Besides, if the constraint gradients $\nabla c_i(\mathbf{x}^*),\ i = 1,\dots,m$ are linearly independent, then $\mathbf{x}^*$ is a KKT point for the problem* (11). *For such points, we have for any infinite subsequence $\mathcal{K}$ such that $\lim_{k\in\mathcal{K}} \mathbf{x}_k = \mathbf{x}^*$ that $\lim_{k\in\mathcal{K}} -\mu_k c_i(\mathbf{x}_k) = \lambda_i^*,\ i = 1,\dots,m$, where $\boldsymbol{\lambda}^*$ is the multiplier vector that satisfies the KKT conditions for the problem* (11).

If now we particularize these general theorems to our case, we can obtain stronger theorems. Theorem 3.1 is generally not applicable, because optimization problems involving nested functions are typically not convex and have local minima. Theorem 3.2 is applicable to prove convergence in the nonconvex case. We assume the functions $\mathbf{f}_1,\dots,\mathbf{f}_{K+1}$ in eq. (1) have continuous first derivatives w.r.t. both its input and its weights, so $E(\mathbf{W},\mathbf{Z})$ is differentiable w.r.t. $\mathbf{W}$ and $\mathbf{Z}$.

Theorem 3.3 (Convergence of MAC/QP for nested problems). *Consider the constrained problem* (2) *and its quadratic-penalty function $E_Q(\mathbf{W},\mathbf{Z};\mu)$ of* (4). *Given a positive increasing sequence $(\mu_k) \rightarrow \infty$, a nonnegative sequence $(\tau_k) \rightarrow 0$, and a starting point $(\mathbf{W}^0,\mathbf{Z}^0)$, suppose the QP method finds an approximate minimizer $(\mathbf{W}^k,\mathbf{Z}^k)$ of $E_Q(\mathbf{W}^k,\mathbf{Z}^k;\mu_k)$ that satisfies $\left\|\nabla_{\mathbf{W},\mathbf{Z}} E_Q(\mathbf{W}^k,\mathbf{Z}^k;\mu_k)\right\| \leq \tau_k$ for $k = 1,2,\dots$ Then, $\lim_{k\rightarrow\infty}(\mathbf{W}^k,\mathbf{Z}^k) = (\mathbf{W}^*,\mathbf{Z}^*)$, which is a KKT point for the problem* (2), *and its Lagrange multiplier vector has elements $\boldsymbol{\lambda}_n^* = \lim_{k\rightarrow\infty} -\mu_k\left(\mathbf{Z}_n^k - \mathbf{F}(\mathbf{Z}_n^k,\mathbf{W}^k;\mathbf{x}_n)\right),\ n = 1,\dots,N$.*

Proof. It follows by applying theorem 3.2 to the constrained problem (2) and by noting that $\lim_{k\rightarrow\infty}(\mathbf{W}^k,\mathbf{Z}^k) = (\mathbf{W}^*,\mathbf{Z}^*)$ exists and that the constraint gradients are linearly independent. We prove these two statements in turn.

The limit of the sequence $((\mathbf{W}^k,\mathbf{Z}^k))$ exists because the objective function $E(\mathbf{W},\mathbf{Z})$ of the MAC-constrained problem (hence the QP function $E_Q(\mathbf{W},\mathbf{Z};\mu)$) are lower bounded and have continuous derivatives.

The constraint gradients are l.i. at any point $(\mathbf{W},\mathbf{Z})$ and thus, in particular, at the limit $(\mathbf{W}^*,\mathbf{Z}^*)$. To see this, let us first compute the constraint gradients. There is one constraint $C_{nkh}(\mathbf{W},\mathbf{Z}) = z_{nkh} - f_{kh}(\mathbf{z}_{n,k-1};\mathbf{W}_k) = 0$ for each point $n = 1,\dots,N$, layer $k = 1,\dots,K$ and unit $h \in \mathcal{I}(k)$, where we define $\mathcal{I}(k)$ as the set of auxiliary coordinate indices for layer k and $\mathbf{z}_{n0} = \mathbf{x}_n,\ n = 1,\dots,N$. The gradient of this constraint is:

$$\frac{\partial C_{nkh}}{\partial \mathbf{W}_{k'}} = -\delta_{kk'}\frac{\partial f_{kh}}{\partial \mathbf{W}_k},\ k = 1,\dots,K$$

$$\frac{\partial C_{nkh}}{\partial z_{n'k'h'}} = \delta_{nn'}\left(\delta_{kk'}\delta_{hh'} - \delta_{k-1,k'}\frac{\partial f_{kh}}{\partial z_{n,k-1,h}}\right),\ n = 1,\dots,N,\ k = 1,\dots,K,\ h \in \mathcal{I}(k).$$

Now, we will show that these gradients are l.i. at any point $(\mathbf{W},\mathbf{Z})$. It suffices to look at the gradients w.r.t. $\mathbf{Z}$. Call $\alpha_{nkh} = \partial f_{kh}/\partial z_{n,k-1,h}$ for short. Constructing a linear combination of them and setting it to zero:

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{h\in\mathcal{I}(k)} \lambda_{nkh}\frac{\partial C_{nkh}}{\partial \mathbf{Z}'} = \mathbf{0}.$$

This implies, for the gradient element corresponding to $z_{n'k'h'}$:

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{h\in\mathcal{I}(k)} \lambda_{nkh}\delta_{nn'}\left(\delta_{kk'}\delta_{hh'} - \delta_{k-1,k'}\alpha_{nkh}\right) = \lambda_{n'k'h'} - \sum_{h\in\mathcal{I}(k'+1)} \lambda_{n',k'+1,h}\alpha_{n',k'+1,h} = 0$$
$$\Longrightarrow \lambda_{n'k'h'} = \sum_{h\in\mathcal{I}(k'+1)} \lambda_{n',k'+1,h}\alpha_{n',k'+1,h}.$$

Applying this for $k' = K,\dots,1$:

- For $k' = K$: $\lambda_{n'Kh'} = 0,\ n' = 1,\dots,N,\ h' \in \mathcal{I}(K)$.
- For $k' = K-1$: $\lambda_{n',K-1,h'} = \sum_{h\in\mathcal{I}(K)} \lambda_{n',K,h}\alpha_{n',K,h} = 0,\ n' = 1,\dots,N,\ h' \in \mathcal{I}(K-1)$.
- ...
- For $k' = 1$: $\lambda_{n',1,h'} = \sum_{h\in\mathcal{I}(2)} \lambda_{n',2,h}\alpha_{n',2,h} = 0,\ n' = 1,\dots,N,\ h' \in \mathcal{I}(1)$.

Hence, all the coefficients λ_{nkh} are zero and the gradients are l.i. □

In practice, as with any continuous optimization problem, convergence may occur in pathological cases to a stationary point of the constrained problem rather than a minimizer.

In summary, MAC/QP defines a continuous path $(\mathbf{W}^*(\mu),\mathbf{Z}^*(\mu))$ which, under some mild assumptions (essentially, that we minimize $E_Q(\mathbf{W},\mathbf{Z};\mu)$ increasingly accurately as $\mu\to\infty$), converges to a stationary point (typically a minimizer) of the constrained problem (2), and thus to a minimizer of the nested problem (1).

References

M. Á. Carreira-Perpiñán and W. Wang. Distributed optimization of deeply nested systems. Unpublished manuscript, arXiv:1212.5921, Dec. 24 2012.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York, second edition, 2006.