

*Scuola CIRO 2002*, pp. 1-000  
A. Agnetis, G. Di Pillo, Editors

# Le Reti Neurali e le Altre Reti Approssimanti nei Problemi di Ottimizzazione Funzionale <sup>1</sup>

Marcello Sanguineti (marcello@dist.unige.it)

Riccardo Zoppoli (rzop@dist.unige.it)

*Dipartimento di Informatica, Sistemistica e Telematica (DIST)*  
*Università di Genova - Via Opera Pia 13, 16145 Genova*

## Sommario

Nei problemi di ottimizzazione funzionale è necessario minimizzare o massimizzare un funzionale rispetto a soluzioni ammissibili rappresentate da funzioni appartenenti a spazi di dimensione infinita. Tali problemi possono essere risolti analiticamente solo se sono soddisfatte ipotesi piuttosto restrittive. Se queste non sono verificate, si impongono procedure risolutive approssimate. La metodologia di ottimizzazione approssimata proposta consiste nel vincolare le funzioni ammissibili ad assumere la struttura di combinazioni lineari di funzioni di base contenenti parametri “liberi”. Sostituendo tali combinazioni lineari di funzioni di base parametrizzate nel funzionale, si ottiene una funzione dipendente da un numero finito di variabili reali e quindi un problema di programmazione non lineare. Chiameremo “reti approssimanti” le combinazioni lineari ottenute in corrispondenza di particolari scelte delle funzioni di base, che garantiscono proprietà di approssimazione particolarmente utili nella risoluzione dei problemi di ottimizzazione funzionale.

La tecnica che presenteremo generalizza il metodo classico di Ritz del calcolo delle variazioni, nel quale le funzioni di base non contengono parametri liberi.

---

<sup>1</sup>Questo lavoro è stato finanziato in parte dal contratto di ricerca CNR-Agenzia 2000 “Nuovi algoritmi e metodologie per la risoluzione approssimata di problemi di ottimizzazione funzionale in ambiente stocastico” e in parte dal contratto di ricerca MIUR “Nuove tecniche per l’identificazione e il controllo di sistemi industriali”. I principali concetti, esposti nella trattazione, sono stati elaborati congiuntamente con Thomas Parisini (Università di Trieste) a partire dalle prime fasi della ricerca. L’esempio riportato nella sezione 12 è stato sviluppato in collaborazione con Angela Di Febbraro (Politecnico di Torino) e Simona Sacone (Università di Genova).

Abbiamo denominato detta tecnica “Metodo di Ritz Estesio” o ERIM (“Extended Ritz Method”). La superiorità dell’ERIM rispetto al metodo di Ritz si manifesta con una riduzione della velocità di crescita del numero di parametri liberi, necessari per ottenere una data accuratezza di ottimizzazione, al crescere del numero di componenti del vettore argomento delle funzioni ammissibili. Più precisamente, il grave inconveniente di una crescita “veloce” (ad esempio esponenziale, con il conseguente insorgere del fenomeno della “maledizione della dimensionalità”) affligge spesso il metodo di Ritz, ma può essere attenuato dall’ERIM. Qualitativamente parlando, la presenza dei parametri liberi nelle funzioni di base aumenta fortemente la “flessibilità” di queste ultime, consentendo all’ERIM di sfruttare meglio le caratteristiche di regolarità degli insiemi di funzioni a cui si suppone che la soluzione ottima appartenga.

Se il problema di ottimizzazione funzionale è formulato in ambiente aleatorio, si propone inoltre una tecnica di approssimazione stocastica per la risoluzione del problema di programmazione non lineare al quale viene ricondotto.

## 1 Introduzione

Nel formulare un problema di ottimizzazione, si richiede di individuare un insieme  $S$  di uno spazio  $H$ , i cui elementi sono chiamati *soluzioni ammissibili* del problema (in pratica, le alternative a disposizione di un decisore), e di associare ad ogni soluzione ammissibile un costo mediante un funzionale  $F$ , che chiamiamo per l’appunto *funzionale di costo*. L’obiettivo è dunque la minimizzazione di  $F$  in  $S$  (per semplicità di esposizione, facciamo riferimento alla minimizzazione di funzionali di costo; la metodologia che presenteremo si applica, con modifiche banali, al caso in cui  $F$  rappresenta una figura di merito da massimizzare).

Per le finalità espositive che ci proponiamo, è utile distinguere i problemi di ottimizzazione in due classi. La prima classe è costituita da quelli le cui soluzioni ammissibili appartengono ad un insieme di  $\mathbb{R}^n$  e consistono quindi in un numero finito  $n$  di componenti di un vettore. I problemi di questa prima classe sono i *problemi di programmazione matematica*. Per essi il termine “funzionale di costo” è comunemente sostituito dal termine “funzione di costo”.

La seconda classe di problemi di ottimizzazione è invece costituita da quei problemi in cui le soluzioni ammissibili sono funzioni appartenenti a spazi di dimensione infinita. Chiamiamo i corrispondenti problemi di ottimizzazione “*problemi di ottimizzazione funzionale*” per analogia con il termine “analisi funzionale”, riservato a quel settore della matematica che studia gli spazi lineari a dimensione infinita. Molte delle difficoltà incontrate in questo contesto sono dovute al fatto che l’ottimizzazione va effettuata nell’ambito di spazi di funzioni: la dimensione infinita rende inutilizzabili o comunque inadeguati gli strumenti analitici tipici della programmazione matematica.

I problemi di ottimizzazione funzionale costituiscono un insieme più ampio di quanto talvolta si creda. Oltre ai problemi classici del calcolo delle variazioni, innumerevoli sono infatti le situazioni teoriche e applicative in cui si richiede di determinare una funzione, in qualche senso “ottima”, nell’ambito di più funzioni ammissibili. Tale funzione può esprimere la dipendenza della soluzione ottima di un problema di programmazione matematica da alcuni parametri del problema stesso soggetti a variazioni, può essere la caratteristica ingresso/uscita di un riconoscitore di immagini o di segnali di varia natura, può rappresentare il modello di un sistema per la rilevazione e la classificazione di guasti, può descrivere la corrispondenza dinamica tra i segnali di ingresso e di uscita di un regolatore o di un filtro che si voglia sintetizzare o di un impianto che si voglia identificare, e così via.

La risoluzione per via analitica di un problema di ottimizzazione funzionale è purtroppo possibile in ben pochi casi. Considerando ad esempio i problemi di controllo ottimo, è noto che tali problemi ammettono soluzioni analitiche di agevole determinazione se valgono le ipotesi di linearità nei vincoli differenziali e algebrici (rispettivamente le equazioni di stato e di misura), di quadraticità nel funzionale di costo e, se il problema è formulato in ambiente stocastico, di gaussianità nelle variabili aleatorie. Se dette ipotesi (chiamate ipotesi LQG) non sono verificate, i problemi di controllo ottimo non possono essere risolti per via analitica (salvo qualche caso particolare), ma soltanto mediante procedure numeriche.

Vi sono diverse modalità per risolvere in forma approssimata i problemi di ottimizzazione funzionale. Il metodo approssimato descritto nel seguito consiste nel vincolare le funzioni decisionali ad assumere una struttura prefissata, in cui viene inserito un numero finito (eventualmente molto grande) di parametri “liberi”. Sostituendo tali funzioni nel funzionale di costo e nei vincoli, si ottiene un problema di programmazione non lineare. La sua risoluzione è poi affidata a un adeguato algoritmo di discesa.

La scelta della struttura delle funzioni decisionali è praticamente illimitata. Funzioni a struttura assegnata particolarmente semplici sono costituite da combinazioni lineari di funzioni di base “fisse”: in tal caso i parametri liberi sono dati dai coefficienti della combinazione lineare. La scelta di detta struttura conduce al metodo classico di Ritz del calcolo delle variazioni. Il metodo risale al 1909 [40]. Non sembra tuttavia che il metodo di Ritz abbia conseguito, in quasi un secolo, importanti successi in problemi in cui le funzioni decisionali dipendono da un numero elevato di variabili (si vedano [12, 19, 42] e i riferimenti in essi contenuti; si veda anche, per un metodo strettamente connesso a quello di Ritz, [8]). Ciò sembra doversi imputare al fatto che il metodo di Ritz può essere soggetto al cosiddetto fenomeno della “maledizione della dimensionalità” [9], dove la “dimensionalità” è rappresentata, nel nostro caso, dal numero  $d$  di variabili delle funzioni ammissibili. Sia infatti  $\varepsilon$  l’errore massimo accettabile nel determinare un’approssimazione della soluzione ottima. Può allora accadere che, per un valore fissato di  $\varepsilon$ , il numero di funzioni di base delle combinazioni lineari utilizzate dal metodo di Ritz (e quindi il numero dei coefficienti di tali combinazioni)

cresca in modo inaccettabile con  $d$ , tipicamente con una velocità di ordine  $O(1/\varepsilon^d)$ .

Per contro, da oltre un decennio, studi sull'approssimazione di funzioni hanno dimostrato la possibilità di approssimare funzioni dotate di opportune caratteristiche di regolarità mediante combinazioni lineari di funzioni di base contenenti parametri liberi (al posto delle funzioni di base fisse), in modo da soddisfare la seguente proprietà fondamentale: fissato  $\varepsilon$ , il numero di parametri da ottimizzare cresce “moderatamente” con il numero di variabili, ad esempio polinomialmente o anche solo linearmente. Esempi di funzioni di base *parametrizzate* sono le funzioni sigmoidali, le “radial basis function”, le “spline” a nodi mobili, ecc..

Tali risultati ci hanno indotto a scegliere, come funzioni ammissibili a struttura fissata, le combinazioni lineari di funzioni di base parametrizzate. Chiameremo “reti approssimanti” le funzioni ottenute in corrispondenza di scelte delle funzioni di base che garantiscono proprietà di approssimazione particolarmente utili nella risoluzione dei problemi di ottimizzazione funzionale; tra queste vi sono le funzioni di base parametrizzate citate poco sopra. Ne deriva un metodo approssimato di ottimizzazione che abbiamo chiamato “Metodo di Ritz Estes” o, per brevità, ERIM (“Extended RItz Method”). Un’impressionante quantità di risultati sperimentali ci spinge ad esprimere la congettura che la proprietà di crescita “moderata” del numero di parametri da ottimizzare al crescere del numero  $d$  di variabili delle funzioni ammissibili, dimostrata nella teoria dell'approssimazione di funzioni, valga anche nel campo dell'ottimizzazione funzionale approssimata. Si vedano non solo i problemi di controllo ottimo (ad esempio, [34, 35, 39, 44]), ma anche altri problemi di ottimizzazione funzionale (ad esempio, [1, 5, 6, 45]).

Risultati teorici della stessa estensione di quelli ottenuti nel campo dell'approssimazione di funzioni non sono ancora disponibili per l'ERIM. Si tenga del resto presente che il problema della minimizzazione di un funzionale di costo è in generale assai più difficile del problema dell'approssimazione di una funzione. Nell'ottimizzazione funzionale, infatti, il funzionale di costo  $F$  è di solito ben più complesso della norma con cui si misura l'errore di approssimazione. Per giunta, possono comparire vincoli sulle funzioni ammissibili, rappresentati dalla regione di ammissibilità  $S$ . Risultati teorici preliminari [21, 27, 28], oltre all'evidenza sperimentale ricordata, ci inducono tuttavia a ritenere che la “parsimonia parametrica” di opportune combinazioni lineari di funzioni di base parametrizzate sussista anche nella risoluzione approssimata di problemi di ottimizzazione funzionale.

## 2 Formulazione generale di un problema di ottimizzazione funzionale ed un esempio

Si consideri un funzionale di costo  $F: S \mapsto \mathbb{R}$ , dove l'insieme  $S$  delle funzioni ammissibili è un *sottoinsieme di uno spazio di funzioni  $H$ , lineare, reale e a dimensione*

*infinita*<sup>2, 3</sup>

$$\underline{\gamma}(x) : \mathcal{B} \mapsto \mathbb{R}^{n_2}, \quad \text{dove } \mathcal{B} \subseteq \mathbb{R}^{n_1}. \quad (1)$$

Possiamo allora formulare il seguente problema di ottimizzazione funzionale.

**Problema P:**

$$\inf_{\underline{\gamma} \in S} F(\underline{\gamma}).$$

In molti problemi di ottimizzazione è necessario considerare più funzioni di decisione, che appartengono ad un insieme di funzioni ammissibili  $S_M \subseteq H_1 \times \cdots \times H_M$ . Per ogni  $i = 1, \dots, M$ ,  $H_i$  è uno spazio lineare, reale e a dimensione infinita di funzioni  $\underline{\gamma}_i : \mathcal{B}_i \mapsto \mathbb{R}^{n_{2i}}$ ,  $\mathcal{B}_i \subseteq \mathbb{R}^{n_{1i}}$  e  $F$  è un funzionale di costo definito su  $S_M$ . Introduciamo allora la funzione  $\underline{\Gamma} \triangleq \text{col}(\underline{\gamma}_1, \dots, \underline{\gamma}_M)$ , ottenuta “incolonnando” le funzioni  $\underline{\gamma}_1, \dots, \underline{\gamma}_M$ , ed estendiamo come segue la formulazione del Problema P.

**Problema PM:**

$$\inf_{\underline{\Gamma} \in S_M} F(\underline{\Gamma}).$$

Come risulterà chiaro nel seguito, le proprietà teoriche del Problema P, di interesse per la nostra trattazione, possono essere estese al Problema PM. Per non appesantire le notazioni, faremo quindi riferimento al Problema P.

Il metodo che presenteremo ha l’obiettivo di risolvere in modo approssimato il Problema P nella sua forma generale; concentreremo tuttavia la nostra attenzione su problemi di ottimizzazione funzionale di tipo stocastico. Prenderemo dunque in considerazione funzionali di costo esprimibili come valori attesi, aventi cioè la forma di funzionali integrali, in cui l’integrazione viene eseguita rispetto a una misura di probabilità. Si avrà cioè

---

<sup>2</sup>La distinzione tra quantità (variabili e funzioni) scalari e quantità a più componenti richiederà al lettore una certa attenzione. Inoltre, la dimensione di alcuni vettori, tra i quali il vettore argomento delle funzioni ammissibili, giocherà un ruolo fondamentale nella trattazione, con particolare riferimento alle problematiche connesse con la maledizione della dimensionalità. Abbiamo quindi scelto di indicare esplicitamente la natura vettoriale delle quantità mediante simboli sottolineati.

<sup>3</sup>Si ricordi che  $\nu$  elementi  $\underline{\gamma}_1, \dots, \underline{\gamma}_\nu$  di uno spazio vettoriale lineare reale  $H$  si dicono *linearmente indipendenti* se non è possibile trovare  $\nu$  numeri reali  $c_1, \dots, c_\nu$  non tutti nulli, tali che  $c_1 \underline{\gamma}_1 + \dots + c_\nu \underline{\gamma}_\nu = \underline{0}$ . Se invece esistono  $\nu$  numeri reali  $c_1, \dots, c_\nu$  non tutti nulli, che soddisfano la condizione sopra riportata, gli elementi  $\underline{\gamma}_1, \dots, \underline{\gamma}_\nu$  sono detti *linearmente dipendenti*; in tal caso, ogni elemento  $\underline{\gamma}_j$  con  $c_j \neq 0$  può essere espresso come una combinazione lineare degli altri elementi. Se, per ogni intero positivo  $\nu$ , esistono  $\nu$  elementi linearmente indipendenti in  $H$ , si dice che  $H$  ha *dimensione infinita*. Se  $H$  non ha dimensione infinita, esiste un intero positivo  $\nu$  per il quale vale quanto segue: in  $H$  vi sono  $\nu$  elementi linearmente indipendenti, ma, per ogni intero  $\mu > \nu$ ,  $\mu$  elementi di  $H$  sono linearmente dipendenti; l’intero  $\nu$  è detto *dimensione* di  $H$  e si dice che  $H$  è uno spazio a *dimensione finita*. In tal caso, ogni insieme di  $\nu$  elementi linearmente indipendenti di  $H$  viene chiamato *base* per  $H$ .

$$F(\underline{\gamma}) = \mathbb{E}_{\underline{z}} \{J[\underline{\gamma}(\underline{x}), \underline{z}]\}, \quad (2)$$

dove  $J(\cdot, \cdot)$  è un funzionale di costo assegnato,  $\underline{x}$  è un vettore aleatorio dipendente da  $\underline{z}$  attraverso una funzione nota e  $\underline{z}$  è un vettore aleatorio “primitivo”.

Presentiamo ora un classico problema di controllo ottimo in ambiente stocastico: vedremo che tale problema può essere considerato un esempio di Problema PM. Un sistema dinamico a tempo discreto (in generale non lineare) è retto dall’equazione di stato

$$\underline{x}_{t+1} = \underline{f}_t(\underline{x}_t, \underline{u}_t, \underline{\xi}_t), \quad t = 0, 1, \dots, T-1, \quad (3)$$

dove  $\underline{x}_t \in \mathbb{R}^n$ ,  $\underline{u}_t \in U_t \subseteq \mathbb{R}^m$  e  $\underline{\xi}_t \in \mathbb{R}^q$  sono rispettivamente i vettori di stato, di controllo e di disturbo e  $U_t$  è l’insieme dei controlli ammissibili.  $T$  rappresenta il numero di stadi temporali su cui si articola il processo decisionale. Di solito, non tutte le componenti del vettore di stato sono misurabili; vi è inoltre la possibilità che non tutte siano misurabili senza essere affette da disturbi. Definiamo allora il vettore  $\underline{y}_t$  dei segnali provenienti dai dispositivi di misura. Supponendo che questi siano privi di dinamica,  $\underline{y}_t$  è in generale una funzione non lineare dello stato e di un vettore di disturbi  $\underline{\eta}_t$ . Possiamo cioè scrivere

$$\underline{y}_t = \underline{g}_t(\underline{x}_t, \underline{\eta}_t), \quad t = 0, 1, \dots, T-1. \quad (4)$$

Facciamo l’ipotesi che le proprietà statistiche dello stato iniziale  $\underline{x}_0$  (considerato un vettore aleatorio) e delle sequenze dei disturbi  $\{\underline{\xi}_t\}_{t=0}^{T-1}$ ,  $\{\underline{\eta}_t\}_{t=0}^{T-1}$  siano note. Si definisca infine il funzionale di costo

$$J = \sum_{t=0}^{T-1} h_t(\underline{x}_t, \underline{u}_t) + h_T(\underline{x}_T), \quad (5)$$

dove  $h(\underline{x}_t, \underline{u}_t)$  rappresenta il costo della generica transizione dallo stato  $\underline{x}_t$  allo stato  $\underline{x}_{t+1}$  sotto l’azione del controllo  $\underline{u}_t$  e  $h_T(\underline{x}_T)$  è il costo finale.

Osserviamo ora che l’insieme delle informazioni, di cui il sistema di controllo dispone all’istante  $t$ , è costituito dalle misure  $\underline{y}_0, \underline{y}_1, \dots, \underline{y}_t$  e dai controlli  $\underline{u}_0, \underline{u}_1, \dots, \underline{u}_{t-1}$  generati fino all’istante  $t-1$ . Definiamo allora il *vettore delle informazioni*

$$\underline{I}_t = \text{col}(\underline{y}_0, \dots, \underline{y}_t, \underline{u}_0, \dots, \underline{u}_{t-1}). \quad (6)$$

Una *legge di controllo* che aspiri ad essere ottima dovrà necessariamente tener conto di tutte le informazioni acquisite stadio dopo stadio ed essere quindi costituita dalle *funzioni di controllo*

$$\underline{u}_0 = \underline{\mu}_0(\underline{I}_0), \underline{u}_1 = \underline{\mu}_1(\underline{I}_1), \dots, \underline{u}_{T-1} = \underline{\mu}_{T-1}(\underline{I}_{T-1}). \quad (7)$$

È ora possibile formalizzare il seguente problema di controllo ottimo.

**Problema C.** Determinare la sequenza delle funzioni ottime  $\underline{\mu}_0^\circ, \underline{\mu}_1^\circ, \dots, \underline{\mu}_{T-1}^\circ$ , che minimizzano il valor medio del funzionale di costo (5), rispettando i vincoli (3), (4) e  $\underline{\mu}_t(\underline{I}_t) \in U_t$ ,  $t = 0, 1, \dots, T - 1$ .

Ovviamente il valor medio del costo (5) va calcolato rispetto a tutti i vettori aleatori precedentemente introdotti. Ponendo dunque  $M = T$  e  $\underline{\gamma}_t^\circ = \underline{\mu}_{t-1}^\circ$  ( $t = 1, \dots, M$ ), il Problema C risulta avere la forma del Problema PM.

È noto che se le (3) e le (4) sono lineari, se le funzioni che compaiono nel costo (5) sono forme quadratiche e se tutti i vettori aleatori sono tra loro mutuamente indipendenti e gaussiani (se sono cioè verificate le ipotesi LQG già ricordate), allora il Problema C può essere risolto per via analitica. Le funzioni di controllo ottime risultano essere  $\underline{u}_t^\circ = -L_t \hat{x}_t$ , dove  $L_t$  è una matrice di guadagno e  $\hat{x}_t \triangleq E(\underline{x}_t | \underline{I}_t)$  è il valor medio condizionato del vettore di stato, calcolato mediante il filtro di Kalman.

Ben diversa è la situazione in cui le ipotesi LQG non siano soddisfatte. Se i vettori aleatori sono tra loro mutuamente indipendenti, il problema è formalmente risolvibile applicando la programmazione dinamica. A tal fine, si richiede il calcolo delle densità di probabilità condizionate  $p(\underline{x}_0 | \underline{I}_0), p(\underline{x}_1 | \underline{I}_1), \dots, p(\underline{x}_{T-1} | \underline{I}_{T-1})$ , la cui determinazione implica oneri computazionali in pratica insostenibili, a meno che il numero di componenti dei vettori coinvolti e il numero di stadi temporali siano molto ridotti o il sistema dinamico sia particolarmente semplice. Una delle prime trattazioni su tale argomento può essere trovata in [3]. Nella sezione 11 risolveremo il Problema C con l'ERIM e nella sezione 12 ne presenteremo un esempio applicativo.

### 3 Il Metodo di Ritz Esteso per l'ottimizzazione funzionale approssimata: linee generali

Una possibile procedura di risoluzione approssimata del Problema P consiste nel far ricorso ad una sequenza di problemi approssimanti di più semplice risoluzione. Un esempio di tale metodologia è dato dal metodo delle funzioni di penalità nei problemi di programmazione non lineare vincolata. L'approccio che seguiremo consiste nell'imporre preliminarmente una struttura assegnata alle funzioni ammissibili (1). Detta  $\underline{\gamma}_\nu$  tale struttura, le funzioni decisionali assumono quindi la forma

$$\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu) : \mathcal{B} \times \mathbb{R}^N \mapsto \mathbb{R}^{n_2}, \quad \mathcal{B} \subseteq \mathbb{R}^{n_1}, \quad (8)$$

dove  $\nu \in \mathbb{Z}^+$ ,  $\mathbb{Z}^+$  indica l'insieme dei numeri interi positivi e  $\underline{w}_\nu$  è un vettore di parametri liberi, da determinare in modo da minimizzare il funzionale di costo. L'intero  $N$  rappresenta il numero di componenti del vettore  $\underline{w}_\nu$ , cioè il numero totale di parametri. Come risulterà chiaro nel seguito, le strutture  $\underline{\gamma}_\nu(\cdot, \cdot)$  da noi scelte ci consentiranno di esprimere  $N$  in funzione di  $\nu$ . Scriviamo quindi

$$N(\nu) : \mathbb{Z}^+ \mapsto \mathbb{Z}^+. \quad (9)$$

L'ERIM si basa sul seguente concetto guida. La minimizzazione rispetto a tutte le funzioni ammissibili  $\underline{\gamma} \in S$  (si veda la (1)) viene sostituita da una successione di minimizzazioni rispetto a funzioni  $\underline{\gamma}_\nu$  (si veda la (8)) che appartengono a famiglie di funzioni parametrizzate. All'aumentare del numero  $N(\nu)$  di parametri, tali famiglie danno origine a una struttura di insiemi a inclusione, che “invade” l'insieme  $S$  delle funzioni ammissibili.

Sostituendo le funzioni  $\underline{\gamma}_\nu$  nel funzionale  $F$  ed effettuando le operazioni da esso richieste (derivate, somme, integrali, ecc.),  $F$  diventa una funzione degli  $N(\nu)$  parametri. In tal modo, il Problema P di ottimizzazione funzionale viene ridotto ad un problema di programmazione non lineare, consistente nel minimizzare una funzione di  $N(\nu)$  variabili reali. La possibilità di ottenere una soluzione approssimata del Problema P, con accuratezza arbitraria e con un onere computazionale limitato, dipende dalla scelta delle funzioni (8).

La metodologia generale proposta si articola nelle seguenti fasi.

i) Si sceglie una struttura sufficientemente semplice per le funzioni parametrizzate  $\underline{\gamma}_\nu$ . Tale scelta porta a definire le *reti a uno strato nascosto* o *reti OHL* (dall'inglese “One Hidden Layer”) (sezione 4).

Le fasi successive sfruttano poi:

a) il criterio che viene scelto per misurare l'accuratezza di una soluzione approssimata;  
 b) le peculiarità della particolare istanza del Problema P, dove per “istanza” si intende la terna  $(H, S, F)$ , che definisce il problema.

ii) Si correda lo spazio  $H$  con una norma  $\|\cdot\|$  scelta sulla base del criterio di cui al punto a). Tale norma induce una distanza tra le funzioni di  $\mathcal{H} \triangleq (H, \|\cdot\|)$  e consente di valutare l'errore commesso approssimando una funzione di  $\mathcal{H}$  con una rete OHL. Le reti OHL vengono “personalizzate” allo spazio  $\mathcal{H}$  richiedendo che soddisfino la proprietà di densità rispetto alla norma  $\|\cdot\|$ . In altre parole, l'impiego di un numero sufficientemente elevato di parametri deve consentire di approssimare arbitrariamente bene qualunque funzione di  $\mathcal{H}$ . In tal modo, si definiscono le *reti dense in  $\mathcal{H}$*  o  *$\mathcal{H}$ -DN* (dall'inglese “ $\mathcal{H}$ -Dense Networks”); per riflettere questo nelle notazioni, scriveremo  $\mathcal{S} \subseteq \mathcal{H}$  invece di  $S \subseteq H$  (sezione 5).

iii) Si restringe la classe di reti OHL a reti che, in relazione al criterio di misura dell'accuratezza e ad un certo insieme di funzioni  $\mathcal{S}$ , abbiano complessità ridotta. Più precisamente, si ricercano reti OHL che consentano di approssimare con accuratezza arbitraria le funzioni (1) usando un numero “non troppo grande” di parametri, anche quando tali funzioni dipendono da un numero elevato di variabili. Con ciò si definiscono le *reti a complessità polinomiale in  $\mathcal{S}$*  (sezione 6).

iv) Si impiegano le reti OHL per definire una sequenza di problemi approssimanti e si stabilisce in che senso tali problemi devono approssimare un'istanza del Problema P.

In tal modo, si introduce il concetto di *reti P-ottimizzanti* o *reti P-ON* (dall'inglese "P-Optimizing Networks") (sezione 8).

v) Si impone che l'ottimizzazione, di cui alla fase precedente, sia ottenibile usando reti OHL con un numero di parametri "non troppo grande", anche quando le funzioni (1) dipendono da un numero elevato di variabili. Questa fase porta a definire le *reti P-ottimizzanti a complessità polinomiale* o *P-ON a complessità polinomiale* (sezione 9).

vi) In presenza di problemi di ottimizzazione funzionale stocastica, per evitare il calcolo del valore atteso della funzione di costo (in pratica non eseguibile), si risolve il problema finale di programmazione non lineare con tecniche di approssimazione stocastica (sezione 10).

Per motivi di chiarezza espositiva, è opportuno riportare sin d'ora la figura che stabilisce le interconnessioni tra le vari famiglie di reti sopra citate. Si faccia dunque riferimento alla figura 1. Il significato di tale figura verrà chiarito via via che il Metodo di Ritz Estesò sarà descritto e commentato. Si può comunque osservare che la figura è divisa in due parti, entrambe ottenute a partire da reti OHL. A sinistra sono evidenziate le reti ottenute richiedendo la capacità di approssimare arbitrariamente bene ogni funzione di  $\mathcal{H}$  o di  $\mathcal{S}$ , indipendentemente dall'istanza del Problema P. Le reti appartenenti agli insiemi della parte destra, invece, consentono di determinare in modo approssimato le soluzioni ottime del Problema P e dipendono quindi dalla particolare istanza del problema, specificata dalla terna  $(H, S, F)$

Il termine "reti approssimanti", usato nel titolo di questo lavoro, intende indicare genericamente reti OHL che godono delle proprietà di approssimazione sopra illustrate.

## 4 Reti a uno strato nascosto

Una scelta piuttosto naturale per le funzioni a struttura fissata  $\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu)$  consiste nell'impiegare, per ciascuna delle componenti  $\gamma_{\nu j}(\underline{x}, \underline{w}_\nu)$ , una combinazione di funzioni di base preassegnate avente la forma

$$\gamma_{\nu j}(\underline{x}, \underline{w}_\nu) = \sum_{i=1}^{\nu} c_{ij} \varphi_i(\underline{x}), \quad c_{ij} \in \mathbb{R}, \quad j = 1, \dots, n_2, \quad (10)$$

dove  $\varphi_1(\underline{x}), \dots, \varphi_\nu(\underline{x}) \in H$  sono le prime  $\nu$  funzioni di una sequenza di funzioni di base e  $\underline{w}_\nu \triangleq \text{col}(c_{ij} : i = 1, \dots, \nu; j = 1, \dots, n_2)$ . La (8) assume quindi la forma

$$\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu) = \text{col} \left[ \sum_{i=1}^{\nu} c_{ij} \varphi_i(\underline{x}) : j = 1, \dots, n_2 \right] \quad (11)$$

e il numero di parametri liberi (si veda la (9)) è dato semplicemente da  $N(\nu) = \nu n_2$ .

Figura 1: relazioni tra gli insiemi di reti approssimanti ottenuti a partire da reti OHL.

La scelta delle combinazioni lineari (10) per approssimare le funzioni  $\underline{\gamma}(\underline{x})$  è motivata dal teorema di Weierstrass. Da tale teorema sappiamo infatti che funzioni  $\underline{\gamma}(\underline{x})$  continue e definite su insiemi compatti possono essere approssimate arbitrariamente bene mediante polinomi di grado sufficientemente elevato.

Sostituendo poi le funzioni (11) nel funzionale  $F$ , si ottiene la funzione

$$F_\nu(\underline{w}_\nu) \triangleq F[\underline{\gamma}_\nu(\cdot, \underline{w}_\nu)].$$

In tal modo, per ogni  $\nu$ , il Problema P di ottimizzazione funzionale è ridotto ad un problema di programmazione non lineare, da risolversi mediante un opportuno algoritmo di discesa. Qualora si usino le combinazioni lineari (10), il metodo ora descritto coincide in pratica con il metodo classico di Ritz del calcolo delle variazioni [13]. Torneremo su questo punto nelle sezioni successive.

Concentriamo ora la nostra attenzione su possibili estensioni delle funzioni a struttura fissata (11). Un'estensione abbastanza naturale consiste nell'inserire alcuni parametri liberi nelle funzioni di base. Indicando con  $\underline{\kappa}_i$  il vettore di tali parametri liberi, le nuove funzioni di base assumono la forma  $\varphi_i(\underline{x}, \underline{\kappa}_i)$ . Poichè il vettore  $\underline{\kappa}_i$  consente di aumentare notevolmente la “flessibilità” delle funzioni  $\varphi_i(\underline{x}, \underline{\kappa}_i)$ , nel seguito non useremo funzioni diverse  $\varphi_1(\cdot, \cdot), \dots, \varphi_\nu(\cdot, \cdot)$ , ma un'unica funzione  $\varphi(\cdot, \cdot)$ .

Chiamiamo le funzioni  $\varphi(\underline{x}, \underline{\kappa}_i)$ , tali che per ogni  $\underline{\kappa}_i$  si abbia  $\varphi(\cdot, \underline{\kappa}_i) \in H$ , “funzioni di base parametrizzate”. Le (10) vengono quindi sostituite da

$$\gamma_{\nu j}(\underline{x}, \underline{w}_\nu) = \sum_{i=1}^{\nu} c_{ij} \varphi(\underline{x}, \underline{\kappa}_i), \quad c_{ij} \in \mathbb{R}, \underline{\kappa}_i \in \mathbb{R}^k, j = 1, \dots, n_2, \quad (12)$$

dove  $\underline{w}_\nu \triangleq \text{col}(c_{ij}, \underline{\kappa}_i : i = 1, \dots, \nu; j = 1, \dots, n_2)$ . Il numero di parametri liberi risulta ora  $N(\nu) = \nu(k + n_2)$ . La (11) è dunque sostituita dalla

$$\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu) = \text{col} \left[ \sum_{i=1}^{\nu} c_{ij} \varphi(\underline{x}, \underline{\kappa}_i) : j = 1, \dots, n_2 \right]. \quad (13)$$

Poichè le funzioni (11) e (13) giocheranno un ruolo fondamentale nel seguito, introduciamo le seguenti definizioni.

**Definizione 4.1** Una sequenza  $\{A_\nu\}_{\nu=1}^\infty$ , dove  $A_\nu \triangleq \{\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu) : \underline{w}_\nu \in \mathbb{R}^{N(\nu)}\}$ ,  $\nu = 1, 2, \dots$ , e le funzioni  $\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu) : \mathbb{R}^{n_1} \times \mathbb{R}^{N(\nu)} \mapsto \mathbb{R}^{n_2}$  hanno la struttura (11) o (13), è detta “schema OHL”. Le funzioni di ciascun insieme  $A_\nu$  sono dette “reti a uno strato nascosto” o “reti OHL”.

**Definizione 4.2** Le reti OHL sono dette “lineari” (“non lineari”) se hanno la struttura (11) (la struttura (13)). A reti OHL lineari (non lineari) corrispondono schemi OHL “lineari” (“non lineari”).

Il termine “rete OHL” è mutuato dalla terminologia delle reti neurali, in quanto le (13) con funzioni di base sigmoidali hanno la stessa struttura delle reti neurali “feedforward” costituite da un solo “strato nascosto” di sigmoidi e da unità di uscita lineari (si veda la (18)). Sebbene il termine “rete OHL” non abbia tale giustificazione nel caso delle combinazioni lineari (11), è utile indicare le funzioni (11) e (13) con una denominazione comune. Poichè si ha  $N(\nu) = \nu n_2$  nel caso di reti lineari e  $N(\nu) = \nu(k + n_2)$  nel caso di reti non lineari, l’intero  $\nu$  risulta idoneo per misurare la complessità delle reti OHL: in entrambi i casi, infatti, il numero  $N(\nu)$  di parametri è proporzionale a  $\nu$ .

La sequenza  $\{A_\nu\}_{\nu=1}^\infty$  presenta evidentemente la struttura a inclusione

$$A_1 \subseteq A_2 \subseteq \dots \subseteq A_\nu \subseteq \dots. \quad (14)$$

Considerando poi l’intersezione di ciascun insieme  $A_\nu$  con l’insieme  $S$  delle funzioni ammissibili per il Problema P, si ottiene la sequenza  $\{A_\nu \cap S\}_{\nu=1}^\infty$  data da

$$A_\nu \cap S = \left\{ \underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu) : \underline{w}_\nu \in W_\nu \subseteq \mathbb{R}^{N(\nu)} \right\}, \quad \nu = 1, 2, \dots. \quad (15)$$

$W_\nu \subseteq \mathbb{R}^{N(\nu)}$  è l'insieme dei vettori ammissibili  $\underline{w}_\nu$ , definito dai vincoli che individuano l'insieme delle funzioni ammissibili  $S$ . Risulta cioè

$$W_\nu \triangleq \{ \underline{w}_\nu : \underline{\gamma}_\nu(\cdot, \underline{w}_\nu) \in A_\nu \cap S \}. \quad (16)$$

Anche la sequenza  $\{A_\nu \cap S\}_{\nu=1}^\infty$  è caratterizzata da una struttura a inclusione (si veda la figura 2).

Figura 2: la struttura a inclusione delle sequenze  $\{A_\nu\}_{\nu=1}^\infty$  e  $\{A_\nu \cap S\}_{\nu=1}^\infty$ .

Le funzioni di base parametrizzate  $\varphi(\underline{x}, \underline{\kappa}_i)$  possono essere costruite in vari modi. Seguendo [43], descriviamo tre modalità di frequente impiego, che ottengono le funzioni di base multivariabili a partire da una singola funzione monovariabile  $h : \mathbb{R} \mapsto \mathbb{R}$  o da una famiglia di funzioni  $h_1, \dots, h_{n_1} : \mathbb{R} \mapsto \mathbb{R}$ , anch'esse monovariabili.

(i) *Costruzione di tipo tensoriale*:  $\varphi(\underline{x}, \underline{\kappa}_i)$  viene ottenuta come prodotto di  $n_1$  funzioni monovariabili (eventualmente identiche):

$$\varphi(\underline{x}, \underline{\kappa}_i) = h_1(x_1, \tau_{i1}) \cdots h_{n_1}(x_{n_1}, \tau_{in_1}),$$

dove  $\tau_{is} \in \mathbb{R}^{l_{is}}$  è un vettore di parametri e  $\underline{\kappa}_i \triangleq \text{col}(\tau_{is} : s = 1, \dots, n_1)$ . Questa costruzione è utilizzata nelle basi di polinomi e di “spline” per approssimazioni basate su grigliature dello spazio del vettore argomento  $\underline{x}$ .

(ii) *Costruzione di tipo “ridge”*, in cui il vettore  $\underline{x}$  viene “contratto” in una variabile scalare mediante il prodotto scalare:

$$\varphi(\underline{x}, \underline{\kappa}_i) = h(\underline{x}^\top \underline{\alpha}_i + \beta_i), \quad (17)$$

dove  $\underline{\kappa}_i \triangleq \text{col}(\underline{\alpha}_i, \beta_i)$ . Le reti neurali di tipo “feedforward” con uno strato nascosto e unità di uscita lineari sono esempi di reti OHL basate sulla costruzione “ridge”; la  $j$ -esima componente di tali reti è data da

$$\gamma_{\nu j}(\underline{x}, \underline{w}_\nu) = \sum_{i=1}^{\nu} c_{ij} \sigma(\underline{x}^\top \underline{\alpha}_i + \beta_i), \quad (18)$$

dove  $\sigma(\cdot)$  è una funzione sigmoideale, cioè una funzione  $\sigma : \mathbb{R} \mapsto \mathbb{R}$  limitata, misurabile e tale che  $\lim_{z \rightarrow +\infty} \sigma(z) = 1$  e  $\lim_{z \rightarrow -\infty} \sigma(z) = 0$ <sup>4</sup>.

(iii) *Costruzione radiale*, in cui il vettore  $\underline{x}$  viene “contratto” in una variabile scalare per mezzo di una norma in  $\mathbb{R}^{n_1}$ . Una scelta tipica è

$$\varphi(\underline{x}, \underline{\kappa}_i) = h(\|\underline{x} - \underline{\tau}_i\|_{\Gamma_i}^2), \quad (19)$$

dove  $\Gamma_i \in \mathbb{R}^{n_1 \times n_1}$ ,  $\Gamma_i = \Gamma_i^\top$ ,  $\Gamma_i > 0$  e  $\underline{\kappa}_i = \text{col}(\underline{\tau}_i, \text{elementi non ridondanti di } \Gamma_i)$ ; i vettori  $\underline{\tau}_i$  sono detti “centri” o “centroidi”. Un esempio di rete OHL ottenuta con la costruzione radiale è la combinazione lineare di gaussiane

$$\gamma_{\nu j}(\underline{x}, \underline{w}_\nu) = \sum_{i=1}^{\nu} c_{ij} e^{-\|\underline{x} - \underline{\tau}_i\|_{\Gamma_i}^2}. \quad (20)$$

Va notato che la scelta degli esempi (18) e (20) non è casuale. In effetti vedremo che, tra le possibili scelte di reti OHL, queste (insieme ad altre) presentano proprietà di approssimazione particolarmente utili per la risoluzione approssimata del Problema P.

## 5 Reti dense

Nulla è stato sinora richiesto alle reti OHL circa la possibilità di approssimare le funzioni decisionali  $\underline{\gamma} \in S$  secondo le già citate proprietà di accuratezza e di “parsimonia” nel numero di parametri. Il concetto di accuratezza implica un criterio di misura dell’errore di approssimazione, che può essere formalizzato matematicamente introducendo una nozione di distanza tra le funzioni di  $H$ . In particolare, possiamo misurare l’errore di approssimazione mediante una norma  $\|\cdot\| : H \mapsto \mathbb{R}^+$ , che induce una distanza tra ogni coppia di funzioni di  $H$  nel Problema P. Dotiamo quindi lo spazio  $H$  di una norma  $\|\cdot\|$  e misuriamo la distanza tra  $\underline{\gamma}(\underline{x})$  e  $\underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu)$  mediante la quantità  $\|\underline{\gamma}(\underline{x}) - \underline{\gamma}_\nu(\underline{x}, \underline{w}_\nu)\|$ .

Da questo punto in poi, invece di  $H$ , considereremo quindi lo *spazio di funzioni normato*  $\mathcal{H} \triangleq (H, \|\cdot\|)$ , che “conserva memoria” della norma scelta. Scriveremo

<sup>4</sup>Abbiamo riportato una delle definizioni più diffuse di funzione sigmoideale (si vedano, ad esempio, [7], [38] e i riferimenti in essi contenuti), che tuttavia non è l’unica usata. Ad esempio, alcuni autori richiedono anche la continuità e/o la monotonicità (o addirittura la stretta monotonicità) di  $\sigma(\cdot)$  su  $\mathbb{R}$ .

esplicitamente la norma  $\|\cdot\|$  solo quando ciò sarà necessario per evitare ambiguità. Per mantenere coerenti le notazioni, nel seguito l'insieme delle funzioni ammissibili verrà indicato con  $\mathcal{S} \subseteq \mathcal{H}$ . Invece, per evitare un eccesso di simboli, manterremo la notazione  $A_\nu$  per gli insiemi  $A_\nu$  con norma, considerati quindi sottoinsiemi di  $\mathcal{H}$ .

Ci si chiede ora quali reti OHL siano in grado di approssimare con accuratezza arbitraria ogni funzione di  $\mathcal{H}$ . A tale scopo è importante la seguente definizione.

**Definizione 5.1** Una sequenza  $\{A_\nu\}_{\nu=1}^\infty$  tale che l'insieme  $\bigcup_{\nu=1}^\infty A_\nu$  è denso in  $\mathcal{H}$  è detta “schema di densità in  $\mathcal{H}$ ”. In tal caso, le reti OHL di ciascun insieme  $A_\nu$  sono dette “reti  $\mathcal{H}$ -dense” o “ $\mathcal{H}$ -DN”.

Definiamo inoltre “*schemi lineari (non lineari) di densità in  $\mathcal{H}$* ” quelli i cui elementi sono reti OHL lineari (non lineari). In corrispondenza di tali schemi, si hanno “*reti  $\mathcal{H}$ -dense lineari*” e “*reti  $\mathcal{H}$ -dense non lineari*”.

È chiaro che una  $\mathcal{H}$ -DN non corrisponde soltanto alla scelta della struttura di una funzione  $\underline{\gamma}$ , come nel caso di una rete OHL, ma soddisfa anche un importante requisito per l'approssimazione di funzioni. Nelle applicazioni sono particolarmente importanti i due spazi  $\mathcal{H}$  descritti di seguito.

i) Lo spazio  $H = C(K, \mathbb{R}^{n_2})$  delle funzioni continue  $\underline{\gamma}(\underline{x}) : K \mapsto \mathbb{R}^{n_2}$  ( $K \subset \mathbb{R}^{n_1}$  è un insieme compatto), con la cosiddetta “norma del sup”

$$\|\underline{\gamma}\|_\infty = \max_{\underline{x} \in K} \|\underline{\gamma}(\underline{x})\|_{\mathbb{R}^{n_2}}, \quad (21)$$

dove  $\|\cdot\|_{\mathbb{R}^{n_2}}$  indica una norma su  $\mathbb{R}^{n_2}$  (che non specifichiamo, dal momento che in spazi a dimensione finita tutte le norme sono equivalenti). Per brevità useremo la notazione  $\mathcal{C}(K, \mathbb{R}^{n_2}) \triangleq (C(K, \mathbb{R}^{n_2}), \|\cdot\|_\infty)$ . Nella letteratura sulle reti neurali, le reti  $\mathcal{C}$ -dense sono anche chiamate (con un termine piuttosto altisonante) “approssimatori universali”.

ii) Lo spazio  $H = L_2(K, \mathbb{R}^{n_2})$  delle funzioni misurabili di quadrato integrabile  $\underline{\gamma}(\underline{x}) : K \mapsto \mathbb{R}^{n_2}$ , con la cosiddetta “norma  $\mathcal{L}_2$ ”

$$\|\underline{\gamma}\|_2 = \left[ \int_K \|\underline{\gamma}(\underline{x})\|_{\mathbb{R}^{n_2}}^2 d\underline{x} \right]^{1/2}.$$

Nel seguito useremo la notazione  $\mathcal{L}_2(K, \mathbb{R}^{n_2}) \triangleq (L_2(K, \mathbb{R}^{n_2}), \|\cdot\|_2)$ .

Associate a questi spazi avremo quindi reti  $\mathcal{C}$ - ed  $\mathcal{L}_2$ -dense. Altri spazi  $\mathcal{H}$  di interesse sono, tra gli altri, gli spazi  $\mathcal{L}_p(K, \mathbb{R}^{n_2})$ ,  $1 \leq p \leq \infty$ . La densità delle reti OHL basate sulle costruzioni di tipo “ridge” (18) e radiale (20) è dimostrata, ad esempio, rispettivamente in [38] e [36].

Nel seguito useremo spesso la sigla “DN” senza specificare lo spazio  $\mathcal{H}$ . Lo faremo ogniqualvolta risulterà chiaro dal contesto quale spazio si considera, o si esamineranno le reti dense senza voler fare riferimento ad uno spazio specifico.

## 6 Reti a complessità polinomiale

La densità nello spazio  $\mathcal{H}$  può essere considerata una proprietà “desiderabile” che le reti OHL (lineari o non lineari) è opportuno che abbiano per essere utilizzate con successo nella risoluzione approssimata del Problema P.

Ulteriori proprietà di approssimazione sono poi richieste quando si considera l’impiego delle reti OHL in problemi di ottimizzazione funzionale nei quali le funzioni ammissibili dipendono da un numero elevato di variabili. In questo caso, importanti differenze possono intervenire tra reti OHL lineari e reti OHL non lineari. Infatti, la relazione tra l’accuratezza di approssimazione ottenibile mediante un certo schema OHL e la complessità delle reti associate a tale schema (misurata dal numero  $\nu$  di funzioni di base) è in generale sostanzialmente diversa per i due tipi di rete. Tale relazione descrive la velocità con cui l’errore di approssimazione tende a zero quando  $\nu$  tende all’infinito, cioè quando si usa un numero di funzioni di base arbitrariamente elevato. Da questo punto di vista, numerosi risultati sperimentali e teorici evidenziano vantaggi sostanziali degli schemi OHL non lineari.

Per non appesantire eccessivamente le notazioni, prenderemo in considerazione (senza perdita di generalità) spazi  $\mathcal{H}$  i cui elementi sono funzioni scalari  $\gamma : \mathbb{R}^d \mapsto \mathbb{R}$  (cioè,  $n_1 = d$  e  $n_2 = 1$  nel Problema P). Misureremo le prestazioni di un dato schema OHL  $\{A_\nu\}_{\nu=1}^\infty$  in termini dell’errore di approssimazione di tipo “worst-case”, commesso nell’approssimare funzioni di  $\mathcal{S}$  mediante funzioni di  $\{A_\nu\}_{\nu=1}^\infty$ . Tale errore è formalizzato dal concetto di *deviazione* tra insiemi, che particolarizziamo al contesto di nostro interesse mediante la seguente definizione.

**Definizione 6.1** *Dati i due sottoinsiemi  $\mathcal{S}$  e  $A_\nu$  dello spazio normato  $\mathcal{H} = (H, \|\cdot\|)$ , si definisce “deviazione di  $\mathcal{S}$  da  $A_\nu$ ” la quantità*

$$\delta(\mathcal{S}, A_\nu) \triangleq \sup_{\gamma \in \mathcal{S}} \inf_{\underline{\gamma} \in A_\nu} \|\underline{\gamma} - \gamma\|. \quad (22)$$

Una volta che si è scelto un certo schema OHL  $\{A_\nu\}_{\nu=1}^\infty$  e si è fissato il numero  $\nu$  di funzioni di base, la quantità (22) misura la distanza, nella norma di  $\mathcal{H}$ , tra la funzione ammissibile “più difficile da approssimare” e la funzione in  $A_\nu$  che la approssima “nel modo migliore”.

Nel caso di reti OHL lineari, fissato un certo intero  $\nu$ , è importante determinare quale sia il più piccolo valore della deviazione (22) conseguibile. Questo significa lasciare libera la scelta del tipo di funzioni di base  $\varphi_1, \dots, \varphi_\nu$  nella (11) e determinare l’estremo inferiore di (22) rispetto alla scelta di tali funzioni. Più precisamente, si consideri una rete OHL lineare le cui  $\nu$  funzioni di base siano linearmente indipendenti. Ciascun insieme  $A_\nu$  è allora un sottospazio di  $\mathcal{H}$  di dimensione  $\nu$ , le cui funzioni sono esprimibili come combinazioni lineari di funzioni  $\varphi_1, \dots, \varphi_\nu$ . Indichiamo con  $span(\varphi_1, \dots, \varphi_\nu)$  tale sottospazio. Sostituiamo allora  $A_\nu = span(\varphi_1, \dots, \varphi_\nu)$  nella (22) e valutiamone l’estremo inferiore rispetto alle funzioni di base  $\varphi_1, \dots, \varphi_\nu$ .

Ciò porta in modo naturale al concetto di “ $\nu$ -width” di un insieme, introdotto da Kolmogorov [23]. Ne riportiamo la definizione, particolarizzata al contesto di nostro interesse (si vedano anche [7, p. 942] e [37, Capitolo I]).

**Definizione 6.2** *Dato un sottoinsieme  $\mathcal{S}$  di uno spazio normato  $\mathcal{H} = (H, \|\cdot\|)$ , si definisce “Kolmogorov  $\nu$ -width” (o semplicemente “ $\nu$ -width”) di  $\mathcal{S}$  in  $\mathcal{H}$  la quantità*

$$d_\nu(\mathcal{S}) = \inf_{\varphi_1, \dots, \varphi_\nu \in \mathcal{H}} \delta(\mathcal{S}, \text{span}(\varphi_1, \dots, \varphi_\nu)). \quad (23)$$

Si noti che la definizione di “ $\nu$ -width” richiede di determinare un estremo inferiore in più rispetto alla definizione di deviazione.

L’errore “worst-case” (23) va poi posto a confronto con quello ottenibile tramite lo schema non lineare di interesse. Le proprietà approssimanti di due diversi schemi OHL non lineari  $\{A_\nu\}_{\nu=1}^\infty$  e  $\{A'_\nu\}_{\nu=1}^\infty$  saranno quindi confrontate mediante le quantità  $\delta(\mathcal{S}, A_\nu)$  e  $\delta(\mathcal{S}, A'_\nu)$ . Un dato schema non lineare  $\{A_\nu\}_{\nu=1}^\infty$  e il miglior schema lineare saranno invece confrontati mediante le quantità  $\delta(\mathcal{S}, A_\nu)$  e  $d_\nu(\mathcal{S})$ .

Nello sviluppo di algoritmi per la risoluzione approssimata del Problema P, vogliamo evitare schemi OHL la cui complessità  $\nu$ , per una data accuratezza di approssimazione  $\varepsilon$ , cresca “troppo velocemente” con il numero  $d$  di variabili delle funzioni ammissibili. L’espressione “una data accuratezza di approssimazione  $\varepsilon$ ” significa che gli errori “worst-case” (22) e (23) sono al più  $\varepsilon$ . La seguente distinzione è di importanza fondamentale nel valutare il comportamento di (22) e (23) in funzione del numero  $d$  di variabili nelle funzioni decisionali ammissibili.

- 1) Supponiamo che per  $\delta(\mathcal{S}, A_\nu)$  o per  $d_\nu(\mathcal{S})$  esista un “upper bound” di ordine  $O(d^p/\nu^q)$ , dove  $p, q \in \mathbb{R}^+$ . In corrispondenza di valori sufficientemente elevati di  $\nu$ , esiste allora una costante  $c > 0$  tale che  $\varepsilon \leq c d^p/\nu^q$ . Possiamo pertanto scrivere

$$\nu \leq \left(\frac{c}{\varepsilon}\right)^{1/q} d^{p/q}. \quad (24)$$

Poichè in tal caso il numero  $\nu$  di funzioni di base, necessarie per garantire una data accuratezza di approssimazione  $\varepsilon$ , deve crescere al più come una potenza di  $d$ , si può concludere che lo schema OHL  $\{A_\nu\}_{\nu=1}^\infty$  presenta un comportamento “favorevole” rispetto al numero  $d$  di variabili nelle funzioni decisionali  $\underline{\gamma}(x)$ .

- 2) Supponiamo che per  $\delta(\mathcal{S}, A_\nu)$  o  $d_\nu(\mathcal{S})$  esista un “lower bound” di ordine  $O(1/\nu^{1/d})$ . Ne segue che, per garantire un’accuratezza di approssimazione  $\varepsilon$ , sono necessarie reti OHL di complessità  $O(1/\varepsilon^d)$ . Tale dipendenza esponenziale della complessità da  $d$  conduce alla maledizione della dimensionalità [9].

La scelta dei due ordini  $O(d^p/\nu^q)$  e  $O(1/\nu^{1/d})$  per effettuare la classificazione di cui sopra è motivata dal fatto che molte stime dell’errore di approssimazione di tipo

“worst-case” sono espresse in questa forma. La differenza sostanziale tra i due comportamenti descritti consente di distinguere tra reti OHL computazionalmente utilizzabili e reti non utilizzabili per approssimare funzioni decisionali ammissibili. Tale differenza costituisce quindi un valido criterio per individuare, nella risoluzione approssimata di istanze del Problema P, un certo tipo di rete OHL piuttosto che un altro. A questo proposito introduciamo la seguente definizione.

**Definizione 6.3** *Una sequenza  $\{A_\nu\}_{\nu=1}^\infty$  per la quale sussiste l’“upper bound” (24) è detta “schema a complessità polinomiale in  $\mathcal{S}$ ”. In tal caso, le reti OHL di ciascun insieme  $A_\nu$  sono dette “reti a complessità polinomiale in  $\mathcal{S}$ ”.*

Da un punto di vista qualitativo, reti OHL caratterizzate dall’“upper bound” polinomiale (24) sulla complessità  $\nu$  rendono “computazionalmente trattabile” l’approssimazione di funzioni in  $\mathcal{S}$ .

Si noti che, per un valore fissato di  $d$ , l’“upper bound”  $\varepsilon \leq cd^p/\nu^q$  implica che  $\varepsilon \rightarrow 0$  quando  $\nu \rightarrow \infty$ ; la proprietà di densità in  $\mathcal{S}$  deriva quindi dalla complessità polinomiale in  $\mathcal{S}$ . Osserviamo che si possono avere reti a complessità polinomiale in un sottoinsieme proprio  $\mathcal{S}$  di  $\mathcal{H}$  (e quindi dense in  $\mathcal{S}$ ), ma non necessariamente dense nell’insieme differenza  $\mathcal{H} \setminus \mathcal{S}$ . Ne risulta che l’insieme delle reti a complessità polinomiale in  $\mathcal{S}$  può non essere contenuto nell’insieme delle  $\mathcal{H}$ -DN. Tale possibilità è evidenziata nella parte sinistra della figura 1. Tuttavia, in pratica, le reti a complessità polinomiale in un sottoinsieme proprio  $\mathcal{S}$  di  $\mathcal{H}$  sono tipicamente ottenute partendo da reti OHL dense in tutto lo spazio  $\mathcal{H}$ , cioè partendo da  $\mathcal{H}$ -DN (ipotesi relativamente blande sulle funzioni di base consentono infatti di ottenere la densità in tutto  $\mathcal{H}$ , per un’ampia varietà di spazi di interesse nell’ottimizzazione funzionale).

Nel seguito di questa sezione presenteremo alcune classi di reti OHL, che hanno complessità polinomiale per vari spazi  $\mathcal{H}$  ed insiemi  $\mathcal{S}$ , di interesse nelle applicazioni. Useremo spesso l’espressione “rete a complessità polinomiale”, senza specificare lo spazio  $\mathcal{H}$  o l’insieme  $\mathcal{S}$ . Faremo questo ogniqualvolta risulterà chiaro dal contesto quale spazio e quale insieme si considerano, o si vorrà fare riferimento a reti a complessità polinomiale senza considerare una specifica scelta di  $\mathcal{H}$  e di  $\mathcal{S}$ .

Una prima classe di reti a complessità polinomiale di frequentissimo impiego è costituita dalle reti OHL con funzioni di base sigmoidali (si veda la (18)). Un interessante confronto teorico tra gli errori di approssimazione di tipo “worst-case” (22) e (23) per reti OHL lineari e non lineari di tipo sigmoidale è presentato in [7], in cui viene considerata la seguente classe di funzioni:

$$G_c^d \triangleq \left\{ \gamma: \mathbb{R}^d \mapsto \mathbb{R} \quad \text{tale che} \quad \int_{\mathbb{R}^d} \|\underline{\omega}\|_2 |\tilde{\gamma}(\underline{\omega})| d\underline{\omega} \leq c \right\}, \quad (25)$$

dove  $\tilde{\gamma}(\underline{\omega})$  è la trasformata di Fourier di  $\gamma$ ,  $\|\underline{\omega}\|_2 = (\underline{\omega}^\top \underline{\omega})^{1/2}$  e  $c$  è uno scalare positivo. In [7] è dimostrato che, misurando l’errore di approssimazione in norma  $\mathcal{L}_2$  nella palla  $B_1^d$  di raggio unitario in  $\mathbb{R}^d$  rispetto alla norma euclidea, si ha <sup>5</sup>

<sup>5</sup>Le (26) e (27) fanno riferimento a definizioni di deviazione e  $\nu$ -width leggermente estese rispetto a

$$\delta(\tilde{G}_c^d, A_\nu) \leq \frac{2c}{\sqrt{\nu}}, \quad (26)$$

dove  $\tilde{G}_c^d = \{\gamma|_{B_1^d} : \gamma \in G_c^d\}$ . Data una funzione  $\gamma : \mathbb{R}^d \mapsto \mathbb{R}$ ,  $\gamma|_{B_1^d}$  indica la restrizione di  $\gamma$  a  $B_1^d$ , cioè  $\gamma|_{B_1^d} : B_1^d \mapsto \mathbb{R}$  e  $\gamma|_{B_1^d}(\underline{x}) = \gamma(\underline{x})$ ,  $\forall \underline{x} \in B_1^d$ .

Per le reti OHL lineari, in [7] viene poi dimostrato che, misurando l'errore di approssimazione in norma  $\mathcal{L}_2$  in  $[0, 1]^d$ , risulta

$$d_\nu(\bar{G}_c^d) \geq b(\nu, c, d) \triangleq \frac{\kappa c}{d^d \sqrt{\nu}}, \quad (27)$$

dove  $\kappa \geq 1/(8\pi e^{\pi-1})$  e  $\bar{G}_c^d = \{\gamma|_{[0,1]^d} : \gamma \in G_c^d\}$ . Data una funzione  $\gamma : \mathbb{R}^d \mapsto \mathbb{R}$ ,  $\gamma|_{[0,1]^d}$  indica la restrizione di  $\gamma$  a  $[0, 1]^d$ , cioè  $\gamma|_{[0,1]^d} : [0, 1]^d \mapsto \mathbb{R}$  e  $\gamma|_{[0,1]^d}(\underline{x}) = \gamma(\underline{x})$ ,  $\forall \underline{x} \in [0, 1]^d$ .

L'interpretazione dei “bound” (26) e (27) è resa complessa dal fatto che, come notato in [7],  $c$  può non essere una costante, ma una funzione di  $d$ ; la dipendenza di  $c$  da  $d$  può inoltre variare a seconda di quali funzioni in  $G_c^d$  si considerano. Per fissare le idee, indichiamo con  $G_{cP}^d$  il sottoinsieme di  $G_c^d$  di funzioni per le quali lo scalare  $c$  cresce polinomialmente con  $d$  (esempi di classi di funzioni con una dipendenza polinomiale di  $c$  da  $d$  sono presentati in [7] e in [30]). In tal caso, dalla (26) segue che le reti OHL neurali sigmoidali beneficiano della proprietà di essere *reti a complessità polinomiale in  $\tilde{G}_{cP}^d$*  rispetto alla norma  $\mathcal{L}_2$ .

Vediamo ora quale rischio possono correre in  $\bar{G}_{cP}^d$  le reti OHL lineari. Supponiamo che  $c$  cresca anche solo linearmente con  $d$ . Per un dato valore di  $d_\nu(\bar{G}_c^d)$ , dalla (27) si può notare che il numero  $\nu$  di funzioni di base aumenta esponenzialmente con  $d$ , causando la maledizione della dimensionalità. Si noti anche che, nei casi in cui  $c$  non dipende da  $d$ , il fattore  $1/d$  fa decrescere a zero  $b(\nu, c, d)$ ; quindi tale “lower bound”, in corrispondenza di valori crescenti di  $d$ , diventa sempre meno significativo.

La figura 3 rappresenta graficamente la discussione di cui sopra, evidenziando l'andamento dei “bound”  $2c \left(\frac{1}{\nu}\right)^{1/2}$  e  $b(\nu, c, d)$  al crescere di  $\nu$ , per un dato valore di  $d$ . Nel caso particolare  $c = \alpha d$  ( $\alpha$  è una costante positiva) e per un medesimo valore assegnato di  $\delta(\tilde{G}_c^d, A_\nu)$  e  $d_\nu(\bar{G}_c^d)$ , la stessa figura descrive anche (nei due grafici più piccoli) l'andamento dell’“upper bound”  $\left(\frac{2\alpha}{\delta}\right)^2 d^2$  su  $\nu$  e del “lower bound”  $\left(\frac{\kappa\alpha}{d_\nu}\right)^d$  su  $\nu$  al crescere di  $d$ .

È importante notare che anche le reti OHL lineari possono beneficiare dell’“upper bound”  $O(1/\sqrt{\nu})$  nell'approssimare funzioni caratterizzate da opportune proprietà di regolarità (si vedano [7, p. 941] e [37, pp. 232-233]). Tali sono le funzioni con derivate parziali a quadrato integrabile fino all'ordine  $s$  (e quindi appartenenti agli spazi di

---

(22) e (23), in quanto si consente, come viene fatto in [7, p. 942], che  $\mathcal{S}$  non sia un sottoinsieme di  $\mathcal{H}$ . Le (22) e (23) sono state introdotte facendo riferimento al caso  $\mathcal{S} \subseteq \mathcal{H}$ , visto che, nell'ottimizzazione funzionale,  $\mathcal{S}$  “provviene” dal Problema P e quindi è un sottoinsieme di  $\mathcal{H}$ .

Figura 3: confronto tra reti OHL non lineari di tipo neurale sigmoidale e reti OHL lineari per l'insieme  $G_c^d$ .

Sobolev  $\mathcal{W}_2^s$ ), se  $s \geq \lfloor d/2 \rfloor + 2$ . È stato infatti dimostrato [7] che dette funzioni appartengono a  $G_c^d$  per opportuni valori di  $c$ . Pertanto, se  $c$  è tale che  $G_c^d \supset \mathcal{W}_2^s$ , le reti OHL di tipo neurale sigmoidale hanno prestazioni superiori a quelle delle reti OHL lineari almeno negli insiemi  $G_c^d \setminus \mathcal{W}_2^s$  (la cui “estensione” è comunque di difficile valutazione). Un confronto più generale tra le proprietà di densità delle reti OHL lineari e non lineari è stato recentemente sviluppato in [26].

È di fondamentale importanza osservare che ciascun tipo di rete a complessità polinomiale è dotata della proprietà di complessità polinomiale su insiemi di funzioni che soddisfano particolari assunzioni di regolarità, diverse tra loro. Questa problematica è stata affrontata in [17], ove si descrivono vari schemi OHL non lineari e i corrispondenti insiemi di funzioni, che tali schemi possono approssimare usando un numero di funzioni di base  $\nu$  caratterizzato da una crescita polinomiale con il numero  $d$  di variabili. Dall'analisi sviluppata in [17] risulta che, in generale, esistono intersezioni tra tali insiemi, ma non inclusioni. In altre parole, tra i vari schemi OHL considerati in [17] (che comprendono approssimatori diffusi nelle applicazioni), non esiste uno schema in grado di ottenere una crescita “moderata” di  $\nu$  in un insieme di funzioni, che includa tutti gli insiemi in cui una crescita dello stesso ordine venga ottenuta mediante altri schemi.

La tabella 1 (ottenuta modificando leggermente la tabella in [16, p. 255]) presenta varie classi di reti OHL non lineari caratterizzate da un decremento dell’“upper bound” sull'errore “worst-case” (22) di ordine  $O(1/\sqrt{\nu})$ . La funzione  $\sigma$  nella riga 2 è sigmoidale. Nella riga 3,  $|\cdot|_+$  indica la funzione “rampa”, cioè  $|z|_+ = 0$  se  $z < 0$ ,  $|z|_+ = z$  se  $z \geq 0$ . La funzione  $\zeta$  nella riga 4 deve verificare una

condizione tecnica (si veda [32]), che è soddisfatta, ad esempio, dalle multiquadriche generalizzate, dalle “thin plate spline” e dalla funzione gaussiana.  $K$  è un insieme compatto in  $\mathbb{R}^d$  e  $\mathcal{W}_p^s(K)$  è lo spazio di Sobolev di ordine  $s$  in norma  $\mathcal{L}_p(K)$ . Infine,  $\tilde{\gamma}$  indica la trasformata di Fourier di  $\gamma$  e  $\mathcal{B}_m$  è un potenziale di Bessel, cioè una funzione tale che  $\tilde{\mathcal{B}}_m(\omega) = 1/(1 + \|\omega\|_2^2)^{m/2}$ ,  $m > 0$ .

Spazio $\mathcal{H}$	Funzione di base parametrizzata $\varphi(\cdot, \cdot)$ in $\{A_\nu\}_{\nu=1}^\infty$	Insieme di funzioni	Rif.
$(\mathcal{L}_2(K), \mathbb{R})$	$\sin(\underline{x}^\top \underline{\alpha} + \theta)$	$\left\{ \gamma : \int_{\mathbb{R}^d}  \tilde{f}(\omega)  d\omega < +\infty \right\}$	[20]
$(\mathcal{L}_2(K), \mathbb{R})$	$\sigma(\underline{x}^\top \underline{\alpha} + \theta)$	$\left\{ \gamma : \int_{\mathbb{R}^d} \ \omega\ _2  \tilde{\gamma}(\omega)  d\omega < +\infty \right\}$	[7]
$(\mathcal{L}_2(K), \mathbb{R})$	$ \underline{x}^\top \underline{\alpha} + \theta _+$	$\left\{ \gamma : \int_{\mathbb{R}^d} \ \omega\ _2^2  \tilde{\gamma}(\omega)  d\omega < +\infty \right\}$	[10]
$(\mathcal{L}_p(K), \mathbb{R})$ $1 \leq p \leq \infty$	$\zeta(\underline{x}^\top \underline{\alpha} + \theta)$	$\mathcal{W}_p^{d/2}(K)$	[32]
$(\mathcal{L}_\infty(\mathbb{R}^d), \mathbb{R})$	$\mathcal{B}_m\left(\ \underline{x} - \underline{\alpha}\ _{\mathbb{R}^d}^2\right)$	$\mathcal{W}_1^{2m}(\mathbb{R}^d)$ , $2m > d$	[15]
$(\mathcal{L}_2(\mathbb{R}^d), \mathbb{R})$	$e^{-\ \underline{x} - \underline{\alpha}\ _{\mathbb{R}^d}^2 / \delta^2}$	$\mathcal{W}_1^{2m}(\mathbb{R}^d)$ , $2m > d$	[14]

Tabella 1: schemi OHL e insiemi di funzioni  $\gamma : \mathbb{R}^d \mapsto \mathbb{R}$  con velocità di decremento  $O(1/\sqrt{\nu})$  dell’“upper bound” (22) sull’errore di approssimazione, in vari spazi  $\mathcal{H}$ .

Gli schemi OHL presentati nella tabella 1 condividono una caratteristica fondamentale, che va evidenziata per non trarre in inganno il lettore circa la “sparizione” della maledizione della dimensionalità: il loro impiego non consente di eliminare tale grave inconveniente, ma di evitarlo considerando insiemi di funzioni che sono sempre più “condizionati” all’aumentare del numero di variabili. I benefici che derivano dall’utilizzo degli schemi OHL in questione sono legati alla possibilità di sfruttare detti “condizionamenti” in relazione agli insiemi di ammissibilità che di volta in volta si incontrano nei problemi di ottimizzazione funzionale.

Per quanto riguarda la possibilità di migliorare l’“upper bound” di ordine  $O(1/\sqrt{\nu})$ , si rimanda all’analisi sviluppata in [25, 29] e in alcuni riferimenti ivi contenuti.

## 7 La sequenza di problemi di programmazione non lineare ottenuta mediante reti OHL

In questa sezione utilizziamo una sequenza di reti OHL a complessità crescente per definire una sequenza di problemi che approssimano sempre più accuratamente il Problema P. A tal fine, invece di minimizzare il funzionale  $F$  rispetto a tutte le

funzioni decisionali ammissibili, cioè a tutti gli elementi di  $S$ , consideriamo, per ogni  $\nu \in \mathbb{Z}^+$ , la minimizzazione di  $F$  rispetto alle sole funzioni in  $S \cap A_\nu$ . In altre parole, tra tutte le funzioni ammissibili, consideriamo solo quelle esprimibili come reti OHL  $\underline{\gamma}_\nu$ . Poichè dette funzioni contengono  $N(\nu)$  parametri liberi, sostituendole nel funzionale  $F$  ed effettuando le operazioni da esso richieste (derivate, somme, integrali, ecc.), il funzionale diventa una funzione di  $N(\nu)$  variabili reali e cioè delle componenti del vettore  $\underline{w}_\nu$ . Indichiamo tale funzione come

$$F_\nu(\underline{w}_\nu) \triangleq F[\underline{\gamma}_\nu(\cdot, \underline{w}_\nu)].$$

La procedura appena descritta riduce la risoluzione del Problema P di ottimizzazione funzionale alla risoluzione di una sequenza  $P_1, P_2, \dots, P_\nu$  di problemi “approssimanti” di *programmazione* (in generale) *non lineare*. Ciascuno di tali problemi può essere formulato nel modo seguente.

**Problema  $P_\nu$ :**

$$\inf_{\underline{w}_\nu \in W_\nu} F_\nu(\underline{w}_\nu).$$

L'insieme  $W_\nu \subseteq \mathbb{R}^{N(\nu)}$  è l'insieme dei vettori ammissibili  $\underline{w}_\nu$  (si veda la (16)). È doveroso osservare che la determinazione analitica della funzione  $F_\nu(\underline{w}_\nu)$  può essere complessa. Tale complessità è accresciuta nei problemi di ottimizzazione funzionale di tipo stocastico, poichè il funzionale di costo contiene il valor medio rispetto a variabili aleatorie (si consideri il funzionale (2)). Diciamo subito, tuttavia, che in molti di tali problemi è possibile evitare il calcolo esplicito del valor medio, utilizzando le tecniche di approssimazione stocastica che saranno descritte nella sezione 10.

Ricordiamo ancora una volta la differenza tra l'ERIM e il metodo classico di Ritz. In quest'ultimo, le funzioni ammissibili sono date dalle combinazioni lineari (11) di funzioni di base fisse, che non contengono i vettori di parametri liberi  $\underline{k}_j$ . I vantaggi, che possono derivare dall'inserimento di tali vettori e quindi dall'utilizzo delle combinazioni lineari (13) di funzioni di base parametrizzate, sono basati sulle proprietà descritte nelle sezioni 5 e 6 e, soprattutto, su quelle che esporremo nelle sezioni 8 e 9.

## 8 Reti ottimizzanti

Una parte della teoria che presenteremo in questa sezione e in quella successiva non richiede che gli estremi inferiori nei Problemi P e  $P_\nu$  siano ottenuti in corrispondenza di una funzione ammissibile (che siano, cioè, dei minimi); altri sviluppi teorici invece richiedono tale ipotesi. Risulta quindi conveniente formulare sotto forma di assunzioni esplicite le richieste di esistenza dei minimi, in modo da farvi agevolmente riferimento quando necessario.

A1. Esiste una soluzione ottima  $\underline{\gamma}^\circ$  per il Problema P, cioè l'estremo inferiore è un minimo in corrispondenza di  $\underline{\gamma}^\circ$ .

A2. Per ogni  $\nu \in \mathbb{Z}^+$ , esiste una soluzione ottima  $\underline{w}_\nu^\circ$  per il Problema  $P_\nu$ , cioè l'estremo inferiore è un minimo in corrispondenza di  $\underline{w}_\nu^\circ$ .

Qui giunti, occorre precisare formalmente il significato dell'espressione (già riportata in precedenza) “i Problemi  $P_\nu$  approssimano il Problema P”. Quando le assunzioni A1 e A2 sono verificate, per semplificare le notazioni è utile introdurre le seguenti definizioni:

$$F^\circ \triangleq F(\underline{\gamma}^\circ) = \min_{\underline{\gamma} \in \mathcal{S}} F(\underline{\gamma})$$

e

$$F_\nu^\circ \triangleq F[\underline{\gamma}_\nu(\cdot, \underline{w}_\nu^\circ)] = \min_{\underline{w}_\nu \in W_\nu} F_\nu(\underline{w}_\nu).$$

Per brevità scriveremo anche  $\underline{\gamma}_\nu^\circ \triangleq \underline{\gamma}_\nu(\cdot, \underline{w}_\nu^\circ)$ . Una volta verificate le assunzioni A1 e A2, diremo che “i Problemi  $P_\nu$  approssimano il Problema P” quando sono verificate le due condizioni seguenti:

$$\lim_{\nu \rightarrow \infty} F(\underline{\gamma}_\nu^\circ) = F^\circ, \quad (28)$$

$$\lim_{\nu \rightarrow \infty} \|\underline{\gamma}_\nu^\circ - \underline{\gamma}^\circ\| = 0, \quad (29)$$

dove  $\|\cdot\|$  è la norma di  $\mathcal{H}$ .

È importante notare che, in generale, la convergenza espressa dalla (28) non implica la convergenza espressa dalla (29) e viceversa. Il concetto, che meglio descrive la convergenza della sequenza dei problemi approssimanti  $P_\nu$  al Problema P, è l'*epi-convergenza* (si veda, ad esempio, [4]). Se gli epigrafici associati ai Problemi  $P_\nu$  convergono all'epigrafico associato al Problema P, allora le sequenze  $\{F_\nu^\circ\}_{\nu=1}^\infty$  e  $\{\underline{\gamma}_\nu^\circ\}_{\nu=1}^\infty$  convergono rispettivamente a  $F^\circ$  e  $\underline{\gamma}^\circ$ . Si noti che la proprietà di densità, discussa nella sezione 5, garantisce soltanto che la soluzione ottima  $\underline{\gamma}^\circ$  del Problema P sia un punto di accumulazione di una sequenza  $\{\underline{\gamma}_\nu^*\}_{\nu=1}^\infty$ , ma non necessariamente della sequenza  $\{\underline{\gamma}_\nu^\circ\}_{\nu=1}^\infty$ . Nella maggior parte delle applicazioni del metodo di Ritz, dimostrare che  $\underline{\gamma}_\nu^\circ \xrightarrow{\nu \rightarrow \infty} \underline{\gamma}^\circ$  è piuttosto difficile.

Le condizioni (28) e (29) motivano l'introduzione della seguente classe di reti OHL.

**Definizione 8.1** *Data un'istanza del Problema P e supposto che le assunzioni A1 e A2 siano verificate, una sequenza  $\{A_\nu\}_{\nu=1}^\infty$  che soddisfa la (28) e la (29) è detta “schema P-ottimizzante”. In tal caso, le reti OHL di ciascun insieme  $A_\nu$  sono dette “reti P-ottimizzanti” o “P-ON”.*

È ora opportuno qualche commento sul ruolo giocato nell'ERIM dalle  $\mathcal{H}$ -DN e dalle P-ON. Ricordiamo innanzitutto che le reti OHL diventano  $\mathcal{H}$ -DN dopo essere state “personalizzate” allo spazio  $\mathcal{H}$ ; a tal fine, devono soddisfare la proprietà di densità nello spazio  $\mathcal{H}$ , nel senso specificato dalla definizione 5.1. Invece, per diventare una P-ON, una rete OHL deve essere “personalizzata” non solo allo spazio  $\mathcal{H}$ , ma anche all'insieme  $\mathcal{S}$  delle funzioni ammissibili e al funzionale  $F$ . Infatti, come sappiamo, un'istanza del Problema P è identificata dalla terna  $(H, S, F)$ ; pertanto, una P-ON deve essere “costruita” tenendo conto non solo della norma  $\|\cdot\|$ , ma del Problema P nella sua completezza. Quindi, una  $\mathcal{H}$ -DN non è necessariamente una P-ON.

D'altronde, una P-ON può non essere una  $\mathcal{H}$ -DN, dal momento che la densità di  $\bigcup_{\nu=1}^{\infty} A_{\nu}$  in  $\mathcal{H}$  non è in generale necessaria per soddisfare le condizioni (28) e (29).

Infatti, la (29) richiede soltanto che mediante  $\{\underline{\gamma}_{\nu}^{\circ}\}_{\nu=1}^{\infty}$  si possa approssimare con accuratezza arbitraria la soluzione ottima  $\underline{\gamma}^{\circ}$ . Quanto detto spiega la separazione tra approssimazione di funzioni e ottimizzazione funzionale nella figura 1. Comunque, visto che la soluzione ottima è incognita, di solito viene richiesto che l'impiego dello schema OHL  $\{A_{\nu}\}_{\nu=1}^{\infty}$  consenta di approssimare arbitrariamente bene una qualunque funzione di  $\mathcal{S}$ .

È evidente che le P-ON giocano un ruolo fondamentale nell'ERIM. Tuttavia, mentre esiste una grande quantità di risultati teorici sulle DN (varie classi di tali reti sono state ampiamente studiate nella teoria dell'approssimazione), la letteratura su quelle che abbiamo definito “P-ON” è estremamente ridotta [21, 27, 28]. Questo si spiega pensando al fatto che tali reti entrano in gioco solo nell'ERIM e cioè solo in un particolare metodo di risoluzione dei problemi di ottimizzazione funzionale.

## 9 Reti ottimizzanti a complessità polinomiale

Nella sezione precedente si è visto che l'impiego delle P-ON garantisce la convergenza delle soluzioni ottime dei Problemi  $P_1, P_2, \dots$  alla soluzione ottima del Problema P. Tuttavia, la convergenza delle sequenze  $\{F_{\nu}^{\circ}\}_{\nu=1}^{\infty}$  e  $\{\underline{\gamma}_{\nu}^{\circ}\}_{\nu=1}^{\infty}$  deve essere sufficientemente veloce, affinché l'ERIM renda computazionalmente trattabile il Problema P. L'accuratezza di approssimazione desiderata va cioè conseguita in corrispondenza di valori di  $\nu$  “sufficientemente piccoli”: le P-ON devono contenere un numero di parametri non troppo elevato, così da poter essere ottimizzate con un onere computazionale accettabile.

È chiaro che la velocità di convergenza delle sequenze  $\{F_{\nu}^{\circ}\}_{\nu=1}^{\infty}$  e  $\{\underline{\gamma}_{\nu}^{\circ}\}_{\nu=1}^{\infty}$  dipende dalle proprietà dello schema OHL  $\{A_{\nu}\}_{\nu=1}^{\infty}$ , dalla norma  $\|\cdot\|$  e dalle caratteristiche della terna  $(H, S, F)$  che identifica l'istanza del Problema P. Abbiamo visto che, nel metodo di Ritz, gli insiemi  $A_{\nu}$  sono costituiti da reti OHL lineari. In letteratura sono riportate poche applicazioni di tale metodo per la risoluzione di problemi in cui le funzioni decisionali dipendono da un elevato numero di variabili. Nei problemi di controllo ottimo deterministico a tempo continuo, ad esempio, le funzioni di controllo,

approssimate da reti OHL lineari, sono “ad anello aperto” e dipendono quindi dalla sola variabile tempo. Si parla in tal caso di “*control parameterization Ritz method*” [42].

Dal punto di vista teorico, stime della velocità con cui le sequenze  $\{\|\underline{\gamma}^\circ - \underline{\gamma}_\nu^\circ\|\}_{\nu=1}^\infty$  e  $\{F(\underline{\gamma}^\circ) - F_\nu^\circ\}_{\nu=1}^\infty$  convergono a zero sono relative al caso  $d = 1$  o forniscono “upper bound” che, pur essendo applicabili al caso multivariabile, non esplicitano la dipendenza di detta velocità di convergenza dal numero  $d$  di variabili (si vedano, ad esempio, [12, 19, 42] e i riferimenti in essi contenuti). I risultati teorici disponibili non sono quindi in grado di dire se il metodo di Ritz può risolvere efficientemente, utilizzando un numero ridotto  $\nu$  di funzioni di base, problemi di ottimizzazione funzionale con funzioni ammissibili che dipendono da un elevato numero di variabili.

D'altronde, i risultati sperimentali non sembrano indicare questa possibilità, a causa di limitazioni dovute essenzialmente alla maledizione della dimensionalità. Si noti che tali limitazioni sussistono anche per altre procedure classiche di ottimizzazione funzionale approssimata, sviluppate nell'ambito del calcolo delle variazioni, che usano reti OHL lineari e sono quindi correlate con il metodo di Ritz. Fra queste ricordiamo il metodo di Galerkin [22, pp. 258-262]. Anche le più recenti applicazioni di tale metodo a problemi di controllo ottimo dimostrano che la maledizione della dimensionalità è tuttora un punto dolente (si veda, ad esempio, [8], in cui si ricerca la soluzione di una classe di problemi di ottimizzazione funzionale risolvendo in modo approssimato la corrispondente equazione di Hamilton-Jacobi-Bellman).

Nella sezione 6 abbiamo visto che le reti OHL non lineari possono risultare preferibili alle reti OHL lineari per la capacità di approssimare funzioni utilizzando un numero ridotto di parametri. Ci chiediamo ora se un'analogia capacità si abbia ancora quando, invece di limitarci a individuare reti OHL in grado di approssimare arbitrariamente bene funzioni decisionali candidate ad essere ottime (senza nessuna garanzia di poterle realmente trovare), puntiamo direttamente a determinare una soluzione approssimata del Problema P (nel senso specificato dalle (28) e (29)), nei casi in cui le funzioni decisionali dipendano da molte variabili.

Analogamente a quanto fatto nella sezione 6, in cui abbiamo introdotto le reti a complessità polinomiale imponendo un “upper bound” sulla velocità con cui l'errore di approssimazione tende a zero, richiediamo ora che la velocità di convergenza sia dei valori subottimi  $F_\nu^\circ$  a  $F^\circ$  sia delle funzioni subottime  $\underline{\gamma}_\nu^\circ$  a  $\underline{\gamma}^\circ$  sia “sufficientemente veloce”. Più precisamente, richiediamo che:

1) la sequenza  $\{\underline{\gamma}_\nu^\circ\}_{\nu=1}^\infty$  sia tale che

$$F(\underline{\gamma}_\nu^\circ) - F^\circ \leq O(d^{p'}/\nu^{q'}) \quad (30)$$

dove  $p', q' \in \mathbb{R}^+$ ;

2) la sequenza  $\{\underline{\gamma}_\nu^\circ\}_{\nu=1}^\infty$  abbia  $\underline{\gamma}^\circ$  come funzione limite e

$$\|\underline{\gamma}_\nu^\circ - \underline{\gamma}^\circ\| \leq O(d^{p''}/\nu^{q''}), \quad (31)$$

dove  $p'', q'' \in \mathbb{R}^+$  e  $\|\cdot\|$  è la norma di  $\mathcal{H}$ .

Se valgono le (30) e (31), il numero  $\nu$  di funzioni di base, necessarie per garantire un'accuratezza di ottimizzazione  $\varepsilon$ , cresce al più come una potenza di  $d$ ; con ciò si scongiura il fenomeno della maledizione della dimensionalità. Definiamo quindi la seguente classe di reti OHL.

**Definizione 9.1** *Data un'istanza del Problema P e verificate le assunzioni A1 e A2, uno schema P-ottimizzante  $\{A_\nu\}_{\nu=1}^\infty$ , per il quale valgono le (30) e (31), è detto "schema P-ottimizzante a complessità polinomiale". In tal caso, le reti P-ottimizzanti di ciascun insieme  $A_\nu$  sono dette "reti P-ottimizzanti a complessità polinomiale" o "P-ON a complessità polinomiale".*

Risultati preliminari sulla costruzione di reti OHL che, per certe classi di funzionali  $F$  e per certi insiemi  $\mathcal{S}$  di funzioni ammissibili, diventano P-ON a complessità polinomiale, sono contenuti in [21, 27, 28].

Dalla definizione 9.1 segue che le P-ON a complessità polinomiale sono un sottoinsieme delle P-ON. Ciò è evidenziato nella parte destra della figura 1.

## 10 Approssimazione stocastica

Come abbiamo anticipato nella sezione 2 (si veda la (2)), i funzionali di costo dei problemi di ottimizzazione funzionale di tipo stocastico possono essere scritti nella forma  $F(\underline{\gamma}) \triangleq \underset{\underline{z}}{\mathbb{E}}\{J[\underline{\gamma}(\underline{x}), \underline{z}]\}$ . Sostituendo le reti OHL  $\underline{\gamma}_\nu$  nel funzionale  $F$ , come descritto nella sezione 7, si ottiene la funzione di costo  $F_\nu(\underline{w}_\nu) \triangleq \underset{\underline{z}}{\mathbb{E}}[J_\nu(\underline{w}_\nu, \underline{z})]$ . Per ogni intero positivo  $\nu$ , si può formulare una versione "stocastica" del Problema  $P_\nu$  ( $\nu = 1, \dots$ ). Per evidenziare la presenza del valore atteso nella funzione di costo  $F_\nu(\underline{w}_\nu)$ , cambiamo lievemente le notazioni usando  $\mathcal{P}_\nu$  invece di  $P_\nu$ .

**Problema  $\mathcal{P}_\nu$  :**

$$\inf_{\underline{w}_\nu \in W_\nu} \underset{\underline{z}}{\mathbb{E}}[J_\nu(\underline{w}_\nu, \underline{z})].$$

Introduciamo ora un'assunzione analoga alla A2.

A2'. Per ogni  $\nu \in \mathbb{Z}^+$  esiste una soluzione ottima  $\underline{w}_\nu^\circ$  per il Problema  $\mathcal{P}_\nu$ .

Nel seguito supporremo che le assunzioni A1 e A2' siano verificate. Per risolvere il Problema  $\mathcal{P}_\nu$  scegliamo algoritmi di discesa basati sul metodo del gradiente, essenzialmente per la loro semplicità. Vedremo infatti che questa scelta ci porterà in modo del tutto naturale a introdurre il concetto di "approssimazione stocastica". Inoltre, per non appesantire eccessivamente le notazioni, utilizzeremo tecniche di risoluzione

che impiegano funzioni di penalità; tali tecniche corrispondono a un'interpretazione “soft” dei vincoli sul vettore  $\underline{w}_\nu$ , vincoli che definiscono l'insieme  $W_\nu \subseteq \mathbb{R}^{N(\nu)}$ . Questo approccio è accettabile in molte situazioni applicative e, come risulterà chiaro poco oltre, non rappresenta un limite concettuale per la nostra trattazione. Supporremo quindi  $W_\nu = \mathbb{R}^{N(\nu)}$ , riducendo in tal modo il Problema  $\mathcal{P}_\nu$  a un problema di programmazione non lineare non vincolata.

L'impiego di algoritmi basati sull'uso del gradiente richiede ovviamente la seguente ipotesi.

A3.  $J_\nu(\underline{w}_\nu, \underline{z})$  è una funzione di classe  $\mathcal{C}^1$  rispetto a  $\underline{w}_\nu$  per ogni  $\underline{z}$ .

Quando l'assunzione A3 è verificata, alcune ipotesi di regolarità aggiuntive consentono di concludere che anche  $F_\nu(\underline{w}_\nu)$  è una funzione  $\mathcal{C}^1$  (si veda, ad esempio, [41]).

A causa del contesto molto generale in cui il Problema  $\mathcal{P}_\nu$  è formulato, di solito non si è in grado di esprimere analiticamente il gradiente  $\nabla_{\underline{w}_\nu} \mathbb{E}_{\underline{z}}[J_\nu(\underline{w}_\nu, \underline{z})]$ . Ciò è dovuto essenzialmente alla necessità di calcolare, in generale per via numerica, l'integrale multiplo associato alla determinazione del valor medio  $\mathbb{E}_{\underline{z}}[J_\nu(\underline{w}_\nu, \underline{z})]$  (la dimensione di  $\underline{z}$  può essere molto elevata). Come si vedrà nell'esempio presentato nella sezione successiva, il gradiente  $\nabla_{\underline{w}_\nu} J_\nu(\underline{w}_\nu, \underline{z})$  può essere invece calcolato in modo relativamente semplice.

Questi due fatti indicano con evidenza l'opportunità di utilizzare una tecnica di *approssimazione stocastica*. L'idea base di tale tecnica consiste infatti nell'impiegare la realizzazione del gradiente  $\nabla_{\underline{w}_\nu} J_\nu(\underline{w}_\nu, \underline{z})$  invece di  $\nabla_{\underline{w}_\nu} \mathbb{E}_{\underline{z}}[J_\nu(\underline{w}_\nu, \underline{z})]$ . In sintesi, per l'ottimizzazione del vettore  $\underline{w}_\nu$  usiamo l'algoritmo

$$\underline{w}_\nu(k+1) = \underline{w}_\nu(k) - \alpha_k \nabla_{\underline{w}_\nu} J_\nu[\underline{w}_\nu(k), \underline{z}(k)], \quad k = 0, 1, \dots, \quad (32)$$

dove la sequenza  $\{\underline{z}(k)\}_{k=1}^\infty$  è generata aleatoriamente usando la densità di probabilità di  $\underline{z}$  e  $\alpha_k$  è un passo di discesa che decresce in modo opportuno. L'algoritmo (32) è uno degli esempi più semplici e diffusi della famiglia degli algoritmi di approssimazione stocastica.

Facciamo ora alcune osservazioni sull'algoritmo (32). Ciò che ha consentito di scriverlo è l'aver supposto  $W_\nu = \mathbb{R}^{N(\nu)}$  e cioè che i vincoli siano assenti o “trasferibili” nella funzione di costo. Esistono tuttavia tecniche di approssimazione stocastica che consentono di soddisfare “in modo esatto” eventuali vincoli sul vettore  $\underline{w}_\nu$  (si veda, ad esempio, [24]).

Per quanto riguarda la convergenza (in probabilità) dell'algoritmo (32), condizioni sufficienti per ottenerla sono riportate, ad esempio, in [11] e in [24]. Alcune di tali condizioni fanno riferimento all'andamento della sequenza  $\{\alpha_k\}_{k=1}^\infty$ . Solitamente si richiede che

$$\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \quad (33)$$

Nell'esempio presentato nella sezione 12 si è scelto  $\alpha_k = c_1/(c_2 + k)$ ,  $c_1, c_2 > 0$ , che soddisfa le (33). Le altre condizioni hanno a che fare con la forma della superficie della funzione di costo  $F_\nu(\underline{w}_\nu)$  e in generale sono ben più difficili da verificare, a causa della complessità di tale superficie.

Esistono numerose tecniche per accelerare la convergenza dell'algoritmo (32); si veda, ad esempio, [24]. Qualora vengano usate reti neurali, non vanno poi dimenticati i numerosi algoritmi esistenti per l'ottimizzazione dei parametri di tali reti (si rinvia, tra gli altri, a [2], [18] e ai riferimenti in essi riportati). Va infine ricordato che la pratica sperimentale dimostra assai spesso la notevole efficacia di alcuni algoritmi euristici, per i quali non è possibile, in generale, fornire condizioni di convergenza.

## 11 Applicazione dell'ERIM a un problema di controllo ottimo stocastico

In questa sezione applichiamo l'ERIM alla risoluzione approssimata del Problema C introdotto nella sezione 2. Per semplificare le notazioni, facciamo l'ipotesi che tale problema sia formulato in un contesto stazionario. Per far sì che l'assunzione A3 sia soddisfatta, supponiamo che le funzioni  $f, g, h$  e  $h_T$  e le loro derivate prime siano continue. Seguendo le linee generali dell'ERIM, imponiamo a ciascuna funzione di controllo  $\underline{\gamma}_t(\underline{I}_t)$  di assumere la struttura di una rete OHL non lineare. Scriviamo quindi (si veda la (13)):

$$\underline{w}_t = \underline{\gamma}_{t, \nu_t}(\underline{I}_t, \underline{w}_{t, \nu_t}) = \text{col} \left[ \sum_{i=1}^{\nu_t} c_{ijt} \varphi_t(\underline{I}_t, \underline{k}_{it}) : j = 1, \dots, m \right], \quad t = 0, 1, \dots, T-1, \quad (34)$$

dove  $\underline{k}_{it} \in \mathbb{R}^{k_i}$  e  $\underline{w}_{t, \nu_t} \triangleq \text{col}(c_{ijt}, \underline{k}_{it} : i = 1, \dots, \nu_t, j = 1, \dots, m)$ . Definiamo inoltre il vettore  $\underline{\nu} \triangleq \text{col}(\nu_t : t = 0, 1, \dots, T-1)$  e gli insiemi  $A_{\underline{\nu}} \triangleq \{\underline{\gamma}_{t, \nu_t}(\underline{I}_t, \underline{w}_{t, \nu_t}) : \underline{w}_{t, \nu_t} \in \mathbb{R}^{\nu_t(k_i+m)}, t = 0, 1, \dots, T-1\}$ . Ogni insieme  $A_{\underline{\nu}}$  è quindi costituito da sequenze di  $T$  reti OHL.

Organizzando opportunamente la "crescita" degli insiemi  $A_{\underline{\nu}}$  (che potrebbe essere evidenziata scrivendo  $A_{\underline{\nu}_1} \subseteq A_{\underline{\nu}_2}, \dots$ ), si può facilmente verificare [45] la possibilità di ottenere una struttura a inclusione del tipo

$$A_{\nu^\circ} \subseteq A_{\nu^\circ+1} \subseteq \dots \subseteq A_{\nu^\circ+i} \subseteq \dots, \quad (35)$$

dove i pedici scalari  $\nu^\circ + i, i = 0, 1, \dots$ , sostituiscono i pedici vettoriali  $\underline{\nu}_i$ . Una sequenza di reti appartenente a  $A_{\nu^\circ+i+1}$  contiene, in una sola delle sue  $T$  reti OHL, una sola funzione di base in più rispetto alla sequenza di  $T$  reti appartenenti all'insieme  $A_{\nu^\circ+i}$ . Il pedice  $\nu^\circ$  rappresenta il numero di funzioni di base contenute nella più piccola catena di  $T$  reti. Visto che tale catena è composta da reti con un'unica funzione

di base, risulta  $\nu^\circ = T$ . Si noti che la struttura a inclusione (35) è riconducibile alla (14).

Sostituendo le (34) nelle (3), (5) e (6), le (3) nella (5) ed esprimendo tutti i vettori, che costituiscono il vettore  $\underline{I}_t$ , in funzione delle variabili aleatorie “primitive”  $\underline{x}_0$ ,  $\underline{\xi} \triangleq \text{col}(\underline{\xi}_0, \dots, \underline{\xi}_{T-1})$  e  $\underline{\eta} \triangleq (\underline{\eta}_0, \dots, \underline{\eta}_{T-1})$ , il funzionale di costo (5) diventa una funzione  $J_\nu(\underline{w}_\nu, \underline{x}_0, \underline{\xi}, \underline{\eta})$ , dove  $\underline{w}_\nu \triangleq \text{col}(\underline{w}_{t,\nu_t} : t = 0, 1, \dots, T-1)$ . La funzione di costo  $F_\nu(\underline{w}_\nu)$  è data quindi da

$$F_\nu(\underline{w}_\nu) \triangleq \mathbb{E}_{\underline{x}_0, \underline{\xi}, \underline{\eta}} [J_\nu(\underline{w}_\nu, \underline{x}_0, \underline{\xi}, \underline{\eta})].$$

In tal modo, il Problema C di ottimizzazione funzionale è ridotto ad una sequenza di problemi di programmazione non lineare della forma del Problema  $\mathcal{P}_\nu$ , con  $\underline{z} \triangleq \text{col}(\underline{x}_0, \underline{\xi}, \underline{\eta})$ . Non essendo presenti vincoli sulle variabili del Problema C, neppure il vettore dei parametri  $\underline{w}_\nu$  è vincolato. Per ogni  $\nu \geq \nu_0$ , ciascun Problema  $\mathcal{P}_\nu$  assume dunque la forma seguente.

**Problema  $\mathcal{C}_\nu$ :**

$$\inf_{\underline{w}_\nu} \mathbb{E}_{\underline{x}_0, \underline{\xi}, \underline{\eta}} [J_\nu(\underline{w}_\nu, \underline{x}_0, \underline{\xi}, \underline{\eta})].$$

Al fine di non appesantire eccessivamente le notazioni, nel seguito ometteremo il pedice  $\nu_t$  nelle funzioni  $\gamma_{t,\nu_t}$  e nei vettori  $\underline{w}_{t,\nu_t}$ ; ometteremo inoltre il pedice  $\nu$  nel costo  $J_\nu$  e nel suo argomento  $\underline{w}_\nu$ . Per risolvere il Problema  $\mathcal{C}_\nu$  mediante l'algoritmo (32), a ciascuna iterazione è necessario determinare le componenti del gradiente  $\nabla_{\underline{w}} J[\underline{w}(k), \underline{x}_0(k), \underline{\xi}(k), \underline{\eta}(k)]$  e cioè le derivate parziali

$$\frac{\partial}{\partial w_t^l} J[\underline{w}(k), \underline{x}_0(k), \underline{\xi}(k), \underline{\eta}(k)], \quad t = 0, 1, \dots, T-1, \quad l = 1, \dots, N(\nu_t),$$

dove  $w_t^l$  è la componente  $l$ -esima del vettore  $\underline{w}_t$  e  $N(\nu_t) = \dim(\underline{w}_t)$ . Possiamo scrivere

$$\frac{\partial J}{\partial w_t^l} = \frac{\partial J}{\partial \underline{u}_t} \frac{\partial \gamma_t(\underline{I}_t, \underline{w}_t)}{\partial w_t^l}. \quad (36)$$

Con alcuni semplici passaggi [35] si ottiene

$$\frac{\partial J}{\partial \underline{u}_t} = \frac{\partial}{\partial \underline{u}_t} h(\underline{x}_t, \underline{u}_t) + \underline{\lambda}_{t+1}^\top \frac{\partial}{\partial \underline{u}_t} f(\underline{x}_t, \underline{u}_t, \underline{\xi}_t) + \sum_{j=t+1}^{T-1} \frac{\partial J}{\partial \underline{I}_j^{u_t}}, \quad t = 0, 1, \dots, T-1, \quad (37)$$

dove  $\underline{\lambda}_t^\top \triangleq \frac{\partial J}{\partial \underline{x}_t}$  e  $\underline{I}_j^{u_t}$  rappresenta l'ingresso, corrispondente a  $\underline{u}_t$ , della  $j$ -esima rete OHL ( $\underline{I}_j^{u_t} = \underline{u}_t$  è cioè una parte del vettore  $\underline{I}_j$ ). Analogamente, indichiamo con  $\underline{I}_j^{y_t}$

l'ingresso, corrispondente a  $\underline{y}_t$ , della  $j$ -esima rete. Quindi,  $\underline{\lambda}_t$  può essere calcolato come segue.

$$\underline{\lambda}_t^\top = \frac{\partial}{\partial \underline{x}_t} h(\underline{x}_t, \underline{u}_t) + \underline{\lambda}_{t+1}^\top \frac{\partial}{\partial \underline{x}_t} f(\underline{x}_t, \underline{u}_t, \underline{\xi}_t) + \left[ \sum_{j=t+1}^{T-1} \frac{\partial J}{\partial \underline{I}_j^{y_t}} \right] \frac{\partial}{\partial \underline{x}_t} g(\underline{x}_t, \underline{\eta}_t),$$

$$t = 0, 1, \dots, T-1$$
(38)

$$\underline{\lambda}_T^\top = \frac{\partial}{\partial \underline{x}_T} h_T(\underline{x}_T).$$

Nella (38) si può riconoscere la forma classica dell'equazione aggiunta del controllo ottimo a  $T$  stadi, con l'inserimento di un termine (il terzo) per tener conto della parametrizzazione delle funzioni di controllo ad anello chiuso. Si noti anche che la struttura della (38) non dipende dal tipo di rete OHL. Il particolare tipo di rete OHL entra in gioco quando vengono esplicitati il vettore colonna  $\frac{\partial \gamma_t(\underline{I}_t, \underline{w}_t)}{\partial w_t^l}$  nelle (36) e i

vettori riga  $\sum_{j=t+1}^{T-1} \frac{\partial J}{\partial \underline{I}_j^{u_t}}$  e  $\sum_{j=t+1}^{T-1} \frac{\partial J}{\partial \underline{I}_j^{y_t}}$  rispettivamente nelle (37) e (38).

Sotto blande ipotesi, le  $T$  reti OHL assumono il ruolo di DN in opportuni spazi  $\mathcal{H}_t$ . Supponiamo, ad esempio, che i vettori  $\underline{x}_0$ ,  $\underline{\xi}_t$  ed  $\underline{\eta}_t$  appartengano ad insiemi compatti. Poichè  $f$  e  $g$  sono funzioni continue, è immediato verificare che i vettori  $\underline{I}_t$  appartengono ad insiemi compatti  $\mathcal{I}_t$ . Sia poi noto a priori che le funzioni ottime  $\gamma_t^\circ(\underline{I}_t)$  sono continue in  $\mathcal{I}_t$ . È allora utile che le reti OHL, candidate ad approssimare arbitrariamente bene le funzioni ottime  $\gamma_t^\circ(\underline{I}_t)$ , siano reti  $\mathcal{C}_t$ -dense, dove  $\mathcal{C}_t(\mathcal{I}_t, \mathbb{R}^m) = (C(\mathcal{I}_t, \mathbb{R}^m), \|\cdot\|_\infty)$ ,  $t = 0, 1, \dots, T-1$  e  $\|\cdot\|_\infty$  è la norma (21). La scelta di reti OHL basate su costruzioni di tipo “ridge” (18) o radiale (20) risulta opportuna, in quanto tali reti sono  $\mathcal{C}_t$ -dense. Per approfondimenti su queste problematiche si veda, ad esempio, [35].

Circa la complessità delle reti OHL, ci si può ricondurre a quanto detto nella sezione 6 e, in particolare, alla definizione 6.3. Supponiamo, ad esempio, che ciascuna delle componenti delle funzioni ottime di controllo  $\gamma_t^\circ(\underline{I}_t)$  appartenga ad uno spazio  $G_{c_t}^{d_t}$  definito dalla (25), dove  $d_t = \dim(\underline{I}_t)$  e  $t = 0, 1, \dots, T-1$ . Supponiamo inoltre che le costanti  $c_t$  non dipendano da  $d_t$  o che crescano al più polinomialmente con  $d_t$ . Le reti OHL neurali sigmoidali (18) risultano idonee ad approssimare le componenti di  $\gamma_t^\circ(\underline{I}_t)$ . Infatti, per quanto detto nella sezione 6, fissato un certo errore di approssimazione in norma  $\mathcal{L}_2$ , il numero  $\nu_t$  di funzioni di base di ciascuna rete cresce non più che polinomialmente al crescere di  $d_t$ . Secondo la terminologia introdotta, le reti OHL neurali sigmoidali  $\gamma_t(\underline{I}_t, \underline{w}_t)$  sono quindi reti a complessità polinomiale in  $\tilde{G}_{c_t}^{d_t}$  rispetto alla norma  $\mathcal{L}_2$ .

È infine da notare che l'applicazione dell'ERIM al Problema C consente di non calcolare le densità di probabilità  $p(\underline{x}_t | \underline{I}_t)$  richieste dalla programmazione dinamica.

Rimane tuttavia la necessità di conservare in memoria i vettori  $\underline{I}_t$ , le cui dimensioni  $d_t$  aumentano con il tempo, e di dover ottimizzare reti OHL di complessità sempre maggiore. Ciò diventa praticamente impossibile quando  $T$  è molto grande o tende addirittura all'infinito. Si impone in tal caso un'ulteriore approssimazione: 1) le funzioni di controllo vengono fatte dipendere soltanto dai più recenti controlli e misure contenuti in  $\underline{I}_t$ , 2) si sostituisce il Problema C con una sequenza di problemi di controllo ottimo "a orizzonte mobile", in cui viene ottimizzato un numero costante di reti OHL, eguale al numero di stadi temporali di detto orizzonte. Per un approfondimento di questa tecnica di approssimazione si rinvia a [35].

## 12 Un esempio applicativo: controllo ottimo del traffico autostradale

Consideriamo un problema di controllo ottimo del traffico autostradale, riconducibile al Problema C. L'esempio è di rilevante interesse nell'ingegneria dei trasporti, non ammette una risoluzione analitica ed è difficilmente risolvibile con tecniche approssimate tradizionali.

Si vuole controllare un tratto di autostrada della lunghezza di 30 km, in cui il flusso veicolare è descritto mediante il noto modello macroscopico di Payne [33]:

$$\begin{aligned}
v_{i,t+1} = & v_{it} + \frac{\delta_T}{\tau} \left\{ V_f b_{it} \left[ 1 - (\rho_{it}/\rho_{\max})^{m(3-2b_{it})} \right]^l - v_{it} \right\} \\
& + \frac{\delta_T}{\Delta_i} v_{it} (v_{i-1,t} - v_{it}) + \frac{\nu \delta_T (\rho_{i+1,t} - \rho_{it})}{\tau \Delta_i (\rho_{it} + \chi)} \\
& - \delta_{\text{on}} \frac{\delta_T}{\Delta_i} v_{it} \frac{r_{it}}{\rho_{it} + \chi},
\end{aligned} \tag{39}$$

$$\begin{aligned}
\rho_{i,t+1} = & \rho_{it} + \frac{\delta_T}{\Delta_i} [\alpha (1 - \gamma_i) \rho_{i-1,t} v_{i-1,t} \\
& + (1 - 2\alpha + \gamma_i \alpha - \gamma_i) \rho_{it} v_{it} - (1 - \alpha) \rho_{i+1,t} v_{i+1,t} + r_{it}], \\
& t = 0, 1, \dots, T - 1; \quad i = 1, \dots, D.
\end{aligned} \tag{40}$$

Il tratto autostradale considerato è suddiviso in  $D = 30$  segmenti di lunghezza  $\Delta = 1$  km ciascuno. In ciascun segmento  $i$ , il traffico è descritto da due variabili:  $v_{it}$  rappresenta la velocità media del traffico e  $\rho_{it}$  la densità media del traffico all'istante  $t$ . Si suppone di controllare il traffico su un arco temporale di 15 minuti generando le variabili di controllo ogni  $\delta_T = 15$  secondi. Il numero di stadi decisionali è dunque  $T = 60$ . I valori dei parametri che compaiono nelle (39) e (40) sono riportati in [33] e [45].

Per completare il modello, occorre descrivere la dinamica delle code di veicoli sulle rampe d'ingresso. Sia dunque  $l_{it}$  il numero di veicoli in tali code; supponiamo

che vi siano 5 rampe d'ingresso (e di uscita) poste nei segmenti 3, 9, 15, 21 e 27. I pedici corrispondenti a detti segmenti costituiscono l'insieme  $\mathcal{I}^r \subset \{1, \dots, 30\}$ . Siano inoltre  $d_{it}$  i flussi aleatori di veicoli in ingresso alle rampe ( $i \in \mathcal{I}^r$ ). Tali rampe possono dunque essere modellate dalle seguenti relazioni:

$$l_{i,t+1} = l_{it} + \delta_T (d_{it} - r_{it}), \quad t = 0, 1, \dots, T-1, i \in \mathcal{I}^r. \quad (41)$$

Le variabili  $r_{it}$  rappresentano i flussi di veicoli che accedono all'autostrada; tali flussi possono essere "modulati" mediante segnali semaforici posti sulle rampe d'ingresso. È inoltre possibile imporre le velocità massime  $b_{it}$  sui segmenti dell'autostrada mediante pannelli a segnaletica variabile, posti nei segmenti 1, 7, 13, 19 e 25. Ciascun segnale indica la velocità massima da osservare su 6 segmenti successivi.

In sintesi, il tratto autostradale può essere modellato mediante il sistema dinamico (3) (ora stazionario), dove  $r_{it}$  e  $b_{it}$  giocano il ruolo di variabili di controllo e  $v_{it}$ ,  $\rho_{it}$  e  $l_{it}$  sono variabili di stato. Si può quindi porre  $\underline{x}_t \triangleq \text{col}(v_{it}, i = 1, \dots, D; \rho_{it}, i = 1, \dots, D; l_{it}, i \in \mathcal{I}^r)$ ,  $\underline{u}_t \triangleq \text{col}(r_{it}, i \in \mathcal{I}^r; b_{it}, i = 1, \dots, D)$  e  $\underline{\xi}_t \triangleq \text{col}(d_{it}, i \in \mathcal{I}^r)$ . Si assuma che i flussi  $d_{it}$  siano mutuamente indipendenti e uniformemente distribuiti su intervalli opportuni. La figura 4 schematizza il tratto di autostrada con il sistema di controllo. Si faccia poi l'ipotesi che tutte le variabili di stato siano misurabili, ma che le misure  $y_{tj}$  siano affette da disturbi additivi. Sia cioè  $y_{tj} = x_{tj} + \eta_{tj}$ ,  $j = 1, \dots, n$ . Si supponga che anche i disturbi  $\eta_{tj}$  siano indipendenti tra loro e da tutte le altre variabili aleatorie e distribuiti uniformemente su intervalli opportuni. Aggregando vettorialmente le variabili, l'equazione di misura (4) risulta essere  $\underline{y}_t = \underline{x}_t + \underline{\eta}_t$ .

Figura 4: sistema di controllo del traffico autostradale.

Il funzionale di costo è costituito dal tempo trascorso complessivamente dai veicoli

sull'autostrada e nelle code; espresso in secondi, assume quindi la forma

$$J = \delta_T \sum_{t=1}^T \left( \Delta \sum_{i=1}^D \rho_{it} + \sum_{i \in \mathcal{I}_r} l_{it} \right). \quad (42)$$

Il modello è completato da numerosi vincoli, dei quali si può tener conto in modo approssimato inglobandoli nel costo (42) mediante opportune funzioni di penalità. Le espressioni analitiche dei vincoli sono riportate in [45]. In definitiva, siamo in presenza del seguente problema di controllo ottimo stocastico.

**Problema C<sub>TRAFF</sub>.** *Determinare la legge di controllo ottima  $\underline{u}_t^\circ = \underline{\mu}_t^\circ(\underline{I}_t)$ ,  $t = 0, 1, \dots, T - 1$ , che minimizza il valore atteso del costo (42) in corrispondenza di una condizione iniziale di traffico  $\underline{x}_0$ , considerata come un vettore aleatorio uniformemente distribuito su un insieme  $X_0$  di condizioni iniziali.*

A causa dell'elevato numero di stadi decisionali, il problema viene risolto utilizzando la tecnica del controllo ad orizzonte mobile, secondo quanto accennato alla fine della sezione 11. Ciò significa che, allo stadio  $t$ , il sistema di controllo genera il vettore  $\underline{u}_t$  utilizzando soltanto le misure  $\underline{y}_{t-N'}, \dots, \underline{y}_t$  (nell'esempio in oggetto si elabora soltanto la misura  $\underline{y}_t$ ) ed un ulteriore vettore che conserva una "traccia" delle misure acquisite prima dello stadio  $t - N'$  (per ulteriori dettagli si rinvia a [35]). Le funzioni di controllo  $\underline{u}_t = \underline{\mu}_t(\underline{I}_t)$  sono vincolate ad assumere la forma di reti neurali "feedforward" sigmoidali. Tali reti vengono ottimizzate mediante l'algoritmo (32), ponendo  $\alpha_k = c_1/(c_2 + k)$ ,  $c_1 = 1$  e  $c_2 = 10^8$ . Come si è detto nella sezione 10, l'ottimizzazione delle reti viene effettuata estraendo randomicamente "pattern" di addestramento costituiti da vari possibili stati iniziali e da una molteplicità di sequenze dei disturbi.

Un esempio simile di controllo del traffico autostradale è descritto in [31], dove tuttavia il problema viene risolto sotto ipotesi molto più semplici, supponendo assenti i disturbi di misura e sostituendo, su un orizzonte temporale mobile, i disturbi  $\underline{\xi}_t$  con i rispettivi valori medi. La tecnica proposta in [31] può dunque risolvere il problema di controllo ottimo (reso di fatto deterministico) mediante una tecnica di programmazione non lineare applicata *in linea* stadio dopo stadio. La tecnica prende il nome di *certainty-equivalent open-loop feedback* (CEOLF) *control*.

L'ERIM e la tecnica CEOLF sono stati posti a confronto nel caso in cui le variabili di stato siano perfettamente misurabili, simulando una forte congestione di traffico all'istante  $t = 0$  nel segmento 11. Nelle figure 5a e 5b sono riportate le velocità  $v_{it}$  e le densità  $\rho_{it}$  in corrispondenza dei vari segmenti autostradali e degli stadi temporali. Le superfici che descrivono  $\rho_{it}$  e  $v_{it}$ , ottenute con l'ERIM, sono praticamente coincidenti con le rispettive superfici ottenute mediante la tecnica CEOLF. Si può constatare la rapida azione di decongestionamento esercitata dal sistema di controllo. Si noti che l'applicabilità del metodo CEOLF è limitata dall'ipotesi, poco realistica, di poter misurare senza disturbi tutte le variabili di stato. Inoltre, l'onere computazionale in

Figura 5: andamento della densità di traffico  $\rho_{it}$  e della velocità media  $v_{it}$  durante il controllo di una forte congestione di traffico nel segmento autostradale 11, sotto l'azione della legge di controllo neurale ottima. a,b: ingressi aleatori nel sistema dinamico e misure esatte del vettore di stato. c,d: ingressi aleatori nel sistema dinamico e misure rumorose del vettore di stato.

linea richiesto dalla tecnica CEOLF può risultare insostenibile. Al contrario, l'ERIM calcola *fuori linea* (quindi senza vincoli stringenti di tempo) le funzioni approssimate di controllo ottimo. Per tale ragione non abbiamo esplorato la possibilità di utilizzare varianti più veloci dell'algoritmo (32), che ha raggiunto la convergenza dopo circa  $3 \cdot 10^5$  iterazioni. Per contro le reti neurali di controllo, una volta *ottimizzate (in media) per ogni possibile condizione iniziale del vettore di stato*, possono generare pressochè *istantaneamente* i vettori di controllo sulla base delle misure via via acquisite.

Il Problema  $C_{\text{TRAFF}}$  è stato poi risolto, applicando l'ERIM, nel caso più generale descritto in questa sezione, supponendo cioè che le misure dello stato siano affette da rumore, secondo il modello  $\underline{y}_t = \underline{x}_t + \underline{\eta}_t$ . Come si è detto, tale caso non è risolvibile mediante la tecnica CEOLF proposta in [31]. La convergenza dell'algoritmo (32) è stata raggiunta dopo  $4 \cdot 10^5$  iterazioni. Le superfici che descrivono le traiettorie delle variabili di stato  $\rho_{it}$  e  $v_{it}$  ottenute con l'ERIM sono riportate nelle figure 5c e 5d.

Desideriamo infine sottolineare che nell'esempio considerato sono presenti ben 65 variabili di stato! Pur supponendo che le misure di dette variabili non siano affette da rumore, un numero così elevato di componenti del vettore di stato esclude la possibilità di impiegare tecniche di programmazione dinamica basate sulla tradizionale discretizzazione dello spazio di stato in reticoli regolari.

## 13 Conclusioni

Il Metodo di Ritz Esteso si è rivelato uno strumento potente per la risoluzione approssimata di problemi di ottimizzazione funzionale. Il metodo manifesta la propria efficacia in contesti caratterizzati da elevata complessità. Ad esempio, situazioni di questo tipo sono le seguenti: le ipotesi LQG non sono soddisfatte; l'elevato numero delle variabili argomento delle funzioni ammissibili (ed il conseguente pericolo della maledizione della dimensionalità) rende problematico l'impiego di strumenti tradizionali quali la programmazione dinamica; vi è una "squadra" composta da più agenti decisionali cooperanti alla minimizzazione di un medesimo funzionale di costo. In quest'ultimo caso, studiato dalla "team theory", è noto che neppure il verificarsi delle ipotesi LQG può essere sufficiente per la determinazione analitica della soluzione ottima (si veda, ad esempio, [5]). Quanto detto è comprovato da numerosi risultati sperimentali, ottenuti su un ampio spettro di problemi applicativi.

Con riferimento al numero di variabili delle funzioni ammissibili, si sono ottenuti per l'ERIM risultati preliminari che trasferiscono all'ottimizzazione funzionale varie proprietà che, con ben più ampio respiro, sono state dimostrate nel settore dell'approssimazione di funzioni. Si fa riferimento alla possibilità di evitare la crescita esponenziale del numero di parametri da ottimizzare nelle reti OHL al crescere del numero di variabili delle funzioni decisionali. Utilizzando la terminologia introdotta in questa trattazione, la letteratura sulla teoria dell'approssimazione di funzioni ha evidenziato che tali reti possono diventare non solo *reti dense in  $\mathcal{H}$* , ma anche *reti a complessità polinomiale in  $\mathcal{S}$* . Analogamente, i risultati teorici preliminari da noi ottenuti av-

valorano la congettura che una scelta opportuna delle reti OHL, ben “calibrata” sul problema di ottimizzazione funzionale considerato, possa rendere dette reti innanzi tutto *P-ottimizzanti* e poi anche *P-ottimizzanti a complessità polinomiale*.

## Riferimenti bibliografici

- [1] A. Alessandri, M. Baglietto, T. Parisini, R. Zoppoli (1999), “A neural state estimator with bounded errors for nonlinear systems,” *IEEE Trans. on Automatic Control*, vol. 44, pp. 2028-2042.
- [2] A. Alessandri, M. Sanguineti, M. Maggiore (2002), “Optimization-based learning with bounded error for feedforward neural networks,” *IEEE Trans. on Neural Networks*, vol. 13, pp. 261-273.
- [3] M. Aoki (1967), *Optimization of Stochastic Systems*, Academic Press, New York.
- [4] M. Attouch (1984), *Variational Convergence for Functions and Operators*, Pitman Publishing, Marshfield, MA.
- [5] M. Baglietto, T. Parisini, R. Zoppoli (2001), “Numerical solutions to the Witsenhausen counterexample by approximating networks,” *IEEE Trans. on Automatic Control*, vol. 46, pp. 1471-1477.
- [6] M. Baglietto, T. Parisini, R. Zoppoli (2001), “Distributed-information neural control: the case of dynamic routing in traffic networks,” *IEEE Trans. on Neural Networks*, vol. 12, pp. 485-502.
- [7] A. R. Barron (1993), “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Trans. on Information Theory*, vol. 39, pp. 930-945.
- [8] R. W. Beard, T. W. McLain (1998), “Successive Galerkin approximation algorithms for nonlinear optimal and robust control,” *Int. J. of Control*, vol. 71, pp. 717-743.
- [9] R. Bellman (1957), *Dynamic Programming*, Princeton University Press, Princeton, N.J..
- [10] L. Breiman (1993), “Hinging hyperplanes for regression, classification, and function approximation,” *IEEE Trans. on Information Theory*, vol. 39, pp. 993-1013.
- [11] Yu. Ermoliev, R. J-B Wets (Eds.) (1980), *Numerical Techniques for Stochastic Optimization*, Springer-Verlag, Heidelberg.

- [12] U. Felgenhauer (1999), “On Ritz type discretizations for optimal control problems,” *Proc. 18th IFIP-ICZ Conference, Res. Notes in Math.*, Chapman-Hall, vol. 386, pp. 91-99.
- [13] I. M. Gelfand, S. V. Fomin (1963), *Calculus of Variations*, Prentice Hall, Englewood Cliffs, N. J..
- [14] F. Girosi (1993), “Regularization theory, radial basis functions and networks,” *From Statistics to Neural Networks. Theory and Pattern Recognition Applications*, J. H. Friedman, V. Cherkassky, H. Wechsler (Eds.), Computer and Systems Sciences Series, Springer-Verlag, Berlino, pp. 166-187.
- [15] F. Girosi, G. Anzellotti (1992), “Rates of convergence of approximation by translates,” *A.I. Memo 1288*, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.
- [16] F. Girosi, M. Jones, T. Poggio (1995), “Regularization theory and neural networks architectures,” *Neural Computation*, vol. 7, pp. 219-269.
- [17] S. Giulini, M. Sanguineti (2000), “On dimension-independent approximation by neural networks and linear approximators,” *Proc. Int. Joint Conf. on Neural Networks*, Como, Italia, pp. I283-I288.
- [18] L. Grippo (2000), “Convergent on-line algorithms for supervised learning in neural networks,” *IEEE Trans. on Neural Networks*, vol. 11, pp. 1284-1299.
- [19] W. W. Hager (1975), “The Ritz-Trefftz method for state and control constrained optimal control problems,” *SIAM J. on Numerical Analysis*, vol. 12, pp. 854-867.
- [20] L. K. Jones (1992), “A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training,” *Annals of Statistics*, vol. 20, pp. 608-613.
- [21] P. C. Kainen, V. Kůrková, M. Sanguineti (2003), “Minimization of error functionals over variable-basis functions,” *SIAM J. on Optimization*, 2003 (di prossima pubblicazione).
- [22] L. V. Kantorovich, V. I. Krylov (1958), *Approximate Methods of Higher Analysis*, P. Noordhoff Ltd, Groningen, Olanda.
- [23] A. N. Kolmogorov (1936), “Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse,” *Annals of Mathematics*, vol. 37, pp. 107-110 (English translation: “On the best approximation of functions of a given class” (1991), *Selected works of A. N. Kolmogorov*, vol. I, V. M. Tikhomirov Ed., Kluwer, pp. 202-205).

- [24] H. J. Kushner, G. G. Yin (1997), *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York.
- [25] V. Kůrková, M. Sanguineti (2001), “Bounds on rates of variable-basis and neural-network approximation,” *IEEE Trans. on Information Theory*, vol. 47, pp. 2659-2665.
- [26] V. Kůrková, M. Sanguineti (2002), “Comparison of worst case errors in linear and neural network approximation,” *IEEE Trans. on Information Theory*, vol. 48, pp. 264-275.
- [27] V. Kůrková, M. Sanguineti (2002), “Error estimates for approximate solution of optimization problems by approximating networks,” *Book of Abstracts of the Workshop Mathematical Diagnostics*, Erice, Italia, pp. 11-12. Si veda anche: V. Kůrková, M. Sanguineti (2002), “Error estimates for approximate optimization over variable-basis functions,” *Research Report ICS-2002-882*, Institute of Computer Science - Academy of Sciences of the Czech Republic.
- [28] V. Kůrková, M. Sanguineti (2003), “Neural network learning as approximate optimization”, *Proc. 6th Int. Conf. on Artificial Neural Networks and Genetic Algorithms (ICANNGA)*, Roanne, Francia, aprile 2003.
- [29] V. Kůrková, M. Sanguineti (2002), “Tight bounds on rates of variable-basis approximation via estimates of covering numbers,” *Research Report ICS-2002-865* - Institute of Computer Science - Academy of Sciences of the Czech Republic.
- [30] V. Kůrková, P. Savický, K. Hlaváčková (1998), “Representations and rates of approximation of real-valued boolean functions by neural networks,” *Neural Networks*, vol. 11, pp. 651-659.
- [31] A. Messner, M. Papageorgiou (1992), “Motorway network control via nonlinear optimization,” *Proc. 1st Meeting of the EURO Working Group on Urban Traffic and Transportation*, Landshut, Germania, pp. 1-24.
- [32] H. N. Mhaskar (1996), “Neural networks for optimal approximation of smooth and analytic functions,” *Neural Computation*, vol. 8, pp. 164-177.
- [33] M. Papageorgiou (1983), *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control*, Lecture Notes in Control and Information Sciences, Springer-Verlag, New York.
- [34] T. Parisini, R. Zoppoli (1994), “Neural networks for feedback feedforward nonlinear control systems,” *IEEE Trans. on Neural Networks*, vol. 5, pp. 436-449.
- [35] T. Parisini, R. Zoppoli (1998), “Neural approximations for infinite-horizon optimal control of nonlinear stochastic systems,” *IEEE Trans. on Neural Networks*, vol. 9, pp. 1388-1408.

- [36] J. Park, I. W. Sandberg (1991), “Universal approximation using radial-basis-function networks,” *Neural Computation*, vol. 3, pp. 246-257, 1991.
- [37] A. Pinkus (1985), *n-Widths in Approximation Theory*, Springer-Verlag, Berlino Heidelberg.
- [38] A. Pinkus (1999), “Approximation theory of the MLP model in neural networks,” *Acta Numerica*, vol. 8, pp. 143-196.
- [39] E. S. Plumer (1996), “Optimal control of terminal processes using neural networks,” *IEEE Trans. on Neural Networks*, vol. 7, pp. 408-418.
- [40] W. Ritz (1909), “Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik,” *J. für die reine und angewandte Mathematik*, vol. 135, pp. 1-61.
- [41] L. Schwartz (1967), *Analyse Mathématique*, Hermann, Parigi.
- [42] H. R. Sirisena, F. S. Chou (1979), “Convergence of the control parameterization Ritz method for nonlinear optimal control problems,” *J. of Optimization Theory and Applications*, vol. 29, pp. 369-382.
- [43] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, P.-Y. Glorennec, B. Delyon, H. Hjalmarsson, A. Juditsky (1995), “Nonlinear black-box modeling in system identification: a unified overview,” *Automatica*, vol. 31, pp. 1691-1724.
- [44] R. Zoppoli, T. Parisini (1992), “Learning techniques and neural networks for the solution of N-stage nonlinear nonquadratic optimal control problems,” *Systems, Models and Feedback: Theory and Applications*, A. Isidori, T. J. Tarn (Eds.), Birkhäuser, Boston, pp. 193-210.
- [45] R. Zoppoli, M. Sanguineti, T. Parisini (2002), “Approximating networks and Extended Ritz Method for the solution of functional optimization problems,” *J. of Optimization Theory and Applications*, vol. 112, pp. 403-440.